Likelihood and Exponential Families

by

Charles J. Geyer

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

1990

Approved by_____
(Chairperson of Supervisory Committee)

Program Authorized
to Offer Degree_____

Date_____

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature_____

Date_____

University of Washington

Abstract

Likelihood and Exponential Families

by Charles J. Geyer

Chairperson of Supervisory Committee: *Professor Elizabeth A. Thompson*
Department of Statistics

A family of probability densities with respect to a positive Borel measure on a finite-dimensional affine space is standard exponential if the log densities are affine functions. The family is convex if the natural parameter set (gradients of the log densities) is convex. In the closure of the family in the topology of pointwise almost everywhere convergence of densities, the maximum likelihood estimate (MLE) exists whenever the supremum of the log likelihood is finite. It is not defective if the family is convex. The MLE is a density in the original family conditioned on some affine subspace (the support of the MLE) which is determined by the "Phase I" algorithm, a sequence of linear programming feasibility problems. Standard methods determine the MLE in the family conditioned on the support ("Phase II").

An extended-real-valued function on an affine space is generalized affine if it is both convex and concave. The space of all generalized affine functions is a compact Hausdorff space, sequentially compact if the carrier is finite-dimensional. A family of probability densities is a standard generalized exponential family if the log densities are generalized affine. The closure of an exponential family is equivalent to a generalized exponential family.

When the likelihood of an exponential family cannot calculated exactly, it can sometimes be calculated by Monte Carlo using the Metropolis algorithm or the Gibbs sampler. The Monte Carlo log likelihood (the log likelihood in the exponential family generated by the Monte Carlo empirical distribution) then hypoconverges strongly (almost surely over sample paths) to the true log likelihood. For a closed convex family the Monte Carlo approximants to the MLE and all level sets of the likelihood

converge strongly to the truth. For nonconvex families, the outer set limits converge.

These methods are demonstrated by an autologistic model for estimation of relatedness from DNA fingerprint data and by isotonic, convex logistic regression for the maternal-age-specific incidence of Down's syndrome, both constrained MLE problems. Hypothesis tests and confidence intervals are constructed for these models using the iterated parametric bootstrap.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Many of the faculty and students in the Departments of Statistics, Biostatistics, and Mathematics at the University of Washington have influenced my thinking about statistics and mathematics and have contributed in some way or another to this thesis and to related research. The Department of Statistics has been an extremely exciting environment for learning about spatial statistics in general and the Gibbs sampler in particular. Lars Andersen, Julian Besag, Sun Wei Guo, John Haslett, Alan Lippman, Antonio Possolo, Nuala Sheehan, Elizabeth Thompson, and Jeremy York all contributed something to what I know about the Gibbs sampler and Monte Carlo maximum likelihood. Alan Lippman, in particular, who worked in this area for his dissertation, spent lots of time talking with me and contributed several useful ideas. The Department has also been an exciting place to learn about the bootstrap. Joe Felsenstein, Nhu Le, David Mason, Michael Newton, Jens Praestgaard, and Jon Wellner all taught me something about the bootstrap. Michael Newton, in particular, has taught me much about the bootstrap and the iterated bootstrap.

The theory of maximum likelihood in the closure of an exponential family presented here owes a great deal to the books of R. T. Rockafellar [62] and O. Barndorff-Nielsen [7]. Rockafellar's course on variational analysis and preprints of his new book with R. Wets [63] changed much of Chapter 5 in the last weeks of writing. Conversations with Rockafellar over the course of my graduate career introduced me to epiconvergence, pointed me toward the optimization software (MINOS, NPSOL, and LSSOL) used for the numerical computations in all of the examples, and steered me away from an unproductive approach to the Phase I problem and toward linear programming. Jim Burke, Alan Lippman, and John McDonald helped me learn to do numerical optimization.

Oliver Ryder of the Zoological Society of San Diego brought the DNA fingerprint problem to Elizabeth Thompson and me, giving us the fingerprint data for the California condors, an extremely interesting problem that was the impetus to work out both the Phase I algorithm and the method of Monte Carlo maximum likelihood. The Down's syndrome data that I used to work out methods for constrained optimization and the iterated parametric bootstrap came to my attention when T. R. Raghunathan assigned it as a homework problem in the graduate level applied statistics class. Some of the ideas about nonparametric regression expressed in Chapter 7 come from Werner Stuetzle and Jon Wellner. Jon Wellner also helped me with some of the probability theory.

In all respects Elizabeth Thompson has been a wonderful advisor. She has has a strong influence on the direction of this work, on my approach to scientific research, my ways of dealing with colleagues, and my philosophy of statistics. She is, of course, not to blame for the remaining flaws.

This research was supported in part by NSF grant BSR-8619760 and USDA grant 88-37151-3958.

In memoriam
My father, Charles J. Geyer, Jr.
He would have enjoyed seeing this.

Chapter 1

ELEMENTARY PROPERTIES OF EXPONENTIAL FAMILIES

1.1 Exponential Families on $\mathbb{R}^n$

The standard modern definition of an exponential family [7, 8, 18] is the following. A family $\mathcal{P} = \{ p_\phi : \phi \in \Phi \}$ of probability densities with respect to a σ-finite measure μ on a measurable space $(\Omega, \mathcal{A})$ is *exponential* if the densities have the form

$$p_\phi(\omega) = a(\phi)h(\omega) \exp\left(\sum_{i=1}^{n} x_i(\omega)\theta_i(\phi) \right), \qquad \omega \in \Omega \tag{1.1}$$

where h and $x_1, \ldots, x_n$ are measurable real functions on Ω and where a and $\theta_1, \ldots, \theta_n$ are real functions on Φ. A family of probability measures is *exponential* if it is dominated by a σ-finite measure and has an exponential family of densities with respect to that measure.

The usual identification of n-tuples with points in $\mathbb{R}^n$ and of the sum in (1.1) with the inner product

$$\langle x, \theta \rangle = \sum_{i=1}^{n} x_i\theta_i \tag{1.2}$$

simplifies the notation in (1.1) to

$$p_\phi(\omega) = a(\phi)h(\omega)e^{\langle x(\omega), \theta(\phi) \rangle}, \qquad \omega \in \Omega$$

where now x is a measurable function from Ω to $\mathbb{R}^n$ and θ is a function from Φ to $\mathbb{R}^n$.

From the point of view of the likelihood approach to statistics, one parametrization is as good as another, since all inferences should be independent of parametrization. Let $\Theta = \{ \theta(\phi) : \phi \in \Phi \}$ and

$$c(\theta) = \int e^{\langle x(\omega), \theta \rangle} \, d\mu(\omega), \qquad \theta \in \Theta.$$

Then

$$\mathcal{P} = \{\, p_\theta : \theta \in \Theta \,\} \tag{1.3}$$

where

$$p_\theta(\omega) = \frac{1}{c(\theta)} h(\omega) e^{\langle x(\omega), \theta \rangle}, \qquad \omega \in \Omega \tag{1.4}$$

defines the same exponential family as the general parametrization above. The parametrization defined by (1.3) and (1.4) is called the *natural* or *canonical* parametrization of the exponential family.

1.2 Standard Exponential Families on $\mathbb{R}^n$

Just as introduction of the natural parameter simplifies the model, so does the introduction of the natural statistic. For an exponential family given by (1.3) and (1.4) the function x is called the *natural* or *canonical* statistic of the family. The family induced by the natural statistic is called a standard exponential family [9, 18].

A family

$$\mathcal{F} = \{\, f_\theta : \theta \in \Theta \,\}$$

of probability densities with respect to a Borel measure λ on $\mathbb{R}^n$ is a *standard exponential family* if Θ is a nonempty subset of $\mathbb{R}^n$ and the densities have the form

$$f_\theta(x) = \frac{1}{c(\theta)} e^{\langle x, \theta \rangle}, \qquad x \in \mathbb{R}^n. \tag{1.5}$$

The measure induced by the density f_θ is denoted F_θ and is defined by

$$F_\theta(B) = \int_B f_\theta \, d\lambda = \int_B \frac{1}{c(\theta)} e^{\langle x, \theta \rangle} \, d\lambda(x), \qquad B \in \mathcal{B} \tag{1.6}$$

where $\mathcal{B}$ denotes the Borel σ-field of $\mathbb{R}^n$.

Note that the measure λ need not explicitly be required to be σ-finite. It is automatically σ-finite if the rest of the definition is satisfied. Let θ be any point of Θ, then the half space

$$A_k = \{\, x \in \mathbb{R}^n : \langle x, \theta \rangle \geq -k \,\}$$

is assigned probability

$$F_\theta(A_k) = \int_{A_k} \frac{1}{c(\theta)} e^{\langle x, \theta \rangle} \, d\lambda \geq \frac{1}{c(\theta)} e^{-k} \lambda(A_k).$$

Hence $\lambda(A_k) < \infty$, and, since $\bigcup_{k=1}^{\infty} A_k = \mathbb{R}^n$, the measure λ is σ-finite.

The relationship between standard and general exponential families is given by the following.

Theorem 1.1 *A family of probability densities with respect to a σ-finite measure is exponential if and only if it has a sufficient statistic that induces a standard exponential family.*

PROOF. By the Fisher-Neyman factorization theorem [5] a family of probability measures dominated by a σ-finite measure μ has a sufficient statistic x if and only if it has densities of the form

$$p_\theta(\omega) = f_\theta\Big(x(\omega)\Big)h(\omega).$$

So it remains only to be shown that if the family $\{p_\theta\}$ is exponential then f_θ has the form (1.5). By the change of variable theorem for integration

$$\int_B f_\theta \, d\lambda = \int_{x^{-1}(B)} f_\theta\Big(x(\omega)\Big)h(\omega) \, d\mu(\omega), \qquad B \in \mathcal{B}$$

where λ is the measure on $\mathbb{R}^n$ defined by

$$\lambda(B) = \int_{x^{-1}(B)} h(\omega) \, d\mu(\omega), \qquad B \in \mathcal{B}.$$

So the sufficient statistic x induces the family $\{f_\theta\}$ of densities with respect to λ. Clearly p_θ has the form (1.4) if and only if f_θ has the form (1.5). □

To a statistician who accepts the principle of sufficiency, there is no need for theoretical discussion of general exponential families. Reduction by sufficiency always produces a standard exponential family from which all inferences are drawn. Thus most of the theory in this dissertation will be stated for standard families only. An application of the change of variable theorem always produces the equivalent result for general families.

This definition of a standard exponential family as a family of densities or distributions on $\mathbb{R}^n$, though traditional, obscures some of the features of exponential families that are made clear when an abstract vector space or an abstract affine space is substituted for $\mathbb{R}^n$. Any finite dimensional vector space or affine space is, or course,

isomorphic to $\mathbb{R}^n$, so this does not change the nature of the mathematical object (exponential families) being defined. It does, however, simplify many definitions and arguments. Insisting that any affine space "is" $\mathbb{R}^n$ is tantamount to insisting that every definition and argument make explicit reference to a fixed origin and basis for the space and that whenever a different basis or origin needs to be considered an explicit change of basis or origin be described. Furthermore, it obscures the duality between the natural parameter space and sample space of the family when both are denoted $\mathbb{R}^n$.

In order to establish notation and give pointers to relevant reference works, Appendix A reviews the mathematical background necessary for the theory of exponential families presented in this dissertation. The topics covered are vector spaces, affine spaces, and convexity.

1.3 Standard Exponential Families on Vector Spaces

The definition of a standard exponential family on an abstract vector space is almost identical to the definition of Section 1.2. The only difference is that the vector space E which serves as the sample space and its dual space E^* which serves as the parameter space are distinguished, and no "standard basis" is singled out in advance for either space. This might be considered a purely notational change; one simply writes E or E^*, whichever is appropriate, in place of $\mathbb{R}^n$ wherever it occurs. The same sort of comment could, however, be made about any mathematical abstraction.

The advantage of using the notation of abstract vector spaces is that it does make a distinction between the sample and parameter spaces. The notation $\langle \cdot, \cdot \rangle$ should not be thought of as indicating an inner product on $\mathbb{R}^n$, a function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$ (as the text just before equation (1.2) says). Rather it should be thought of as indicating a "canonical bilinear form" placing E and E^* in duality (Section A.1), a function from $E \times E^*$ to $\mathbb{R}$. In this view the notation $\langle \cdot, \cdot \rangle$ only makes sense when one of the arguments takes values in the sample space and the other in the parameter space. If one is tempted to write down $\langle x, y \rangle$ when x and y are both points in the sample space, or $\langle \theta, \phi \rangle$ when θ and ϕ are both points in the parameter space, then one is in the process of committing an error or, at best, introducing an irrelevancy.

This position may seem to be a bit extreme, since it might be thought that one cannot even define the Euclidean topology on E without reference to the unit ball

and hence a distance function. This is wrong because the Euclidean topology for a finite-dimensional vector space E is the same as the "weak" topology, the weakest topology that makes all of the linear functions $x \mapsto \langle x, \theta \rangle$, $\theta \in E^*$ continuous. (This would not be true if E were infinite-dimensional, but then the distinction between E and E^* would be crucial.) Another place where one might think that an inner product is necessary is in discussion of asymptotics. The asymptotic distribution of the natural statistic (or the maximum likelihood estimator) has a covariance which induces an inner product on E (or E^*), but this inner product does not need to refer to some preexisting inner product. In any case, whether one finds these arguments convincing or not, the abstract vector space notation will be used throughout this dissertation.

Let λ be a Borel measure on a finite-dimensional real vector space E. A family

$$\mathcal{F} = \{ f_\theta : \theta \in \Theta \} \tag{1.7}$$

probability densities with respect to λ is a standard exponential family if the densities have the form

$$f_\theta(x) = \frac{1}{c(\theta)} e^{\langle x, \theta \rangle}, \qquad x \in E, \tag{1.8}$$

where Θ is a nonempty subset of E^*.

The "normalizing constant" $c(\theta)$ plays an important role in some of the theory. The function $c : E^* \to \overline{I\!R}$ defined by

$$c(\theta) = \int e^{\langle x, \theta \rangle} \, d\lambda(x), \qquad \theta \in E^* \tag{1.9}$$

is called the *Laplace transform* of the measure λ. (The notation $\overline{I\!R}$ denotes the compactified real line $[-\infty, +\infty]$ which is the subject of Section A.6). The theory of Laplace transforms that plays a role in the theory of exponential families is discussed in Section 2.1. For now it is enough to say that c is a convex function that is strictly positive. So a density f_θ is defined by (1.8) whenever $c(\theta)$ is finite, that is whenever θ lies in the set

$$\operatorname{dom} c = \{ \theta \in E^* : c(\theta) < +\infty \}$$

the *effective domain* of the function c (Section A.7). Necessarily, $\Theta \subset \operatorname{dom} c$, so $\operatorname{dom} c$ is not empty whenever c is the Laplace transform of a measure associated with an exponential family.

The family is said to be *full* if $\Theta = \operatorname{dom} c$. The set $\operatorname{dom} c$ is usually referred to as the *natural parameter space* of the family, although it is not a parameter space of the family but is the parameter space of the associated full family in its natural parametrization. This misnomer presumably survives because of the concentration on full families that characterizes most of the literature. Or perhaps one should say that concentration on full families (ignoring constraints on the parameters) survives because of this misnomer. It is still considered permissible by some to call a parameter value that does not lie in the parameter space Θ (a negative variance component, for example) a "maximum likelihood estimate" because it lies in the "natural parameter space" of the family, even though such an estimate is patently ridiculous. The name that will be applied to Θ throughout this dissertation will be *natural parameter set*, this being taken to indicate that Θ is a subset of $\operatorname{dom} c$. It is hoped that this terminology will not be too confusing.

1.4 Standard Exponential Families on Affine Spaces

Although the Laplace transform associated with an exponential family plays an important role in some parts of the theory, it plays no role in others. It simplifies much of the theory when the following definition, which makes no explicit reference to the Laplace transform, is used.

A family of probability densities with respect to a Borel measure λ on a finite-dimensional real affine space E is a *standard exponential family* if it is of the form

$$\mathcal{F} = \exp \mathcal{H} = \{\, e^h : h \in \mathcal{H} \,\} \tag{1.10}$$

for some nonempty subset $\mathcal{H}$ of $A(E)$, where, as defined in Section A.5, $A(E)$ is the set of all affine functions from E to $\mathbb{R}$. This is equivalent to the definition given by (1.7) and (1.8) because if E is a vector space and h is an affine function, then h has the form $x \mapsto \langle x, \theta \rangle + k$ for some θ in E^* where $k = h(0)$, and in order that h be a probability density it must integrate to one, so

$$e^{-k} = c(\theta) = \int e^{\langle x, \theta \rangle} \, d\lambda(x)$$

and

$$f(x) = \frac{1}{c(\theta)} e^{\langle x, \theta \rangle}$$

agreeing with (1.8).

We have finally arrived at a definition of exponential families simple enough to state without symbols. A family of densities on an affine space is standard exponential if the log densities are affine, a family of densities on an arbitrary space is exponential if it has a sufficient statistic that induces a standard family, and a dominated family of measures is exponential if it has an exponential family of densities with respect to the dominating measure.

The two characterizations of an exponential family in this section and the preceding section will be referred to as the "affine" picture and the "vector" picture of exponential families. The sample space is the same in both pictures (a finite-dimensional Euclidean space, though in the vector picture an origin is prescribed). The families of densities and log densities are the same objects (families of real functions on the sample space). But the parameter space, which plays an important role in the vector picture, seems absent from the affine picture. It can, however, be easily defined

$$\Theta = \{ \nabla h : h \in \mathcal{H} \} \tag{1.11}$$

as the set of gradients of the log densities of the family. This obviously agrees with the vector picture, since if h is $x \mapsto \langle x, \theta \rangle - \log c(\theta)$ then $\nabla h = \theta$. The only thing that is a bit tricky is defining the space in which Θ lies. In the vector picture, Θ is a subset of E^*, the dual space of the sample space. In the affine picture things are not so simple. The gradient of an affine function h is a map $z \mapsto h(x + z) - h(x)$ where x is any fixed point in the sample space E. In order that $x + z$ make sense when $x \in E$, z must be an element of the translation space T of the sample space E. Thus a gradient is a linear functional on T and hence an element of T^*. This shows why the vector picture is often the simplest picture to use. The natural parameter is a much simpler object there.

1.5 Uniqueness, Regularity, and Steepness

The main reason for introducing the affine picture is to simplify the treatment of the closure of an exponential family, which is the subject of Chapter 4. There are also simplifications in other areas of the theory, one of which is the problem of "non-uniqueness" or "lack of identifiability." This is usually dealt with by the introduction

of the "minimal" representation of the exponential family as in the monographs of Barndorff-Nielsen [7, pp. 112-113] and Brown [18, pp. 13-16]. This notion introduces much needless complexity in the theory. The blame rests entirely with the vector picture. In the affine picture the situation is very simple.

The *support* of a σ-finite Borel measure λ on a finite-dimensional affine space is the smallest closed set whose complement has λ-measure zero. This is the intersection of all closed sets whose complements have measure zero. This set is closed, hence measurable, and has measure zero because a σ-finite Borel measure on a Euclidean space is Radon. The *convex support* of λ is the smallest convex set whose complement has measure zero. It is the convex hull of the support. The *affine support* of λ is the smallest affine set whose complement has measure zero. It is the affine hull of the support.

If λ is a measure that is associated with a standard exponential family, it is σ-finite because its Laplace transform is not identically $+\infty$, so these supports are well-defined. Let L be the affine support of λ. Clearly, the values the densities in the exponential family take on the complement of L are irrelevant. Only the restrictions to L determine probabilities. But these restrictions constitute another exponential family of densities. To be precise let

$$\mathcal{F}|L = \{\, f|L : f \in \mathcal{F} \,\}$$

denote the restriction of $\mathcal{F}$ to L, and let $\lambda|L$ denote the restriction of λ to L defined by

$$(\lambda|L)(B) = \lambda(B \cap L), \qquad B \in \mathcal{B}.$$

Then $\mathcal{F}|L$ is a standard exponential family of densities with respect $\lambda|L$ and the two families obviously define the "same" probability measures, since

$$\int_B f \, d\lambda = \int_{B \cap L} f \, d\lambda = \int_{B \cap L} f|L \, d(\lambda|L)$$

for any Borel set B in E.

The family $\mathcal{F}|L$ has the desired "uniqueness" or "identifiability" property. Distinct elements of $\mathcal{F}|L$ correspond to distinct probability measures. For if f_1 and f_2 are densities in $\mathcal{F}$ which correspond to the same probability measure, then $f_1 = f_2$ almost everywhere, so the set

$$L_{12} = \{\, x \in E : f_1(x) = f_2(x) \,\}$$

supports λ. But L_{12} is affine, since it is the locus of points where the two affine functions $\log f_1$ and $\log f_2$ are equal. Hence $L \subset L_{12}$ and $f_1|L = f_2|L$.

In the vector picture the situation is more complicated because L need not be a vector subspace. So the operation of reduction to a minimal representation cannot be described as a restriction. It is necessary to introduce a new vector space of the same dimension as L (its translation space, for example) to carry the minimal representation. A more geometric way of thinking of the operation is that the origin is being shifted to lie in L because of the insistence that an exponential family live on a vector space. The situation is even more complicated in the "$\mathbb{R}^n$ picture" because L need not have a basis that is a subset of the standard basis for $\mathbb{R}^n$. It is necessary to find a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^k$ to describe the identification of L with $\mathbb{R}^k$. Again the only reason for introducing this linear transformation is the insistence that an exponential family live on $\mathbb{R}^k$.

Two other concepts that figure prominently in the other presentations of the theory of exponential families are regularity and steepness [7, p. 116–117]. Their sole function is to exclude from consideration all cases where the maximum likelihood estimate cannot be characterized as a point where the gradient of the likelihood is zero. Since one of the main features of this dissertation is an interest in constrained problems, that is in precisely those cases that the concepts of regularity and steepness exclude, neither concept will play any role in the theory presented in this dissertation. It turns out that these concepts are not missed. A completely satisfactory theory can be constructed without them.

Chapter 2

MAXIMUM LIKELIHOOD IN CONVEX EXPONENTIAL FAMILIES

The subject of maximum likelihood estimation in exponential families has a long history, and the characterization of the conditions under which a maximum likelihood estimate exists (that is when there some point in the parameter space where the likelihood achieves its supremum) has received much attention [6, 7, 35, 36, 74, 1, 18, 45]. The question of what to do when the maximum likelihood estimate does not exist in this traditional sense has received less attention, but a complete solution has been worked out for full exponential families supported on a finite set [6, 7]. In this case the maximum likelihood estimate always exists in the closure of the family (closure here meaning closure in the topology of convergence in distribution or of convergence of densities almost everywhere, these being the same for families with finite support). The maximum likelihood estimate is always a distribution in the family obtained by conditioning the original family on the smallest face of the convex hull of the support that contains the observation. This kind of construction of maximum likelihood estimates in the closure of the family has been extended to some full families with countable support [18] and to unconstrained generalized linear models [73]. A characterization of the closure of a convex exponential family with finite support seems to have been known but does not seem to have appeared in print [7, See p. 164].

In this chapter a complete solution will be presented to the problem of finding a maximum likelihood estimate in a convex exponential family (one whose natural parameter set is convex). No unnecessary regularity conditions are assumed. The maximum likelihood estimate in the closure of the family in the topology of almost everywhere convergence of densities is shown to exist for all discrete and all continuous families. It exists if and only if the supremum of the likelihood is finite (this applies to all families, discrete, continuous, or mixed). This chapter will use only the vector picture of exponential families. The affine picture will reappear in Chapter 4.

When it exists the maximum likelihood estimate is always the maximum likelihood estimate in a family obtained by conditioning the original family on some affine set A, which is referred to here as the *support* of the maximum likelihood estimate. The determination of A is a problem in convex geometry that proceeds by techniques completely different from the usual methods employed for maximum likelihood. It is in essence a sequence of linear programming feasibility problems. This suggests for this part of the computation the name "Phase I maximum likelihood problem" (after Phase I of a linear programming problem, finding a feasible point). Once the Phase I problem has been solved resulting in the determination of the support A of the maximum likelihood estimate or in the determination that the supremum of the likelihood is $+\infty$ and no estimate exists, not even in the closure of the family, the maximum likelihood estimate is determined (if it exists) by standard methods applied to the family conditioned on A. This standard part of the procedure is referred to here as the "Phase II" maximum likelihood problem. It is not different in methodology from any other maximum likelihood problem. The only difference is that the problem being solved is not the original one (unless the support A of the estimate is a support of the original family, a *trivial* solution to the Phase I problem), and the solution is known to exist (unless the Phase I calculations discovered that the supremum of the likelihood was not finite). This chapter and the next two chapters are concerned entirely with the Phase I problem and the characterization of the closure of an exponential family. Methods for Phase II will not be discussed until Chapters 6 and 7.

The notion of the Phase I algorithm, of actually calculating maximum likelihood estimates in the closure of the family, seems to be entirely new. That linear programming without iteration could be used to determine whether the maximum likelihood estimate exists (in the traditional sense) in logistic regression has been noted by Albert and Anderson [1].

2.1 Laplace Transforms

Let λ be a nonzero positive Borel measure on a finite-dimensional real vector space E. Then the *Laplace transform* of λ is the function $c : E^* \to \overline{I\!R}$ defined by

$$c(\theta) = \int e^{\langle x, \theta \rangle} \, d\lambda(x), \qquad \theta \in E^*,$$

which just repeats (1.9).

Some fundamental properties of Laplace transforms and log Laplace transforms are summarized in the following theorem, which is Theorem 7.1 in Barndorff-Nielsen [7] and Theorem 1.13 in Brown [18]. Convex and strictly convex functions, including the notion of strict convexity in a direction, are defined in Section A.7.

Theorem 2.1 *The Laplace transform c of a nonzero measure λ and its logarithm are both lower semicontinuous convex functions and are proper convex functions if they are not identically $+\infty$. Both are strictly convex in a direction ϕ if and only if λ is not concentrated on a hyperplane normal to ϕ.*

PROOF. First note that c is strictly positive, because if $c(\theta) = 0$ for any θ, then $e^{\langle x,\theta\rangle} = 0$ for almost all x, which is impossible since $e^{\langle x,\theta\rangle}$ is never zero and λ is assumed to be nonzero. Hence $\log c$ is never $-\infty$. Thus both are proper unless identically $+\infty$.

An application of Hölder's inequality gives

$$c(t\theta_1 + (1-t)\theta_2) \le c(\theta_1)^t c(\theta_2)^{(1-t)}$$

with equality if and only if $\langle x, \theta_1 - \theta_2\rangle$ is constant for almost all x (see, for example, Rudin [65, pp. 63–65] or Barndorff-Nielsen [7, pp. 100–101] for the conditions for equality in Hölder's inequality). This shows that $\log c$ is convex and is strictly convex in a direction $\phi = \theta_1 - \theta_2 \neq 0$ if and only if $\langle x, \phi\rangle$ is not constant almost everywhere, which is the same as saying that λ is not concentrated on a hyperplane normal to ϕ.

That c is convex follows from the fact that e^f is convex whenever f is [62, p. 32]. That c is strictly convex in exactly the same directions that $\log c$ is strictly convex, follows from the fact that $f \mapsto e^f$ is strictly increasing.

That c is lower semicontinuous follows from Fatou's lemma, and that $\log c$ is lower semicontinuous follows from the lower semicontinuity of c and the continuity of the logarithmic function. □

Let K be the convex support of the measure λ, and let σ_K denote its support function (Section A.10)

$$\sigma_K(\phi) = \sup\{\, \langle x, \phi\rangle : x \in K \,\}.$$

For each direction ϕ in E^* define

$$H_\phi = \{\, x \in E : \langle x, \phi \rangle = \sigma_K(\phi) \,\}. \tag{2.1}$$

Then H_ϕ is the supporting hyperplane to the set K with normal vector ϕ [62, p. 100]. Note that the word "support" gets used in two different meanings here. K is a support of λ in the sense of measure theory. H_ϕ is a supporting hyperplane of K in the sense of convex analysis.

The following theorem is fundamental. It is a minor variation of a statement given without proof on page 105 in Barndorff-Nielsen [7].

Theorem 2.2 *Let c be the Laplace transform of λ and K the convex support of λ. Then for any nonzero ϕ*

$$c(\theta + s\phi)e^{-as} \to \begin{cases} 0, & a > \sigma_K(\phi) \\ c(\theta)F_\theta(H_\phi), & a = \sigma_K(\phi) \\ +\infty, & a < \sigma_K(\phi) \end{cases}$$

as $s \to +\infty$, where F_θ is as has been defined in (1.6).

Proof. First let $a = \sigma_K(\phi)$. Note that

$$c(\theta + s\phi)e^{-\sigma_K(\phi)s} = \int e^{\langle x,\theta\rangle - s[\sigma_K(\phi) - \langle x,\phi\rangle]}\, d\lambda(x) \tag{2.2}$$

and that the integrand is nonincreasing in s almost everywhere $[\lambda]$ (that is, for all x such that $\langle x, \phi\rangle \le \sigma_K(\phi)$). Hence the the integral in (2.2) converges by the monotone convergence theorem to

$$\int_{H_\phi} e^{\langle x,\theta\rangle}\, d\lambda(x) = c(\theta)F_\theta(H_\phi) \tag{2.3}$$

since the integrand in (2.2) converges to the integrand in (2.3) for x in H_ϕ and converges to 0 for x such that $\langle x, \phi\rangle < \sigma_K(\phi)$. This proves the case $a = \sigma_K(\phi)$.

The case $a > \sigma_K(\phi)$ follows immediately since

$$\begin{aligned} \lim_{s\to\infty} c(\theta + s\phi)e^{-as} &= \left(\lim_{s\to\infty} c(\theta + s\phi)e^{-\sigma_K(\phi)s}\right)\left(\lim_{s\to\infty} e^{[\sigma_K(\phi)-a]s}\right) \\ &= c(\theta)F_\theta(H_\phi) \cdot 0 = 0 \end{aligned}$$

Thus only the case $a < \sigma_K(\phi)$ remains. If $\lambda(H_\phi) > 0$, then (as in the preceding case)

$$\begin{aligned}\lim_{s\to\infty} c(\theta + s\phi)e^{-as} &= \left(\lim_{s\to\infty} c(\theta + s\phi)e^{-\sigma_K(\phi)s}\right)\left(\lim_{s\to\infty} e^{[\sigma_K(\phi)-a]s}\right)\\ &= c(\theta)F_\theta(H_\phi)\cdot(+\infty) = +\infty\end{aligned}$$

If, however, $\lambda(H_\phi) = 0$, more work is required. Choose $\delta > 0$ small enough so that the sets

$$\begin{aligned}A_\delta &= \{\, x \in E : a \le \langle x, \phi\rangle \le \sigma_K(\phi) - 2\delta \,\}\\ B_\delta &= \{\, x \in E : \sigma_K(\phi) - \delta \le \langle x, \phi\rangle \le \sigma_K(\phi) \,\}\end{aligned}$$

have positive λ-measure (this is always possible because $\bigcup_{\delta>0} A_\delta$ has positive λ-measure and $\lambda(B_\delta) > 0$ for all $\delta > 0$). Then

$$F_{\theta+s\phi}(A_\delta) \le \frac{F_{\theta+s\phi}(A_\delta)}{F_{\theta+s\phi}(B_\delta)} \le e^{-s\delta}\frac{F_\theta(A_\delta)}{F_\theta(B_\delta)} \to 0$$

as $s \to \infty$. But also

$$F_{\theta+s\phi}(A_\delta) = \frac{c(\theta)}{c(\theta + s\phi)}\int e^{s\langle x,\phi\rangle}\, dF_\theta(x) \ge \frac{c(\theta)}{c(\theta + s\phi)}e^{as}F_\theta(A_\delta). \tag{2.4}$$

so the right hand side of (2.4) goes to 0 as $s \to \infty$ and $c(\theta + s\phi)e^{-as} \to \infty$, which establishes the last case. □

The recession function of a convex function is defined in Section A.12.

Theorem 2.3 *Let λ be a measure with proper Laplace transform c and convex support K. Then the recession function of $\log c$ is the support function of K.*

PROOF. Since $\log c$ is lower semicontinuous (Theorem 2.1), for any θ in $\operatorname{dom} c$ (there is one since c is proper)

$$\begin{aligned}(\operatorname{rc}\log c)(\phi) &= \lim_{s\to\infty}\frac{c(\theta + s\phi) - c(\theta)}{s}\\ &= \lim_{s\to\infty}\log\left(\left[\frac{c(\theta + s\phi)}{c(\theta)e^{s\sigma_K(\phi)}}\right]^{1/s} e^{\sigma_K(\phi)}\right)\end{aligned} \tag{2.5}$$

If $\lambda(H_\phi) > 0$ then the term in square brackets converges to $F_\theta(H_\phi)$ by Theorem 2.2 and the right hand side of (2.5) converges to $\sigma_K(\phi)$ as was to be proved.

Otherwise, since the left hand side of (2.4) is less than or equal to one,

$$\frac{c(\theta + s\phi)}{c(\theta)e^{s\sigma_K(\phi)}} \geq e^{[a-\sigma_K(\phi)]s} F_\theta(A_\delta)$$

for some δ such that $F_\theta(A_\delta) > 0$. But since $\langle x, \phi \rangle \leq \sigma_K(\phi)$ a. e. $[\lambda]$,

$$c(\theta + s\phi) \leq c(\theta)e^{s\sigma_K(\phi)}, \qquad s \in I\!R. \tag{2.6}$$

Thus

$$1 \geq \left[\frac{c(\theta + s\phi)}{c(\theta)e^{s\sigma_K(\phi)}} \right]^{1/s} \geq e^{a-\sigma_K(\phi)} F_\theta(A)^{1/s} \to 1$$

as $s \to \infty$, and again the the right hand side of (2.5) converges to $\sigma_K(\phi)$. □

2.2 Directions of Recession and Constancy

The unconstrained log likelihood corresponding to an observation x for a standard exponential family of densities with respect to a measure λ with Laplace transform c is the function $\tilde{l}_x$ defined by

$$\tilde{l}_x(\theta) = \log f_\theta(x) = \langle x, \theta \rangle - \log c(\theta). \tag{2.7}$$

Since it is the sum of a linear function and the negative of a convex function (both of which are convex) it is a concave function. Since $\log c$ is lower semicontinuous, l_x is upper semicontinuous.

Let Θ be the natural parameter set of the family. Then the (constrained) log likelihood of the family is the function l_x defined by

$$l_x(\theta) = \tilde{l}_x(\theta) - \delta_\Theta(\theta) \tag{2.8}$$

where δ_Θ is the indicator function of Θ defined by

$$\delta_\Theta(\theta) = \begin{cases} 0, & \theta \in \Theta \\ +\infty, & \text{otherwise} \end{cases}$$

Maximizing l_x is equivalent to maximizing $\tilde{l}_x$ subject to the constraint that the solution lie in Θ. The term "constrained log likelihood" will be used to denote l_x only when necessary to contrast l_x and $\tilde{l}_x$. Otherwise l_x will be called simply the "log likelihood."

A vector ϕ is a *direction of constancy* of the exponential family if λ is concentrated on a hyperplane normal to ϕ, that is if $y \mapsto \langle y, \phi \rangle$ is constant almost everywhere. The latter characterization makes it obvious that the set of all directions of constancy is a vector subspace (of the dual space E^* of the space E where λ lives). This subspace is called the *constancy space* of the family.

A vector ϕ is a *direction of constancy* of the log likelihood l_x if

$$l_x(\theta + s\phi) = l_x(\theta), \qquad \forall s \in I\!R.$$

In order to get a simple characterization of when a maximum likelihood estimate exists, it is necessary to require that the directions of constancy of the family be the directions of constancy of the log likelihood. Let Φ denote the constancy space. For $\phi \in \Phi$, there is some constant k such the $\langle y, \phi \rangle = k$ for almost all y. In particular $\langle x, \phi \rangle = k$ with probability one (x being the observation). Hence (with probability one)

$$c(\theta + s\phi) = e^{ks} c(\theta)$$

and

$$\tilde{l}_x(\theta + s\phi) = \tilde{l}_x(\theta), \qquad \forall s \in I\!R \tag{2.9}$$

Conversely, if (2.9) holds then $\tilde{l}_x$ is not strictly concave in the direction ϕ and hence c is not strictly convex, so $\phi \in \Phi$. Thus the directions of constancy of the family are exactly the directions of constancy of the unconstrained log likelihood (for almost all observations).

They will also be the directions of recession of the constrained log likelihood if $\theta + s\phi \in \Theta$ for all $s \in I\!R$, for each $\theta \in \Theta$ and for each $\phi \in \Phi$, that is if

$$\Theta = \Theta + \Phi = \{\, \theta + \phi : \theta \in \Theta,\ \phi \in \Phi \,\}. \tag{2.10}$$

A standard exponential family will be said to be *closed* if its natural parameter set Θ and constancy space Φ satisfy (2.10) and if its natural parameter set is closed relative

to the effective domain of its Laplace transform c, that is

$$\Theta = (\operatorname{cl}\Theta) \cap (\operatorname{dom} c). \tag{2.11}$$

Combining these two operations one gets the notion of an "almost everywhere" closure of the natural parameter set

$$\text{a-cl}\,\Theta = \Big(\operatorname{cl}(\Theta + \Phi)\Big) \cap (\operatorname{dom} c).$$

Changing the parameter space from Θ to a-cl Θ does not essentially change the problem. Every density in the family with natural parameter set a-cl Θ is a pointwise almost everywhere limit of densities in the family with natural parameter set Θ.

For a closed convex family, the (constrained) log likelihood is (for almost all observations) a proper upper semicontinuous concave function [from (2.11)] whose directions of constancy are exactly the directions of constancy of the family [from (2.10)] and which is strictly concave in every direction that is not a direction of constancy [from Theorem 2.1].

We also need to know about directions of recession (Section A.12) of the log likelihood l_x of a convex family. In order to avoid a separate definition for the recession function (in the hypographical sense rather than the epigraphical sense) of a concave function, the recession function of $-l_x$, will be calculated.

Theorem 2.4 *The recession function of the negative log likelihood is given by*

$$(\operatorname{rc} -l_x)(\phi) = \sigma_K(\phi) - \langle x, \phi \rangle + \delta_{\operatorname{rc}\Theta}(\phi) \tag{2.12}$$

PROOF. For $\theta \in \text{r-int}\,\Theta$

$$\begin{aligned}(\operatorname{rc} -l_x)(\phi) &= \lim_{s\to\infty} \frac{l_x(\theta) - l_x(\theta + s\phi)}{s} \\ &= \left(\lim_{s\to\infty} \frac{c(\theta + s\phi) - c(\theta)}{s}\right) - \langle x, \phi \rangle + \left(\lim_{s\to\infty} \delta_\Theta(\theta + \phi)\right)\end{aligned}$$

which equals (2.12) by Theorems 2.3 and A.2. □

Corollary 2.4.1 *A nonzero vector ϕ is a direction of recession of the log likelihood l_x for a convex family if and only if*

$$\sigma_K(\phi) \le \langle x, \phi \rangle \tag{2.13}$$

and

$$\phi \in \operatorname{rc} \Theta.$$

PROOF. These are just the ϕ such that $(\operatorname{rc} - l_x)(\phi) \le 0$. □

An equivalent way to state (2.13) is that ϕ is a nonzero normal to the convex support K of the family at the observation x.

2.3 Maximum Likelihood

A theorem for the "existence" of the maximum likelihood estimate in the usual sense can now be stated.

Theorem 2.5 *The maximum likelihood estimate in a closed convex exponential family exists (the supremum of the likelihood is achieved) if and only if every direction of recession of the log likelihood is a direction of constancy. [Recall that "closed" here means that (2.10) and (2.11) are satisfied.]*

PROOF. The log likelihood is a proper upper semicontinuous concave function. By Theorem 27.1 in Rockafellar [62], the supremum is achieved if every direction of recession is a direction of constancy.

Conversely, if there is a direction of recession ϕ that is not a direction of constancy, then the log likelihood l_x is nondecreasing (Theorem A.3) in the direction ϕ and strictly concave (Theorem 2.1 and the definition of "closed"). A nondecreasing strictly convex function is necessarily strictly increasing where it is finite, so $l_x(\theta) < l_x(\theta + s\phi)$ for $s > 0$, and the supremum is not attained. □

This theorem is a dual in a sense to Theorem 8.6 in Brown [18], which characterizes the existence of the maximum likelihood estimate in terms of the tangent cone to K at x and the barrier cone of Θ rather than (as here) the normal cone to K at x and the recession cone of Θ. There are three reasons why this theorem is an improvement. First, it is stronger than Brown's theorem, which is restricted to steep families. Second, it is more suited to computation, as will be shown presently. Third, even when the maximum likelihood estimate does not "exist" in the traditional sense, it may exist in a sense as a limit of a sequence of densities maximizing the likelihood. The theorem presented here almost directly produces this limit.

For any measurable set A such that $\lambda(A) > 0$, define the conditional density given A by

$$f_\theta(x|A) = \begin{cases} \frac{1}{F_\theta(A)}, & x \in A \\ 0, & \text{otherwise} \end{cases}$$

Theorem 2.6 *For a closed convex standard exponential family of densities with respect to a measure λ and ϕ a direction of recession of the natural parameter set, if $\lambda(H_\phi) > 0$ [with H_ϕ given by (2.1)], then*

$$f_{\theta+s\phi}(\,\cdot\,) \to f_\theta(\,\cdot\,|H_\phi), \text{ a. e. } [\lambda] \qquad \text{as } s \to \infty \tag{2.14}$$

holds for each θ in the natural parameter set Θ. Conversely, if $\lambda(H_\phi) = 0$, then

$$f_{\theta+s\phi}(x) \to \infty \qquad \text{as } s \to \infty. \tag{2.15}$$

PROOF. Since ϕ is a direction of recession, $\sigma_K(\phi) < \langle x, \phi \rangle$ by (2.13). But this implies

$$\theta + s\phi \in \operatorname{dom} c, \qquad \forall \theta \in \operatorname{dom} c, \ \forall s \geq 0$$

by (2.6). Also

$$\theta + s\phi \in \operatorname{cl} \Theta, \qquad \forall \theta \in \operatorname{cl} \Theta$$

by Theorem A.2. Thus, since the family is closed, (2.11) holds and

$$\theta + s\phi \in \Theta, \qquad \forall \theta \in \Theta, \ \forall s \geq 0.$$

Now

$$f_{\theta+s\phi}(x) = \frac{c(\theta)e^{\langle x,\phi \rangle}}{c(\theta + s\phi)} f_\theta(x)$$

so (2.14) and (2.15) both follow directly from Theorem 2.2. □

Hence when there exists a direction of recession ϕ that is not a direction of constancy one of two situations can occur.

(a) $\lambda(H_\phi) > 0$. Then the family

$$\mathcal{F}_A = \{\, f_\theta(\,\cdot\,|A) : \theta \in \Theta \,\},$$

where $A = H_\phi$, consists of limits of densities in the original family $\mathcal{F}$. Moreover, $\mathcal{F}_A$ is (up to almost everywhere equivalence) a standard exponential family of densities with respect to the restriction of λ to the affine subspace A. Furthermore, since ϕ is not a direction of constancy, $F_\theta(A) < 1$ for all θ, and each density f_θ in $\mathcal{F}$ is dominated by a density, namely $f_\theta(\,\cdot\,|A)$, in $\mathcal{F}_A$. Thus the likelihood for $\mathcal{F}_A$ dominates the likelihood for $\mathcal{F}$, but the supremum of the likelihood is the same for both (because the limit does not increase the supremum). The maximum likelihood estimate in the almost everywhere closure of $\mathcal{F}_A$ (if it exists) is a pointwise almost everywhere limit of densities in the original family that maximizes the original likelihood.

(b) $\lambda(H_\phi) = 0$. In this case, from Theorem 2.6, the supremum of the likelihood is $+\infty$. Taking almost everywhere limits does not produce a maximum likelihood estimate. In fact, the same analysis applied to the other case shows that instead of (2.14) we get

$$f_{\theta+s\phi}(\,\cdot\,) \to 0 \text{ a. e. } [\lambda] \qquad \text{as } s \to \infty.$$

In case (b) we give up and say that the maximum likelihood estimate does not exist, even as a pointwise almost everywhere limit. In case (a), $\mathcal{F}_A$ is a convex exponential family but it need not be a closed one for either of two reasons. First $\lambda|A$ has a larger constancy space than λ, since ϕ is a direction of constancy of $\lambda|A$ but not of λ. Second, the effective domain of the Laplace transform of $\lambda|A$ may include points that are not in the effective domain of the Laplace transform of λ or even translates of such points by multiples of ϕ. The following example shows this behavior.

Example 2.1 Let λ be a discrete measure in $\mathbb{R}^2$ with atoms of mass one at $(0, k)$, $k = 0, 1, \ldots$ and of mass $e^{\alpha k}$ at $(1, k)$, $k = 0, 1, \ldots$ for some $\alpha > 0$. Then the Laplace transform of λ is

$$c(\theta) = \sum_{i=0}^{\infty} e^{k\theta_2} + \sum_{i=0}^{\infty} e^{\theta_1 + k\theta_2} e^{\alpha k} = \begin{cases} \frac{1}{1-e^{\theta_2}} + e^{\theta_1} \frac{1}{1-e^{\theta_2+\alpha}}, & \theta_2 < -\alpha \\ +\infty, & \theta_2 \geq -\alpha \end{cases}$$

Consider the full family with natural parameter space

$$\Theta = \operatorname{dom} c = \{\, \theta \in \mathbb{R}^2 : \theta_2 < -\alpha \,\},$$

and suppose the observation x has $x_1 = 0$. Then x is on the boundary of the convex support with normal vector $\phi = (-1, 0)$. The maximum likelihood estimate lies in the family conditioned on

$$H_\phi = \{\, (0, k) : k = 0, 1, \ldots \,\}$$

But the Laplace transform of λ conditioned on this set is

$$c_\phi(\theta) = \sum_{i=0}^{\infty} e^{k\theta_2} = \begin{cases} \frac{1}{1 - e^{\theta_2}} & \theta_2 < 0 \\ +\infty, & \theta_2 \geq 0 \end{cases}$$

Hence the closure of the natural parameter set is

$$\text{a-cl}\, \Theta = \{\, \theta \in I\!R^2 : \theta_2 \leq -\alpha \,\}.$$

The additional points on the boundary (where $\theta_2 = \alpha$) can occur as maximum likelihood estimates. The family conditioned on H_ϕ is a geometric family with mean value parameter $p(\theta) = e^\theta$. For

$$x_2 \geq \frac{p(\theta)}{1 - p(\theta)} = \frac{e^\alpha}{1 - e^\alpha}$$

the gradient of the log likelihood is never zero and the maximum of the likelihood occurs at the boundary, that is, $\hat{\theta}_2 = \alpha$, as can be verified by direct calculation.

After taking the almost everywhere closure of the family $\mathcal{F}_A$, we are back in exactly the same situation as we started in. Our problem is to find the maximum likelihood estimate in a closed convex exponential family $\mathcal{F}_A$. The maximum likelihood estimate may exist in $\mathcal{F}_A$. If it does not, there is another direction of recession ϕ that is not a direction of constancy. If H_ϕ has probability zero, the maximum likelihood estimate does not exist, even as a limit, otherwise it may exist in the family conditioned on $A \cap H_\phi$. So A is replaced with $A \cap H_\phi$, and the process continues. This iteration must eventually stop, because the dimension of A decreases in each iteration (ϕ is not normal to the old A but is normal to the new). Dimensions being integer-valued, the dimension of A must eventually become zero if the iteration does not stop with a maximum likelihood estimate found at an earlier step or the supremum of the likelihood being found to be $+\infty$. If the dimension of A reaches zero, A is a single point. The family $\mathcal{F}_A$ has exactly one distribution, the distribution concentrated at

this single point, and this is the maximum likelihood estimate (all of the densities in the family are equal almost everywhere and all are maximum likelihood estimates).

This argument is summarized in the following.

Theorem 2.7 *A maximum likelihood estimate for a closed, convex exponential family $\mathcal{F}$ exists as a pointwise almost everywhere limit of densities in the family if and only if the supremum of the likelihood is finite. If the supremum is finite, the estimate lies in the original family if and only if every direction of recession of the log likelihood is a direction of constancy (that is if every normal to the convex support K at the observation x is also normal to the affine support or is not a direction of recession of the natural parameter set). Otherwise, there is a direction of recession ϕ that is not a direction of constancy, and the estimate lies in the closure of the family $\mathcal{F}_A$ conditioned on the $A = H_\phi$.*

The following concrete example illustrates this iterative process.

Example 2.2 (The Multinomial Distribution) The sample space is the set of all vectors $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$ whose coordinates are natural numbers that sum to N, the multinomial sample size,

$$\mathcal{S} = \{\, x \in \mathbb{R}^n : x_i \in \mathbb{N},\ \forall i,\ \text{and}\ \sum_{i=1}^{n} x_i = N \,\}.$$

The sample space for a multinomial distribution with $n = 3$ (a trinomial distribution) and $N = 4$ is shown in Figure 2.1. The measure λ generating the family has atoms at the points of $\mathcal{S}$ with mass proportional to the multinomial coefficients

$$\lambda(\{x\}) = \binom{N}{x} = \frac{N!}{x_1! x_2! \cdots x_n!}, \qquad x \in \mathcal{S},$$

and hence the Laplace transform is

$$c(\theta) = \int e^{\langle x, \theta \rangle} \lambda(x) = \sum_{x \in \mathcal{S}} \binom{N}{x} \prod_{i=1}^{n} \left(e^{\theta_i}\right)^{x_i} = \left(\sum_{i=1}^{n} e^{\theta_i} \right)^N$$

This gives the densities with respect to λ the form

$$f_\theta(x) = \frac{1}{c(\theta)} e^{\langle x, \theta \rangle} = \frac{\prod_{i=1}^{n} \left(e^{\theta_i}\right)^{x_i}}{\left(\sum_{j=1}^{n} e^{\theta_j} \right)^N} = \prod_{i=1}^{n} \frac{e^{\theta_i}}{\sum_{j=1}^{n} e^{\theta_j}} = \prod_{i=1}^{n} p_i^{x_i}$$

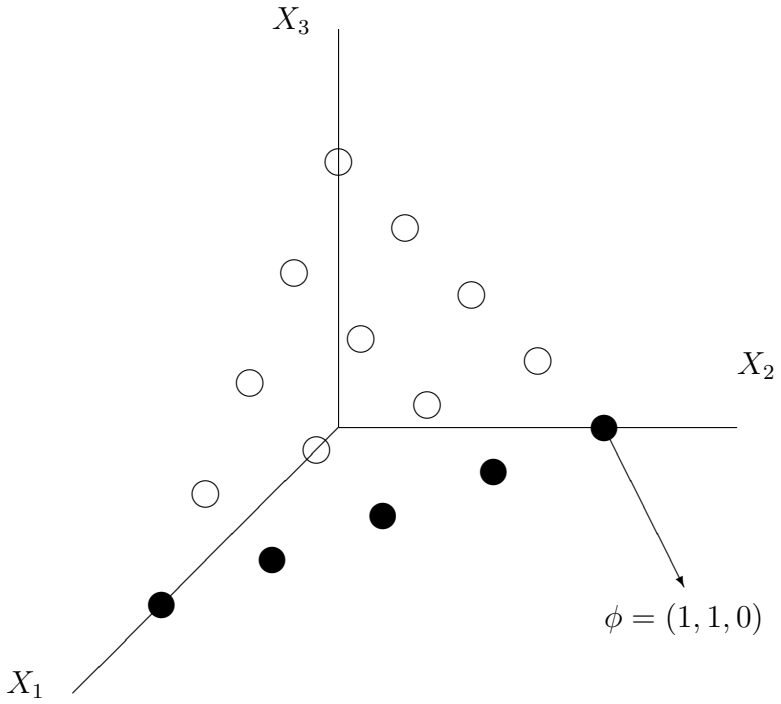


Figure 2.1: Example: The Trinomial Distribution with Sample Size 4. Dots are the points in the natural sample space. The vertices are the points (4,0,0), (0,4,0), and (0,0,4). The observation is (0,4,0). The constraints on the natural parameters are $\theta_1 \geq \theta_2 \geq \theta_3$. A direction of recession is the vector $\phi = (1, 1, 0)$, which is drawn as a normal to the sample space at the observation. The filled-in dots are the support of the maximum likelihood estimate.

where

$$p_i = \frac{e^{\theta_i}}{\sum_{j=1}^{n} e^{\theta_j}}.$$

So this gives the usual form of the multinomial distribution. A point $x \in \mathcal{S}$ is assigned the probability

$$f_\theta(x)\lambda(\{x\}) = \binom{N}{x} \prod_{i=1}^{n} p_i^{x_i}.$$

The convex support K is the convex hull of $\mathcal{S}$. Suppose the natural parameter set is

$$\Theta = \{\, \theta \in I\!R^3 : \theta_1 \geq \theta_2 \geq \theta_3 \,\}$$

and that the observation is $x = (0, 4, 0)$. Then the vector $\phi = (1, 1, 0)$ is a direction of recession because it is normal to K at x and is a direction of recession of Θ. It is not a direction of constancy because it is not normal to the affine hull of $\mathcal{S}$ [$(1, 1, 1)$ being the only direction of constancy]. Hence the maximum likelihood estimate lies in the family conditioned on H_ϕ, that is on the atoms y in $\mathcal{S}$ satisfying $\langle y - x, \phi \rangle = 0$, which are shown in black in the figure.

On iterating the process, no further direction of recession that is not a direction of constancy is found. Such a direction ϕ would be normal to the convex hull of the four black dots in the figure and would satisfy $\langle y - x, \phi \rangle \leq 0$ for $y = (4, 0, 0)$, i. e.

$$\langle (4, 0, 0) - (0, 4, 0), \phi \rangle$$

or $\phi_1 - \phi_2 \leq 0$. But in order to be a direction of recession of Θ, the constraint $\phi_1 \geq \phi_2$ must be satisfied. Hence $\phi_1 = \phi_2$ for any direction of recession ϕ, but any such ϕ is normal to the affine hull of the four black dots, hence a direction of constancy.

The conditional family is recognizable as a binomial distribution with sample size N and natural statistic X_2 (observed to be $N = 4$). The mean value parameter is

$$p_2(\theta) = \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2}}$$

with $\theta_1 \geq \theta_2$, that is $0 < p_2(\theta) < \frac{1}{2}$. The maximum likelihood estimate is $\hat{\theta}_1 = \hat{\theta}_2 = 0$ or $\hat{p}_2 = \frac{1}{2}$.

It is easy to verify that if the original family had been full, the iteration would have continued and the maximum likelihood estimate would have been the distribution concentrated at $(0, 1, 0)$.

2.4 The Phase I Algorithm

Suppose that the natural parameter set Θ is a polyhedral convex set, the set of points satisfying a finite system of linear equalities

$$\langle \theta, b \rangle = \beta(b), \qquad b \in B$$

and inequalities

$$\langle \theta, c \rangle \le \gamma(c), \qquad c \in C$$

where B and C are finite sets of vectors. Then the recession cone of Θ is the set of vectors ϕ satisfying

$$\langle \phi, b \rangle = 0, \qquad b \in B \tag{2.16}$$

and

$$\langle \phi, c \rangle \le 0, \qquad c \in C. \tag{2.17}$$

and is also polyhedral. Whether or not Θ is polyhedral, we assume that its recession cone is polyhedral.

Suppose also that the family is discrete, that is, concentrated on a countable set S of points. This is the case of practical interest.

In order to calculate the MLE, it is necessary to have a routine that, given sets of vectors B and C determining $\operatorname{rc}\Theta$ and given the sample space S, determines a nonzero vector ϕ that is a direction of recession of the constrained log likelihood but not a direction of constancy or determines that no such direction exists.

The vectors in the set

$$T = \{\, y - x : y \in S \,\}$$

are called vectors *tangent* to the convex support K at the observation x. A vector ϕ is a direction of recession of the unconstrained log likelihood if it is normal to T, that is if

$$\langle y - x, \phi \rangle \le 0, \qquad \forall y \in S.$$

It is a direction of recession of the constrained log likelihood if it is also in $\operatorname{rc}\Theta$, that is if (2.16) and (2.17) hold as well.

(The tangent vectors have been called "structure vectors" by Jacobsen [45], who gives a theorem that the MLE exists in a full exponential family if and only if the tangent cone is the whole space. This, as Jacobsen says, follows from Theorem 9.13 in Barndorff-Nielsen [7]).

The set of all nonnegative combinations of the tangent vectors is called the *tangent cone* to K at x. It is not necessary to consider all the tangent vectors. If a convex cone is the set of all nonnegative combinations of some subset, then the subset is said to *generate* the cone. If one can find a subset of the tangent vectors that generates

the tangent cone, then it is only necessary to consider the subset, since any ϕ normal to the subset will be normal to the whole set. In the trinomial example (Figure 2.1), there are fourteen nonzero tangent vectors, one from the observation to each of the other points in the sample space. The tangent cone is generated, however, by just two tangent vectors, the ones lying in the edges of the triangle $(1, -1, 0)$ and $(0, -1, 1)$.

It is often easy to choose such a subset. In a generalized linear model, where the cases are independent, a subset generating the cone consists of vectors $y - x$ where y and x are equal for all except one case, which has been changed to next highest or next lowest value. This reduces the size of the subset to order of the number of cases. Notice that this can reduce an infinite set of tangent vectors to a finite set, thus making possible calculations for some models with denumerable sample spaces (Poisson regression, for example).

Where there is dependence in the model, however, one may be unable to find a subset of the tangent vectors that generates the tangent cone. In this case there may be far too many tangent vectors for available linear programming software to use all of them in constraints. Then one must choose an arbitrary subset, find a ϕ normal to the subset, check to see whether it is normal to all the tangent vectors and, if it is not, increase the subset and try again. Eventually, this must find a normal, if only by increasing the subset to the whole set, but in most cases a normal will be found while employing a small subset of all the tangents.

These considerations give the following algorithm. Finding a direction of recession is essentially a linear programming feasibility problem, with the added twist that 0 is always feasible, and the problem is to find a nonzero feasible vector if one exists. Thus we minimize a linear objective function that is strictly negative on the feasible region, except at zero.

Algorithm R (Finding a Direction of Recession) Given sets of vectors B and C and a sample space S, and a point $x \in S$ (the observation) finds a direction of recession or determines that none exists.

1. Set T to be some subset of the tangent vectors $y - x$, $y \in S$ that generates the whole set.

2. Select a subset T' of T.

3. Solve the linear program

$$\begin{aligned} \text{Minimize:} \quad & \textstyle\sum_{c \in C} \langle \phi, c \rangle \\ \text{Subject to:} \quad & |\phi_i| \leq 1, \quad i = 1, \ \ldots, \ n \\ & \langle \phi, b \rangle = 0, \quad b \in B \\ & \langle \phi, c \rangle \leq 0, \quad c \in C \\ & \langle \phi, y \rangle \leq 0, \quad y \in T' \end{aligned}$$

   producing the solution ϕ.

4. If ϕ is zero, the algorithm terminates with the answer that a direction of recession does not exist. If $T' = T$, the algorithm terminates with the answer that ϕ is a direction of recession.

5. Check whether $\langle \phi, y \rangle \leq 0$, for all $y \in T$. If yes, the algorithm terminates with the answer ϕ is a direction of recession. If no, increase T' adding at least some of the tangent vectors y such that $\langle \phi, y \rangle > 0$. Go to step 3.

Whenever a direction of recession ϕ is found, it becomes a direction of constancy in the next iteration (the problem for the family conditioned on the face of the convex support normal to ϕ). In order to establish the closure condition (2.10), it is necessary to drop inequality constraints determined by vectors c that are not orthogonal to ϕ (if $\langle \phi, c \rangle \neq 0$ then the constraint $\langle \theta + s\phi, c \rangle \leq \gamma(c)$ is violated for some $s \in I\!R$). In order that the algorithm make progress and keep finding directions of recession that are not directions of constancy, it is necessary to add ϕ to the set of equality constraints. This reduces the dimension of the subspace in which ϕ is permitted to lie in each iteration and thus assures termination of the algorithm. This additional constraint does no harm. If ϕ' is a direction of recession of the problem obtained by conditioning on the face normal to ϕ, then so is the projection ϕ'' of ϕ' on the hyperplane normal to ϕ, because the directions of recession form a convex cone [62, p. 69] which contains the whole line along any direction of constancy. Furthermore ϕ' and ϕ'' are normal to the same face of the sample space, since the sample space (of the current problem) is contained in the hyperplane normal to ϕ.

Now the overall phase I algorithm can be can be described as follows.

Algorithm P (Phase I Maximum Likelihood) Given sets of vectors B and C, the sample space S, and the observation x determines the affine subspace A supporting the maximum likelihood estimate.

1. Invoke Algorithm R to find a direction of recession ϕ or determine that none exists. If none exists, terminate with answer $A = \operatorname{aff} S$, otherwise set

$$S = S \cap H_\phi = \{\, y \in S : \langle y - x, \phi \rangle = 0 \,\}.$$

2. (Adjust the constraint set). Remove from C any c such that $\langle \phi, c \rangle \neq 0$. Add ϕ to B. Go to step 1.

2.5 Discussion

Left open by the preceding discussion is the question of whether the maximum likelihood estimate $\hat{f}$ in the closure as constructed by the Phase I–Phase II algorithm is essentially unique. If one takes an arbitrary sequence $\{\theta_n\}$ such that

$$l_x(\theta_n) \to \sup_{\theta \in \Theta} l_x(\theta)$$

does the sequence $\{f_{\theta_n}\}$ necessarily converge pointwise almost everywhere to $\hat{f}$? It turns out that the answer is yes, but a demonstration of this requires techniques that are developed in Chapter 3 and so must wait until Chapter 4.

Note that the properties of being a direction of recession or a direction of constancy depend on the family only through the convex support K and the natural parameter set Θ. There is no dependence on the specific form of the measure λ or its Laplace transform other than through K. Thus one can do the Phase I problem for families for which the functional form of the Laplace transform is unknown. This is very useful in constructing examples and is also important when the Phase II problem is done by Monte Carlo (Chapter 5) because of the intractability of the Laplace transform.

The problem of directions of constancy is usually dealt with by insisting that the natural sufficient statistic have minimal dimension so that no directions of constancy exist. This device, however, does not avoid directions of recession in the phase I

problem. Since conditioning on faces of the convex support produces new directions of constancy, use of this device necessitates a reduction to a minimal representation in each iteration. This necessitates much unnecessary computation and can lead to ill-conditioning of the problem. Moreover, it greatly complicates the algorithms. It is worth doing the little bit of extra theory involved in describing the "almost everywhere" closure to get cleaner algorithms.

The interest of a large majority of statisticians in issues related to the Phase I algorithm (those that have any interest at all) still seems to be in the "existence" question rather that in computation of the support of the maximum likelihood estimate (followed of course by the calculation of the estimate itself in Phase II). It is thus worth pointing out that the "existence" question is settled by the Phase I calculation.

It is hoped, of course, that now that estimates in the closure can be calculated, they will gain some acceptance. It has to be admitted that some of the traditional justification of maximum likelihood estimates does not apply to estimates in the closure. Asymptotic efficiency does not apply (or does so only vacuously) because if the true parameter value is a point in the original family, then the probability of estimates in the closure (not in the original family) goes to zero as the sample size increases. Asymptotically the estimate always exists in the traditional sense. Still the maximum likelihood estimate, whether in the closure or not, represents the hypothesis best supported by the evidence at hand. If the maximum likelihood estimate in the closure is concentrated on a proper subset A of the support of the original family, it can be said that the data provide no evidence whatsoever against the hypothesis that the true distribution is not concentrated on A. Any estimate not concentrated on A is inherently Bayesian (or decision theoretic). It is prior opinion (or the loss function), not the data, that tells one the estimate should not be concentrated on A. Furthermore, if the sample size is large enough to outweigh prior opinion, any reasonable estimate will be "almost" concentrated on A. But that an estimate is "almost" concentrated on A may not be apparent from its parameter value. So whether or not one takes the maximum likelihood estimate in the closure to be one's "official" estimate, it is still of some interest to know its support A. It says something about the data not revealed by other methods.

Chapter 3

GENERALIZED AFFINE FUNCTIONS

Affine functions on a real affine space are the real-valued functions that are are both concave and convex [62, p. 23]. Extended-real-valued functions that are both concave and convex will be of interest in the study of closures of exponential families. They will be called here *generalized* affine functions.

3.1 Compactness of $G(E)$

If E is a real affine space, let $A(E)$ denote the space of all affine functions on E, what is called the *affine dual* of E in Appendix A. Let $G(E)$ denote the space of all generalized affine functions on E, and let $F(E)$ denote the space of all extended-real-valued functions on E. These three spaces will always be given the topology of pointwise convergence. This makes $G(E)$ a topological subspace of $F(E)$, and $A(E)$ a topological subspace of $G(E)$. This topology has a basis consisting of sets of the form

$$\{\, g \in G(E) : g(x_i) \in U_i,\ i = 1, \ldots, n \,\}, \tag{3.1}$$

where $x_1, \ldots, x_n$ are points in E and $U_1, \ldots, U_n$ are open intervals (perhaps semi-infinite intervals) in $\overline{I\!R}$.

Recall that $F(E)$ is the same as the topological product $\overline{I\!R}^E$, and that the topology of pointwise convergence is the same as the product topology. Since $\overline{I\!R}$ is a compact topological space, so is $F(E)$ by Tychonoff's theorem. Furthermore $G(E)$ is also compact.

Theorem 3.1 *The space $G(E)$ of generalized affine functions on a real affine space E is a compact Hausdorff space.*

PROOF. $F(E)$ is a compact Hausdorff space because $\overline{I\!R}$ is compact and Hausdorff and because the product of compact spaces is compact (Tychonoff's theorem) and the product of Hausdorff spaces is Hausdorff. Hence the conclusion follows immediately

if it can be shown that $G(E)$ is closed in $F(E)$, since a closed subset of a compact space is compact and every subspace of a Hausdorff space is Hausdorff.

Let g be any point in the closure of $G(E)$. Then there is a net $\{g_\alpha\}$ in $G(E)$ that converges to g. For any x and y in $\operatorname{dom} g$ and λ in $(0, 1)$, write $z = \lambda x + (1 - \lambda)y$. Then

$$g_\alpha(z) \le \lambda g_\alpha(x) + (1 - \lambda) g_\alpha(y)$$

whenever the right hand side makes sense (is not $\infty - \infty$), which happens eventually, since $g_\alpha(x)$ and $g_\alpha(y)$ both converge to limits that are not $-\infty$. Hence

$$g(z) \le \lambda g(x) + (1 - \lambda) g(y)$$

and g is convex. By symmetry it is also concave and hence is generalized affine. Thus $G(E)$ contains its closure and is closed. □

3.2 The Structure of Generalized Affine Functions

Since a generalized affine function is both convex and concave, all of its level sets are convex—both the level sets in the epigraphical sense and the level sets in the hypographical sense. Thus for any t in $\overline{\mathbb{R}}$ and any generalized affine function g the sets

$$\{\, x \in E : g(x) < t \,\} \qquad \text{and} \qquad \{\, x \in E : g(x) \ge t \,\}$$

are complementary convex sets (and the same is true if the inequalities are reversed). In large part the theory of generalized affine functions is the theory of such sets. Thus a name is needed for them. A subset A of an affine space E is a *half space* of E if A and its complement are both convex sets. In this terminology the preceding remark can be rephrased as: every level set of a generalized affine function is a half space.

Theorem 3.2 (Elementary Structure Theorem). *A function f from an affine space to $\overline{\mathbb{R}}$ is generalized affine if and only if*

(a) *$f^{-1}(+\infty)$ and $f^{-1}(-\infty)$ are both convex sets,*

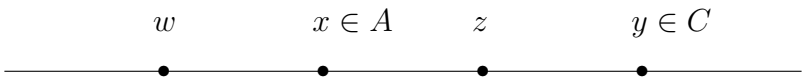
(b) *$A = f^{-1}(\mathbb{R})$ is an affine set, and*

(c) *$f|A$ is an affine function.*

PROOF. First suppose (a), (b), and (c). It suffices to prove that f is convex; that f is concave then follows by symmetry. If $\operatorname{dom} f$ is empty or is a single point, f is obviously convex. Otherwise let x and y be distinct points of $\operatorname{dom} f$, let $z = tx + (1-t)y$ for some t in $(0,1)$, and write $B = f^{-1}(+\infty)$ and $C = f^{-1}(-\infty)$. Clearly the convexity inequality

$$f(z) \leq tf(x) + (1-t)f(y) \tag{3.2}$$

holds whenever $x, y \in C$ (because C is convex) and whenever $x, y \in A$ (because A is affine and f is affine on A). So it remains only to be shown that (3.2) holds when x is in A and y is in C, i. e., that then z is in C. Let w be any point on the line through x and y such that w is on the other side of x from y, as shown in the diagram. It



is obvious from the diagram that $w \in C$ is impossible (then C would not be convex) and $w \in A$ is also impossible (then A would not be affine). Hence $w \in B$. Having established the location of w it is now obvious that $z \in B$ is impossible (then B would not be convex) and that $z \in A$ is impossible (then A would not be affine). Hence $z \in C$. This completes the proof that (a), (b), and (c) imply that f is a generalized affine function.

To see the converse, suppose that f is a generalized affine function. Then (a) is true because B and C are level sets of f. If A is empty or a single point then it is affine. Otherwise let x and y be distinct points of A and let z be any other point on the line determined by x and y. If $z \in (x, y)$ then $z \in A$ because A is convex, being the intersection of $\operatorname{dom}^e h$ and $\operatorname{dom}^h f$ which, being level sets of f, are convex. If $z \notin (x, y)$, suppose without loss of generality that $y \in (x, z)$, and suppose to get a contradiction that $z \in C$. Then $x, z \in \operatorname{dom} f$, and by the convexity inequality $y \in C$, contrary to assumption. Hence $z \notin C$, and $z \notin B$ follows by symmetry. Thus $z \in A$, and (b) is true. Now (c) follows immediately. A real function on an affine space that is both convex and concave is affine [62, p. 23]. □

On a finite-dimensional affine space, the structure of generalized affine functions is simpler still.

Theorem 3.3 (Recursive Structure Theorem). *A generalized affine function f on a finite-dimensional affine space has of one of the following forms.*

(a) *f is affine.*

(b) *$f \equiv -\infty$ or $f \equiv +\infty$.*

(c) *There is an affine function h such that $f(x) = -\infty$ whenever $h(x) < 0$, that $f(x) = +\infty$ whenever $h(x) > 0$, and that the restriction of f to the hyperplane $\{ x : h(x) = 0 \}$ is a generalized affine function.*

PROOF. Let f be a generalized affine function on a finite-dimensional affine space E. As in the preceding theorem, write $A = f^{-1}(\mathbb{R})$, $B = f^{-1}(+\infty)$, and $C = f^{-1}(-\infty)$. The proof now divides into the consideration of three cases, corresponding to the cases of the theorem. Case I is when $B = C = \emptyset$. Then $A = E$, and f is affine, which is (a). Case II is when exactly one of B and C is nonempty. Say for definiteness, that B is empty (the other case follows by symmetry). This implies that $A \neq E$ and hence, A being affine, that the complement of A (that is C) is dense in E. But this implies, C being convex, that $C = E$ and hence $f \equiv -\infty$. Case III is when neither B or C is nonempty. Then, since E is finite-dimensional and B and C are disjoint, nonempty, and convex, there is a hyperplane weakly separating B and C. That is, there is an affine function h such that $h(x) \geq 0$ for $x \in B$ and $h(x) \leq 0$ for $x \in C$ (see, for example, Rockafellar [62, Theorem 11.3] for the relevant separation theorem). As in the argument for case II, the complement of A must be dense in each of the sets $\{ x : h(x) > 0 \}$ and $\{ x : h(x) < 0 \}$ and thus

$$\begin{aligned} f(x) &= +\infty, \qquad h(x) > 0 \\ f(x) &= -\infty, \qquad h(x) < 0 \end{aligned}$$

The theorem now follows from the obvious fact that the restriction of a generalized affine function to an affine subspace (here $\{ x : h(x) = 0 \}$ is a generalized affine function. □

Corollary 3.3.1 *A generalized affine function on a finite-dimensional affine space is Borel measurable.*

PROOF. Affine sets and functions are measurable, since affine functions are continuous and affine sets are closed. Thus it remains only to be shown that the sets $B = f^{-1}(+\infty)$ and $C = f^{-1}(-\infty)$ are measurable. By the theorem B is either empty or the whole space (in either case measurable) or is the union of an open half space (which is measurable) and a subset of the bounding hyperplane which is measurable by induction. □

A similar theorem describes the structure of generalized affine functions on general affine spaces. The proof is similar to the proof of Theorem 2.5 describing the structure of semispaces (half spaces maximal with respect to the property of excluding a specified point) in Klee [49] and will not be given here.

Theorem 3.4 (Klee Representation) *A function f from a real affine space E to $\overline{I\!R}$ is generalized affine if and only if it has a representation of the following form. There exists a totally ordered family $\mathcal{H}$ of affine functions on E, with the notation $g \prec h$ being used to denote that g precedes h in the ordering, having the property that every x in E either lies in the set*

$$A = \{\, x \in E : h(x) = 0,\ \forall h \in \mathcal{H} \,\}$$

or there exists an element h_x of $\mathcal{H}$ such that

$$h_x(x) \neq 0 \quad \text{and} \quad h(x) = 0,\ \forall h \prec h_x,$$

and there exists an affine function g on E, and

$$f(x) = \begin{cases} g(x), & x \in A \\ +\infty, & x \notin A \text{ and } h_x(x) > 0 \\ -\infty, & x \notin A \text{ and } h_x(x) < 0 \end{cases}$$

3.3 Sequential Compactness of $G(E)$

Though $G(E)$ is always compact (so every net has a convergent subnet), it need not be sequentially compact (sequential compactness being the property that every sequence

has a convergent subsequence). If E is finite-dimensional, however, the space $G(E)$ is first countable and hence sequentially compact. This means that all topological arguments can use sequences only, no reference to nets being necessary.

Theorem 3.5 *The space $G(E)$ of generalized affine functions on a finite-dimensional affine space E is first countable.*

PROOF. Let f be a point in $G(E)$. What is to be shown is that there is a countable local base at f, that is a countable family $\mathcal{U}$ of neighborhoods of f such that every neighborhood of f contains an element of $\mathcal{U}$. Since every neighborhood of f contains a basic open neighborhood of the form (3.1) with $f(x_i) \in U_i$, $i = 1, \ldots, n$, it is enough to show that for each x in E and for each open interval W in $\overline{I\!R}$ containing $f(x)$, there is a U in $\mathcal{U}$ such that $g(x) \in W$ for all $g \in U$.

Let D be a countable set in E that is dense in E and in each of the hyperplanes in the representation of f given by Theorem 3.3. Let $\mathcal{U}$ be the collection of all neighborhoods of f of the form (3.1) where the U_i have endpoints that are rational or $\pm\infty$ and the x_i are elements of D. Then $\mathcal{U}$ is countable. We now need to consider the three cases of Theorem 3.3.

In case (a) let $x_i, \ldots, x_n$ be any maximal affinely independent set in D (there exists one since the affine hull of D is all of E). Then x has a representation as an affine combination

$$x = \sum_{i=1}^{n} \lambda_i x_i. \tag{3.3}$$

Let

$$U = \{\, g \in G(E) : r_i < g(x_i) < s_i,\ i = 1, \ldots, n \,\},$$

where r_i and s_i are rational for $i = 1, \ldots, n$ and

$$r_i < f(x_i) < s_i,\ i = 1, \ldots, n.$$

Then U is in $\mathcal{U}$ and is a neighborhood of f, and if $g \in U$

$$|g(x) - f(x)| \le \sum_{i=1}^{n} |\lambda_i||g(x_i) - f(x_i)| \le \sum_{i=1}^{n} |\lambda_i|(s_i - r_i).$$

So for any neighborhood W of $f(x)$ and $s_i - r_i$ chosen small enough, $g(x)$ lies in W for all $g \in U$.

In case (b) let $x_i, \ldots, x_n$ be a finite convex set in D containing x (there exists such a set since the convex hull of D is all of E). Then x has a representation as a convex combination (3.3), where now all of the λ_i are nonnegative. Suppose for definiteness that $f \equiv -\infty$, so that

$$U = \{\, g \in G(E) : g(x_i) < r \,\},$$

where r is rational, is in $\mathcal{U}$ and is a neighborhood of f. Then for $g \in U$

$$g(x) \leq \sum_{i=1}^{n} \lambda_i g(x_i) < r$$

by the convexity inequality. So for any neighborhood W of $f(x)$ and r chosen to be large enough negative number, $g(x)$ lies in W for all $g \in U$.

In case (c) the calculations are the same as in the preceding case. Take $x \in B$ where B is as defined in Theorem 3.3. Then because D is dense in the open convex set B, the convex hull of $D \cap B$ is B and x is a convex combination of points in $D \cap B$, and the conclusion follows as in the preceding case. The case $x \in C$ follows by symmetry. This leaves only the case $x \in A$, which follows by induction on the dimension of the hyperplanes involved in the representation of Theorem 3.3. □

There are two other interesting topological questions that may be asked about the space $G(E)$. Theorem 3.5 implies the sequential compactness of $G(E)$ when E is finite-dimensional. In what cases does $G(E)$ fail to be sequentially compact? If E has an uncountable Hamel basis, then $G(E)$ is not sequentially compact. The proof of this is almost identical to the proof for Counterexample 105 in Steen and Seebach [69] which says that I^I where $I = [0, 1]$ is compact but not sequentially compact. If B is a Hamel basis for E, then the set I^B is homeomorphic to a subset of $G(E)$ consisting of affine functions f such that $0 \leq f(x) \leq 1$ for $x \in B$. This subset has sequences without convergent subsequences, so $G(E)$ has the same property and is not sequentially compact. This rules out the possibility of $G(E)$ being sequentially compact for the most interesting cases where E is an infinite-dimensional topological vector space. If E is an infinite-dimensional F-space (the topology is induced by a complete invariant metric) then E cannot have a countable Hamel basis. This follows from the Baire category theorem (See for example, Exercise 1, p. 52 in Rudin [64]).

The second interesting question concerns the metrizability of $G(E)$ or, more generally, its separation properties. In this context it is useful to eliminate two special cases. If E is empty, then $G(E)$ is a singleton (the empty function). If E is a singleton, say $E = \{x\}$, then $G(E)$ is homeomorphic to $\overline{I\!R}$, a homeomorphism being $f \mapsto f(x)$. Either of these cases is said to be a *trivial* space of generalized affine functions. Other cases, that is when E contains more than one point (hence an uncountable infinity of them), are said to be *nontrivial.* In the trivial cases $G(E)$ is metrizable, since it is a one-point space or homeomorphic to $\overline{I\!R}$.

No nontrivial $G(E)$ is second countable, hence none is metrizable, since any compact metric space is separable, hence second countable. To show this it is necessary only to produce an uncountable discrete set. Let f be any nonconstant affine function, let x and y be any two points such that $f(x) \neq f(y)$, and let $t = y - x$. Define $g \in G(E)$ by

$$g(z) = \begin{cases} +\infty, & f(z) > 0 \\ 0, & f(z) = 0 \\ -\infty, & f(z) < 0 \end{cases}$$

and a family of generalized affine functions

$$S = \{\, g_\lambda : \lambda \in I\!R \,\}$$

by

$$g_\lambda(z) = g(z + \lambda t), \qquad z \in E.$$

Then $g_\lambda(z) = g_\mu(z) = 0$ only if

$$f(z + \lambda t) = f(z + \mu t) = 0$$

which implies

$$f(z) + \lambda\Big(f(y) - f(x)\Big) = f(z) + \mu\Big(f(y) - f(x)\Big)$$

which implies $\lambda = \mu$, because $f(x) \neq f(y)$. So the family S is uncountable and discrete, a neighborhood of g_λ not containing any other element of S being

$$\{\, g \in G(E) : |g(z)| < \epsilon \,\},$$

where $0 < \epsilon < \infty$ and $g_\lambda(z) = 0$, since $|g_\mu(z)| = \infty$ if $\mu \neq \lambda$.

Something stronger is in fact true. No nontrivial space of generalized affine functions is completely normal, though any is normal (being compact and Hausdorff). The proof uses the Baire category theorem and will not be given here.

3.4 Separation and Support Properties

Half spaces of an affine space E were introduced at the beginning of Section 3.2. They are convex sets whose complements are also convex. A special case of complementary half spaces is the pair E, $\emptyset$. These will be referred to as *improper* half spaces. All other half spaces are *proper* half spaces.

The following theorem is well known. Complementary half spaces C and D are said to *exactly separate* sets A and B if $A \subset C$ and $B \subset D$ or if $B \subset C$ and $A \subset D$.

Theorem 3.6 (M. H. Stone) *Let A and B be disjoint convex subsets of an affine space. Then there exist complementary convex sets C and D that exactly separate A and B.*

A proof is given by Klee [50, p. 455]. It requires the axiom of choice. The general idea is that by Zorn's lemma there exists a maximal convex set C containing A but disjoint from B. It then follows from the maximality of C that its complement D convex and contains B.

This can be immediately translated into the language of generalized affine functions as follows. The notion of a half space also applies to $\overline{I\!R}$, the proper half spaces being of the form $[-\infty, x)$ and $[x, +\infty]$ or the reverse $[-\infty, x]$ and $(x, +\infty]$, and the improper half spaces being the pair $\overline{I\!R}$ and $\emptyset$. A generalized affine function is said to *exactly separate* sets A and B if there are complementary half spaces C and D of $\overline{I\!R}$ such that

$$A \subset f^{-1}(C) \quad \text{and} \quad B \subset f^{-1}(D).$$

Corollary 3.6.1 *Let A and B be disjoint convex sets of an affine space E. Then there exists a generalized affine function on E that exactly separates A and B.*

Proof. Apply the theorem to produce complementary half spaces, F and G exactly separating A and B. Define a function h that has the value $-\infty$ on F and $+\infty$ on G. Then h is generalized affine by Theorem 3.2 and exactly separates A and B. □

The following lemma is an obvious analog of Theorem 5.3 in Rockafellar [62].

Lemma 3.7 *Let E be an affine space and A be a half space of $E \times \mathbb{R}$, then the function h defined by*

$$h(x) = \inf\{\, r \in \mathbb{R} : (x, r) \in A \,\}, \qquad x \in E \tag{3.4}$$

is generalized affine.

PROOF. Take any $x, y \in \operatorname{dom}^e h$ and let $z = tx + (1-t)y$ where $0 < t < 1$. Then

$$h(z) \le \inf\{\, tr + (1-t)s : (x, r) \in A \text{ and } (y, s) \in A \,\}$$

by the convexity of A, so h satisfies the convexity inequality. Conversely, take any $x, y \in \operatorname{dom}^h h$ and define z and t as before. Then

$$h(z) \ge \sup\{\, tr + (1-t)s : r < h(x) \text{ and } y < h(s) \,\}$$

by the convexity of the complement of A, so h also satisfies the concavity inequality. □

A function $h : E \to \overline{\mathbb{R}}$ *supports* another function $f : E \to \overline{\mathbb{R}}$ at a point $x \in E$ if $h(x) = f(x)$ and $h \le f$ (meaning $h(z) \le f(z)$ for all $z \in E$).

Theorem 3.8 *Every convex function f on an affine space E has a generalized affine support h at each point x in E. The set of generalized affine supports is a nonempty compact set in $G(E)$.*

PROOF. Any set

$$\{\, g \in G(E) : g(z) \le f(z) \,\}$$

is closed by definition of the topology of pointwise convergence. The same is true if the inequality is reversed. Thus the set of generalized affine supports of f at x

$$\begin{aligned} S &= \{\, g \in G(E) : g \le f \text{ and } f(x) \ge g(x) \,\} \\ &= \{\, g : f(x) \ge g(x) \,\} \cap \left(\bigcap_{z \in E} \{\, g : f(z) \le g(z) \,\} \right) \end{aligned}$$

is the intersection of closed sets, hence closed and compact (since $G(E)$ is compact). It remains to be shown that S is nonempty.

If $f(x) = -\infty$ then $h \equiv -\infty$ is a supporting generalized affine function.

If $f(x) = +\infty$ then $\operatorname{dom} f$ and $\{x\}$ are disjoint convex sets, thus by Theorem 3.6 there are complementary half spaces A and B such that $\operatorname{dom} f \subset A$ and $x \in B$. Define h by

$$h(z) = \begin{cases} -\infty, z \in A \\ +\infty, z \in B \end{cases}$$

Then h is a supporting generalized affine function.

This leaves only the case $x \in f^{-1}(\mathbb{R})$. For each $\epsilon > 0$, the sets $\{(x, f(x) - \epsilon)\}$ and epi f are disjoint convex sets of $E \times \mathbb{R}$. Thus by Theorem 3.6 there are complementary half spaces A and B of $E \times \mathbb{R}$ such that epi $f \subset A$ and $(x, f(x) - \epsilon) \in B$. By Lemma 3.7 the function defined by (3.4) is generalized affine. Let this function be denoted by h_ϵ. Then $h_\epsilon \leq f$ and $h(x) \geq f(x) - \epsilon$. Thus the set

$$S_\epsilon = \{\, g \in G(E) : g \leq f \text{ and } g \geq f(x) - \epsilon \,\}$$

is nonempty. It is also closed (being the intersection of sets defined by inequalities involving single points), hence compact. Thus the intersection

$$S = \bigcap_{\epsilon > 0} S_\epsilon,$$

being the intersection of a nested family of nonempty compact sets, is nonempty. □

The notion of a *face* of a convex set is defined in Section A.11.

Theorem 3.9 *Let C be a convex set in an affine space E. A set D in E is a face of C if and only if it is the locus of points where some generalized affine function h achieves its supremum over C and the supremum is finite, that is if there is an $r \in \mathbb{R}$ such that $D = h^{-1}(r)$ and $h(x) \leq r$ for all $x \in C$.*

Proof. Suppose that h satisfies the condition of the theorem. Choose $x, y \in C$ such that $z = tx + (1 - t)y$ is in D for some t with $0 < t < 1$. Then $h(x) < r$ or $h(y) < r$ would imply $h(z) < r$ by the convexity of h, which contradicts $z \in D$. Thus $h(x) = h(y) = r$ and $x, y \in D$. This shows that D is a face of C.

Conversely, suppose that D is a face of C. Then aff D and $C \setminus D$ are disjoint (Theorem A.1). Let

$$B = \operatorname{con}(C \cup \operatorname{aff} D)$$

and define a function f by

$$f(z) = \begin{cases} -\infty, & z \in B \setminus \operatorname{aff} D \\ 0, & z \in \operatorname{aff} D \\ +\infty, & z \notin B \end{cases}$$

Then f is convex. Let x be any point of aff D and h be a generalized affine support of f at x. Then $h(z) = 0$ for all $z \in \operatorname{aff} D$ and $h(z) = -\infty$ for all $z \in C \setminus D$. Then D is the locus of points where h achieves its supremum over C, because $C \cap \operatorname{aff} D = D$. □

The three preceding theorems show how generalized affine functions capture the local properties of convex sets and functions. It is interesting to compare these theorems with classical results in which generalized affine functions are replaced by affine functions. In each case, a weaker statement results, even in the finite-dimensional case.

Even in $\mathbb{R}^2$ it is not the case that any pair of disjoint convex sets A and B can be exactly separated by an affine function. What is true in the finite-dimensional case is that the sets can be *properly* separated [62, Theorem 11.3]: there is an affine function h such that $h(x) \leq 0$ for $x \in A$ and $h(x) \geq 0$ for $x \in B$ and $A \cup B$ is not contained in the hyperplane $h^{-1}(0)$. Even less is true in the infinite-dimensional case. Then proper separation requires the additional hypothesis that one of the sets, say A, have a nonempty core [50, Theorem 8.10] or, what is equivalent, if there is some point $x \in A$ such that $A - x$ is an absorbing set [64, Exercise 3, p. 81].

An example in $\mathbb{R}^2$ of disjoint convex sets that that cannot be exactly separated by an affine function are the half space

$$\{\, (x, y) : x > 0 \text{ or } x = 0 \text{ and } y \geq 0 \,\}$$

and its complement. Any properly separating affine function is of the form

$$h\colon (x, y) \mapsto \alpha x$$

for some $\alpha \neq 0$. But no such h exactly separates A and its complement, since $h^{-1}(0)$ (the y-axis) contains points of both half spaces.

In classical convexity theory the gradient of an affine function that supports a convex function f at a point x of its effective domain is called a *subgradient* of f at

x. In the finite-dimensional case, every convex function f has a subgradient (hence has a supporting affine function) at each point of the relative interior of its effective domain [62, Theorem 23.4]. Failure to have a supporting affine function can occur even in $\mathbb{R}^1$. Consider the function f defined by

$$f(z) = \begin{cases} +\infty, & z < 0 \\ -\sqrt{z} & z \geq 0 \end{cases}$$

which has an infinite one-sided derivative at zero and hence no supporting affine function there.

In classical theory the analogue of Theorem 3.9 is not a theorem but a definition. If D is the locus of points where some affine function h achieves its supremum over a convex set C, then D is a face of C but not all faces can be constructed this way. Such a face is called an *exposed* face. Convex sets with nonexposed faces can occur in $\mathbb{R}^2$ [62, p. 163].

Chapter 4

CLOSURES OF EXPONENTIAL FAMILIES AND MAXIMUM LIKELIHOOD

In this chapter the notion of the closure of an exponential family in the topology of pointwise convergence of densities is developed. The notion of a closure of an exponential family in the topology of convergence in distribution has been studied by others [6, 7, 18]. Using convergence in distribution rather than pointwise convergence of densities appears at first sight to yield a more general theory, since pointwise convergence of densities implies convergence in total variation (hence in distribution) by Scheffe's theorem. This appearance is deceiving, however, since the closure in the topology of convergence in distribution has been derived only for families with finite support [7] and for some families with countable support [18]. When the support of a family is countable and topologically discrete, the topology of pointwise convergence of densities and the topology of convergence in distribution are the same. Hence no more generality has actually been obtained by studying the weaker topology (of convergence in distribution). In fact, less, the technical difficulties associated with the topology of convergence in distribution have been a barrier to progress.

4.1 Generalized Exponential Families

Suppose $\mathcal{F}$ is a standard exponential family of densities with respect to a measure λ on a finite-dimensional real affine space E, and suppose that $\{f_\alpha\}$ is a net in $\mathcal{F}$ that converges almost everywhere to a density f. That is, there is a measurable set A such that $\lambda(A^c) = 0$ and

$$f_\alpha(x) \to f(x), \qquad x \in A.$$

The value of $f(x)$ for x not in A is irrelevant and need not be defined. Let $h_\alpha = \log f_\alpha$. Then, since the space $G(E)$ of all generalized affine functions on E is compact, there is a subnet $\{h_{\alpha(\beta)}\}$ that converges (pointwise) to some generalized affine function h.

Since

$$h_{\alpha(\beta)}(x) \to h(x) = \log f(x), \qquad x \in A$$

and f may be arbitrarily defined on the complement of A, we may take $f = e^h$. Then $f_\alpha(x) \to f(x)$ for all x in A, which is for almost all x. This argument shows that there is no loss of generality in going from the topology of pointwise almost everywhere convergence of densities (on the space of all measurable functions) to the topology of pointwise convergence in $G(E)$. Every possible limit is equal almost everywhere to a generalized affine function. Furthermore, since the topology of pointwise convergence in $G(E)$ is first countable, it is enough to consider only limits of sequences.

This leads to the notion of a generalized exponential family. A family of probability densities with respect to a Borel measure λ on a finite-dimensional real affine space E is a *standard generalized exponential family* if it is of the form

$$\mathcal{F} = \exp \mathcal{H} = \{\, e^h : h \in \mathcal{H} \,\} \tag{4.1}$$

for some nonempty subset $\mathcal{H}$ of $G(E)$. Note that there is no explicit restriction to measurable members of $\mathcal{H}$ since by Corollary 3.3.1 every element of $G(E)$ is measurable. The family $\mathcal{F}$ is said to be *full* if

$$\mathcal{H} = \{\, h \in G(E) : \int e^h \, d\lambda = 1 \,\}.$$

Let $\operatorname{cl} \mathcal{H}$ denote the closure of $\mathcal{H}$ in $G(E)$ (defined in Section 3.1) and $\text{r-cl}\, \mathcal{H}$ the closure relative to the full family, that is

$$\text{r-cl}\, \mathcal{H} = \{\, h \in \operatorname{cl} \mathcal{H} : \int e^h \, d\lambda = 1 \,\}.$$

The corresponding definitions for the families of densities are

$$\operatorname{cl} \mathcal{F} = \{\, e^h : h \in \operatorname{cl} \mathcal{H} \,\}$$

and

$$\text{r-cl}\, \mathcal{F} = \{\, f \in \operatorname{cl} \mathcal{F} : \int f \, d\lambda = 1 \,\}.$$

By the continuity of the exponential function $\operatorname{cl} \mathcal{F}$ is the closure of $\mathcal{F}$ in the topology of pointwise convergence, and $\text{r-cl}\, \mathcal{F}$ is the closure relative to the full family.

By the first countability of $G(E)$ every element f in $\operatorname{cl}\mathcal{F}$ is the limit of a sequence $\{f_n\}$ in $\mathcal{F}$. Clearly $f(x) \geq 0$ for all x, and by Fatou's lemma

$$\int f\,d\lambda = \int \left(\liminf_{n\to\infty} f_n\right) d\lambda \leq \liminf_{n\to\infty} \int f_n\,d\lambda = 1.$$

Hence every f in $\operatorname{cl}\mathcal{F}$ is a *subprobability density,* that is, a nonnegative measurable function whose integral is less than or equal to one.

4.2 Maximum Likelihood

Let $m = \sup\mathcal{H}$, that is

$$m(x) = \sup\{\, h(x) : h \in \mathcal{H}\,\}, \qquad x \in E \tag{4.2}$$

be the log likelihood supremum function for the family $\mathcal{F}$. Given an observation x, the maximum likelihood estimate $\hat{h}_x$ is the set of log densities maximizing the log likelihood at x

$$\hat{h}_x = \{\, h \in \mathcal{H} : h(x) = m(x)\,\},$$

or, what is equivalent, the set

$$\hat{f}_x = \exp\hat{h}_x = \{\, f \in \mathcal{F} : f(x) = e^{m(x)}\,\}.$$

The maximum likelihood estimate may be empty: the supremum of the likelihood need not be achieved. If, however, we seek the maximum likelihood in the closure of the family, the supremum is always achieved.

Theorem 4.1 *The maximum likelihood estimate*

$$\hat{h}_x = \{\, h \in \operatorname{cl}\mathcal{H} : h(x) = m(x)\,\},$$

in the closure of a generalized exponential family is always nonempty.

Proof. There always exists a sequence $\{h_n\}$ in $\mathcal{H}$ such that $h_n(x) \to m(x)$ as $n \to \infty$. By the sequential compactness of $G(E)$ this sequence has a subsequence converging to a limit h which is necessarily in $\operatorname{cl}\mathcal{H}$ and which satisfies $h(x) = m(x)$. This h is in $\hat{h}_x$. □

The theorem is trivial, since the definition of closure has been designed to remove all difficulties in the proof and generalized affine functions were invented just for this application. Note that any element h of the maximum likelihood estimate $\hat{h}_x$ supports the log likelihood supremum function m at the observation x, that is $h \le m$ (by definition since $m = \sup \mathcal{H}$) and $h(x) = m(x)$. There need not always be a supporting affine function, but by Theorem 3.8 there is always a supporting generalized affine function. The failure of the maximum likelihood estimate to exist in the traditional sense is precisely the failure of the convex function m to be subdifferentiable (have a supporting affine function) at x. A simple example shows this.

Example 4.1 (The Binomial Distribution) Let λ be the measure in $\mathbb{R}^1$ with atoms of mass $\binom{N}{x}$ at the points $x = 0, 1, \ldots, N$. Then the Laplace transform is

$$c(\theta) = \sum_{x=0}^{N} \binom{N}{x} e^{x\theta} = (1 + e^{\theta})^N$$

so

$$f_\theta(x) = \frac{e^{x\theta}}{(1 + e^{\theta})^N} = p(\theta)^x (1 - p(\theta))^{N-x}$$

where

$$p(\theta) = \frac{e^{\theta}}{1 + e^{\theta}}.$$

so

$$F_\theta(\{x\}) = f_\theta(x) \lambda(\{x\}) = \binom{N}{x} p(\theta)^x (1 - p(\theta))^{N-x}$$

the usual form of the binomial distribution.

The log likelihood is

$$l_x(\theta) = x\theta - N \log(1 + e^{\theta}) = x \log p(\theta) + (N - x) \log(1 - p(\theta))$$

has derivative

$$\nabla l_x(\theta) = x - N p(\theta)$$

and so the maximum likelihood estimate satisfies

$$p(\hat{\theta}) = x/N$$

for x such that this equation has a solution for $\hat{\theta}$, that is

$$\hat{\theta} = \log \frac{x}{N - x} \tag{4.3}$$

for $0 < x < N$. The log likelihood supremum function is thus

$$m(x) = x \log \frac{x}{N} + (N - x) \log N - xN$$

for such x. Anticipating the results of Theorem 4.2, m is a lower semicontinuous convex function with $\operatorname{cl}(\operatorname{dom} m) = [0, N]$, the convex support of λ. Thus the values for $x = 0$, N are found by lower semicontinuity to be zero:

$$m(x) = \begin{cases} x \log \frac{x}{N} + (N - x) \log N - xN, & 0 \le x \le N \\ +\infty, & \text{otherwise} \end{cases}$$

By direct calculation the gradient of m is found to be given by (4.3) for $0 < x < N$. So the affine function

$$y \mapsto m(x) + \langle y - x, \hat{\theta} \rangle$$

associated with $\hat{\theta}$ does support m at x. But there is no affine function supporting m at 0 or N. There must be (by the theorem) supporting generalized affine functions, and these are obvious

$$h(x) = \begin{cases} +\infty, & x < 0 \\ 0, & x = 0 \\ -\infty, & x > 0 \end{cases}$$

supports m at 0, and

$$h(x) = \begin{cases} +\infty, & x > N \\ 0, & x = N \\ -\infty, & x < N \end{cases}$$

supports m at N. The distributions associated with these densities are concentrated at 0 and N in accord with the elementary notion of a maximum likelihood estimate being defined by $\hat{p} = x/N$, $0 \le x \le N$.

The behavior of the maximum likelihood estimate in the relative closure is also of interest. Is there a similar theorem for the relative closure, or, what is equivalent, is the maximum likelihood estimate in the closure always a probability density? The answer is no, as the following examples show.

Example 4.2 Let λ be Lebesgue measure on the negative quadrant in $\mathbb{R}^2$. Then the Laplace transform of λ is

$$c(\theta) = \int_{-\infty}^{0} \int_{-\infty}^{0} e^{x_1\theta_1 + x_2\theta_2} \, dx_1 \, dx_2 = \begin{cases} \frac{1}{\theta_1\theta_2}, & \theta_1 > 0, \ \theta_2 > 0 \\ +\infty, & \text{otherwise} \end{cases}$$

The densities of the full family have the form

$$f_\theta(x) = \theta_1\theta_2 e^{x_1\theta_1 + x_2\theta_2}, \qquad \theta \in \operatorname{dom} c. \tag{4.4}$$

Let us investigate the nonfull family with natural parameter set

$$\Theta = \{ \, \theta \in \operatorname{dom} c : \theta_1\theta_2 = 1 \, \}. \tag{4.5}$$

for which the formula for the densities (4.4) reduces to

$$f_\theta(x) = e^{x_1\theta_1 + x_2\theta_2}, \qquad \theta \in \Theta.$$

The closure of this family obviously contains only two limit points, one obtained by letting $\theta_1 = 1/\theta_2$ go to $+\infty$ and one by letting it go to 0. The resulting limit points have the form

$$f_i(x) = \begin{cases} +\infty, & x_i > 0 \\ 1, & x_i = 0 \\ 0, & x_i < 0 \end{cases}$$

for $i = 1, 2$. Note that neither is a probability density, in fact $\int f_i \, d\lambda = 0$.

These limits are maximum likelihood estimates for observations on the boundary of the sample space, i. e. points (x_1, x_2) such that $x_1 = 0$ or $x_2 = 0$. Such points, however, occur with probability zero. A modification of this example produces elements of the maximum likelihood estimate that are subprobability densities and that occur with positive probability.

Example 4.3 Now let λ be the measure of the last example with the addition of an atom of mass one at the origin. Then

$$c(\theta) = 1 + \int_{-\infty}^{0} \int_{-\infty}^{0} e^{x_1\theta_1 + x_2\theta_2} \, dx_1 \, dx_2 = \begin{cases} 1 + \frac{1}{\theta_1\theta_2}, & \theta_1 > 0, \ \theta_2 > 0 \\ +\infty, & \text{otherwise} \end{cases}$$

We keep the same parameter set (4.5) and the densities are now

$$f_\theta(x) = \frac{1}{2} e^{x_1\theta_1 + x_2\theta_2}, \qquad \theta \in \Theta$$

and the limit points are

$$f_i(x) = \begin{cases} +\infty, & x_i > 0 \\ \frac{1}{2}, & x_i = 0 \\ 0, & x_i < 0 \end{cases}$$

Now $\int f_i \, d\lambda = \frac{1}{2}$ since $f_i(0) = \frac{1}{2}$. If $x = 0$ is the observation, then the whole closure of the family is the estimate $\hat{f}_x$ since $f_\theta(0) = \frac{1}{2}$ for all θ in Θ.

Further modification makes one of the limit points the unique maximum likelihood estimate.

Example 4.4 Let λ be the measure of the last example with the addition of another atom at $(0, -1)$. Then

$$c(\theta) = 1 + e^{-\theta_2} + \int_{-\infty}^{0} \int_{-\infty}^{0} e^{x_1\theta_1 + x_2\theta_2} \, dx_1 \, dx_2 = \begin{cases} 1 + e^{-\theta_2} + \frac{1}{\theta_1\theta_2}, & \theta_1 > 0, \ \theta_2 > 0 \\ +\infty, & \text{otherwise} \end{cases}$$

So now, still keeping the same parameter set (4.5), the densities have the form

$$f_\theta(x) = \frac{1}{2 + e^{-\theta_2}} e^{x_1\theta_1 + x_2\theta_2}, \qquad \theta \in \Theta \tag{4.6}$$

Taking the limit as $\theta_1 = 1/\theta_2 \to \infty$ gives

$$f_1(x) = \begin{cases} +\infty, & x_1 > 0 \\ \frac{1}{3}, & x_1 = 0 \\ 0, & x_1 < 0 \end{cases} \tag{4.7}$$

and taking the limit as $\theta_2 = 1/\theta_1 \to \infty$ gives

$$f_2(x) = \begin{cases} +\infty, & x_2 > 0 \\ \frac{1}{2}, & x_2 = 0 \\ 0, & x_2 < 0 \end{cases} \tag{4.8}$$

Now if $x = 0$ is the observation, f_2 is the unique maximum likelihood estimate (or, to precisely follow our definition of the estimate as a set, we have $\hat{f}_x = \{f_2\}$).

Another modification gets the same behavior in a discrete family.

Example 4.5 Let λ be counting measure on the set in $\mathbb{R}^2$

$$S = \{(0,0),(0,-1)\} \cup \{\, (-k,-l) : k = 1,2,\ldots,\ l = 1,2,\ldots \}.$$

Then the Laplace transform of λ is

$$c(\theta) = 1 + e^{-\theta_2} + \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} e^{k\theta_1 + l\theta_2} = \begin{cases} 1 + e^{-\theta_2} + \frac{e^{-\theta_1}}{1-e^{-\theta_1}}\frac{e^{-\theta_2}}{1-e^{-\theta_2}} & \theta_1 > 0,\ \theta_2 > 0 \\ +\infty, & \text{otherwise} \end{cases}$$

We now need to adjust the parameter set so that the third term in the Laplace transform is constant, say

$$\Theta = \{\, \theta \in \mathbb{R}^2 : \theta_1 > 0,\ \theta_2 > 0,\ (e^{\theta_1} - 1)(e^{\theta_2} - 1) = 1 \,\}$$

Then for $\theta \in \Theta$ the densities again have the form (4.6), the closure contains the two additional limit points (4.7) and (4.8), and the unique maximum likelihood estimate for the observation $x = 0$ is the latter, f_2.

4.3 The Structure of Generalized Exponential Families

The log likelihood supremum function (4.2) of a generalized exponential family, being the pointwise supremum of convex functions, is convex. Hence its effective domain

$$M = \operatorname{dom} m = \{\, x \in E : \sup_{h \in \mathcal{H}} h(x) < +\infty \,\},$$

the set of points where the supremum of the log likelihood is not $+\infty$, is a convex set.

More can be said if the generalized exponential family is the closure of an exponential family. For any convex function g on a vector space E, the *conjugate* of g, also called the *Legendre-Fenchel transform* of g is the function g^* from E^* to $\overline{\mathbb{R}}$ defined by

$$g^*(\theta) = \sup\{\, \langle x, \theta \rangle - g(x) : x \in E \,\}$$

and similarly if $E^{**} = E$, as is the case when E is finite-dimensional, and g is a convex function on E^*

$$g^*(x) = \sup\{\, \langle x, \theta \rangle - g(\theta) : \theta \in E^* \,\}$$

([63, Definition 2E.1] or [62, pp. 104] which differ slightly, though not when g is proper, the former definition being the one given here).

When this is applied to exponential families, we see that the log likelihood supremum function of a full exponential family is the Legendre-Fenchel transform of the log Laplace transform.

$$\begin{aligned} m(x) &= \sup_{\theta \in \operatorname{dom} c} \log f_\theta(x) \\ &= \sup_{\theta \in E^*} \langle x, \theta \rangle - \log c(\theta) \qquad (4.9) \\ &= (\log c)^*(x) \end{aligned}$$

(4.10)

the change from taking the sup over $\operatorname{dom} c$ to taking the sup over all of E^* being valid because $\langle x, \theta \rangle - \log c(\theta) = -\infty$ for θ in $E^* \setminus \operatorname{dom} c$.

From this and Theorem 2.3 a partial characterization of the effective domain of m is immediate.

Theorem 4.2 *The log likelihood supremum function m of an exponential family is a lower semicontinuous proper convex function. The closure of the effective domain M of m supports the family. If the family is full, then* $\operatorname{cl} M$ *is the convex support K.*

PROOF. Let m_f denote the log likelihood supremum function of the full family. Then since $\log f_\theta \le m \le m_f$ and m_f, being the conjugate of $\log c$, is proper, m is proper as well. Since m is proper and the pointwise supremum of lower semicontinuous convex functions, it is lower semicontinuous [62, Theorem 9.4]. For any closed convex function g the recession function of g is the support function of $\operatorname{dom} g^*$ [62, Theorem 13.3]. Applied to $g = \log c$ this gives that the support function of $M_f = \operatorname{dom} m_f = (\log c)^*$ is the support function of K by Theorem 2.3. Two convex sets with the same support function have the same closure [62, Corollary 13.1.1]. Hence $K = \operatorname{cl} M_f$. That $\operatorname{cl} M$ supports λ follows from $m \le m_f$ so $M_f \subset M$. □

For a proof that uses less convex analysis (only the separating hyperplane theorem) see Barndorff-Nielsen [7, Theorem 9.1].

Theorem 4.3 (Structure Theorem for Generalized Exponential Families) *Let $\mathcal{H}$ be the family of log densities of a generalized exponential family with respect to a measure*

λ, and let $m = \sup \mathcal{H}$ be the log likelihood supremum function and $M = \operatorname{dom} m$. For any $h \in \mathcal{H}$ the set $D = M \cap h^{-1}(\mathbb{R})$ is a face of M and $\operatorname{cl} D$ supports the distribution associated with h.

PROOF. Fix $h \in \mathcal{H}$. Let $A = h^{-1}(\mathbb{R})$ and $D = M \cap A$. The function that is zero where h is finite, $+\infty$ where h is $+\infty$, and $-\infty$ where h is $-\infty$ is generalized affine (Theorem 3.2) and achieves its maximum over M on D. Thus D is a face of M (Theorem 3.9).

Let K_A denote the convex support of $\lambda|A$, the restriction of λ to A (the smallest closed convex set whose complement in A has λ measure zero). Suppose $x \in \text{r-int}\, K_A$ ($\text{r-int}\, K_A$ is nonempty since K_A is), and suppose to get a contradiction that $x \notin M$. Then there is a sequence $\{h_n\}$ in $\mathcal{H}$ such that $h_n(x) \to +\infty$, and there is a subsequence that converges to a limit h. Then $h(x) = +\infty$ and (by Fatou's lemma) $\int e^h \, d\lambda \le 1$. Hence the half space $B = h^{-1}(+\infty)$, which contains x, has λ-measure zero. But this contradicts the assumption $x \in \text{r-int}\, K$, i. e. that there is a neighborhood of x in $\operatorname{aff} K_A$, every point of which has a neighborhood of nonzero measure. Every neighborhood of x in $\operatorname{aff} K_A$ contains a point in the interior of B, and each such point has a neighborhood contained in B, which thus has measure zero. This contradiction shows that $\text{r-int}\, K_A \subset M$, hence, because convex sets with the same relative interior have the same closure [62, Theorem 6.3],

$$\operatorname{cl}(\text{r-int}\, K_A) = \operatorname{cl} K_A = K_A \subset \operatorname{cl}(M \cap \operatorname{aff} K_A) \subset \operatorname{cl}(M \cap A) = \operatorname{cl} D$$

so

$$\int e^h \, d\lambda = \int_{\operatorname{cl} D} e^h \, d\lambda$$

as was to be proved. □

This theorem says, in effect, that every generalized exponential family "is" a union of exponential families indexed by faces of M. Let $\mathcal{D}$ be the family of nonempty faces of M (note that this includes M itself). Let $\mathcal{H}$ be the set of log densities of the family, and let

$$\mathcal{H}_D = \{\, h \in \mathcal{H} : M \cap h^{-1}(\mathbb{R}) = D \,\}.$$

Then each $\mathcal{H}_D$ "is" an exponential family on $\operatorname{aff} D$. More precisely the restriction of $\mathcal{H}_D$ to $\operatorname{aff} D$ is an exponential family of log densities with respect to the measure

$\lambda|\operatorname{aff} D$, which is the restriction of λ to $\operatorname{aff} D$ defined by

$$(\lambda|\operatorname{aff} D)(B) = \lambda(B \cap \operatorname{aff} D), \qquad B \in \operatorname{aff} D,$$

since each $h \in \mathcal{H}_D$ is concentrated on $\operatorname{cl} D \subset \operatorname{aff} D$. The distributions induced by h and $h|\operatorname{aff} D$ are "the same" in the sense that

$$\int_B e^h \, d\lambda = \int_{B \cap \operatorname{aff} D} e^h \, d\lambda, \qquad B \in \mathcal{B}.$$

The theorem is a completely general version of the construction in Barndorff-Nielsen [7, pp. 154-155] and in Brown [18, pp. 191–194] in which the set K is used instead of M. Barndorff-Nielsen shows that for an exponential family whose support is a finite set of points that the completion of the exponential family "is" a union of exponential families obtained by conditioning the original family on the faces of the convex support K. This is a special case of Theorem 4.3 since when the family has finite support and is full, $M = K$ and each face of M is closed (being the convex hull of a finite set of points). Except in the case of finite support the name "completion" is inappropriate, since the relative closure is not in general complete, and the closure does not consist solely of probability densities. Brown calls the same object defined by Barndorff-Nielsen an "aggregate exponential family" dispensing with its topological origin, and demonstrates that when enough regularity conditions are assumed one can still get Barndorff-Nielsen's result in some cases where the family has countable support. There is, however, no reason to adopt these regularity conditions. The proper object of study in the general case is the set M not K. When M is substituted for K, a completely general result without unnecessary regularity conditions is obtained.

Example 4.6 Let λ be a measure on $\mathbb{R}^2$ concentrated on the atoms $(0, 0)$, $(0, 1)$ and $(1, k)$, $k = 0, \pm 1, \pm 2, \ldots$, any measure that has a Laplace transform that is finite everywhere. Let $p_k = \lambda(\{(1, k)\})$. Then finiteness of the Laplace transform requires

$$\sum_{k=-\infty}^{+\infty} p_k e^{\theta k}$$

be finite for all real θ. For example, the sequence $p_k = e^{-k^2}$ will do.

Consider maximum likelihood in the full family when the observation is $(0, 0)$. Then $\phi = (-1, 0)$ is a direction of recession that is not a direction of constancy. So the support of the maximum likelihood estimate is in the hyperplane

$$H_\phi = \{\, x \in \mathbb{R}^2 : x_1 = 0 \,\}.$$

The family conditioned on H_ϕ is a Bernoulli family. And the distribution concentrated at the observation $(0, 0)$ is the obvious maximum likelihood estimate.

The point of the example is that the convex support K of the original family is the whole strip

$$K = \{\, x \in \mathbb{R}^2 : 0 \le x_1 \le 1 \,\}.$$

But the support A of the maximum likelihood estimate, which is the one-point affine space $\{(0, 0)\}$ is not a face of K. It is, by Theorem 4.3, a face of M, which is

$$M = \{\, x \in \mathbb{R}^2 : 0 < x_1 \le 1 \text{ or } x_1 = 0 \text{ and } 0 \le x_2 \le 1 \,\}.$$

Thus even for full discrete families in $\mathbb{R}^2$ a theory of closures based on the convex support K is unworkable, even if "discrete" means the atoms are topologically discrete (have no points of accumulation) rather than just supported on a countable set. This example improves an example given by Brown [18, Exercise 6.18.1] which shows the same phenomenon for a discrete but not topologically discrete family in $\mathbb{R}^2$ and for a topologically discrete family in $\mathbb{R}^3$.

4.4 The Relative Closure of a Convex Exponential Family

Theorem 4.1 is somewhat unsatisfactory, despite the fact that it is in general the best result obtainable, as the examples which follow it show. From the construction in Chapter 2, we know that the situation is better in the closure of a convex exponential family. Then the Phase I–Phase II algorithm constructs a maximum likelihood estimate in the relative closure whenever the supremum of the likelihood is finite. This is perhaps not clear from the description in Chapter 2, which does not mention generalized affine functions, but (2.14) can be replaced by $f_{\theta+s\phi} \to f_{\theta,\phi}$ pointwise,

where $f_{\theta,\phi}$ is defined by

$$f_{\theta,\phi}(x) = \begin{cases} +\infty, & \langle x, \phi \rangle > \sigma_K(\phi) \\ f_\theta(x|H_\phi), & \langle x, \phi \rangle = \sigma_K(\phi) \\ 0, & \langle x, \phi \rangle < \sigma_K(\phi) \end{cases}$$

which shows that $\log f_{\theta+s\phi}$ converges to a generalized affine function $\log f_{\theta,\phi}$. Similarly, in the next iteration of the Phase I algorithm, $\log f_{\theta+s\phi_2,\phi}$ converges to another generalized affine function, and so forth.

The question of whether the limit thus constructed is essentially unique, which was raised at the end of Chapter 2, is answered affirmatively in the following.

Lemma 4.4 *For a convex exponential family with natural parameter set Θ that is closed relative to the effective domain of the Laplace transform c, the support function of the effective domain M of the log likelihood supremum function is $\sigma_M = \sigma_K + \delta_{\operatorname{rc}\Theta}$, where K is the convex support and Θ the natural parameter set.*

PROOF. In the constrained case (4.9) becomes

$$\begin{aligned} m(x) &= \sup_{\theta \in \Theta} \log f_\theta(x) \\ &= \sup_{\theta \in E^*} \langle x, \theta \rangle - \log c(\theta) - \delta_\Theta(\theta) \\ &= (\log c + \delta_\Theta)^*(x) \end{aligned}$$

Since $\log c + \delta_\Theta$ is a lower semicontinuous proper convex function (see page 17), the support function of $M = \operatorname{dom} m$ is the recession function of $\log c + \delta_\Theta$ [62, Theorem 13.3], which is calculated in Theorem 2.3 to be $\sigma_K + \delta_{\operatorname{rc}\Theta}$. □

Theorem 4.5 *The maximum likelihood estimate in the relative closure of a convex exponential family is nonempty if the supremum of the log likelihood is finite. Let D be the smallest face of the effective domain M of the log likelihood supremum function containing the observation x, then the estimate is concentrated on* $\operatorname{cl} D$. *The maximum likelihood estimate is essentially unique in the sense that distinct estimates are equal almost everywhere.*

PROOF. The first sentence just repeats Theorem 2.7. Suppose h is an element of the maximum likelihood estimate in the relative closure. If h is not actually affine, then there is a hyperplane properly separating $h^{-1}(-\infty)$ and $h^{-1}(+\infty)$, i. e. there is a nonzero affine function g such that

$$\begin{aligned} h(x) &= -\infty, \qquad & g(x) &< 0 \\ h(x) &= +\infty, \qquad & g(x) &> 0 \end{aligned}$$

Since the set $h^{-1}(+\infty)$ lies outside of M, the normal $\phi = \nabla g$ to the bounding hyperplane $g^{-1}(0)$ satisfies

$$\sigma_M(\phi) \le \langle x, \phi \rangle.$$

By the lemma, this occurs if and only if

$$\sigma_K(\phi) + \delta_{\operatorname{rc}\Theta}(\phi) \le \langle x, \phi \rangle,$$

which occurs if and only if $\sigma_K(\phi) \le \langle x, \phi \rangle$ (that is if ϕ is a direction of recession of the unconstrained log likelihood) and $\phi \in \operatorname{rc}\Theta$. Thus taking the limit of the family in the direction ϕ in the sense of Theorem 2.6 we see that h "is" (up to almost everywhere equivalence) a density in the closure of the family conditioned on the hyperplane $H_\phi = g^{-1}(0)$. If the whole problem is restricted to this hyperplane, h is still generalized affine and if not actually affine has a hyperplane separating $h^{-1}(-\infty)$ and $h^{-1}(+\infty)$ from which it follows by the same argument that h "is" in the closure of the exponential family conditioned on this second hyperplane, and so forth. That is, by induction, h "is" in the exponential family obtained by conditioning the original family on $A_1 = h^{-1}(\mathbb{R})$. The log likelihood supremum of this family is just $m|A_1$, because the conditioning increases the likelihood.

Let $D_1 = A_1 \cap M$, let K_1 be the convex support of $\lambda|A_1$, and let Θ_1 be the natural parameter set of the family conditioned on A_1 (the set of gradients of the log densities considered as functions on A_1). Then in order that h be a maximum likelihood estimate it must be true that every direction of recession of the log likelihood is a direction of constancy of the family, i. e. that

$$\sigma_{D_1}(\phi) = \sigma_{K_1}(\phi) + \delta_{\Theta_1}(\phi) \le \langle x, \phi \rangle \tag{4.11}$$

implies that ϕ is a direction of constancy of the family. Now (4.11) implies that x is on the relative boundary of D_1, because (4.11) must hold with equality if $m(x) < +\infty$. Let

$$A_2 = \{\, y \in A_1 : \langle y - x, \phi \rangle = 0 \,\}.$$

Then, since ϕ was a direction of constancy, the family on A_1 is the same up to almost everywhere equivalence as the family conditioned on A_2, and $D_2 = M \cap A_2$ is a face of D_1. If this process is repeated until (4.11) does not hold (with the subscripts 1 replaced by 2, 3, ...), we finally arrive at A_n, D_n. Such that h "is" in the exponential family obtained by conditioning the original family on A_n, and x is in the relative interior of the face D_n [62, Theorem 13.1].

The relative interiors of nonempty faces of a convex set are a partition of the set [62, Theorem 18.2] so there is a unique face D of M containing the observation x in its relative interior. This is obviously the smallest face containing x. Since h was arbitrary, every maximum likelihood estimate is, up to almost everywhere equivalence, the maximum likelihood estimate in the original family conditioned on $\operatorname{cl} D$. But since the maximum likelihood estimate in a convex exponential family is unique, the log likelihood being strictly convex in every direction that is not a direction of constancy, the maximum likelihood estimate in the closure in the topology of pointwise convergence is unique as well. □

The theory of generalized exponential families presented here is a satisfactory solution to the problem of maximum likelihood in an exponential family if the supremum of the likelihood is finite. When the supremum of the likelihood is not finite (which can only happen with nonzero probability if the family is neither discrete nor continuous), the theory fails to produce sensible results. The following two examples show this.

Example 4.7 Let λ be the measure in $\mathbb{R}^2$ uniform on the boundary of the unit circle (hence a one-dimensional measure embedded in a two-dimensional space). For any observation x, there is a direction of recession ϕ that is not a direction of constancy (the normal to the circle at x) but the hyperplane H_ϕ tangent to the circle at x has measure zero. So the supremum of the likelihood is $+\infty$ and the theory produces no estimate. Since this is true for all observations, the maximum likelihood estimate does not exist, with probability one, even in the closure in the topology of almost everywhere convergence of densities.

This example is not as bad as it looks at first, because the maximum likelihood estimate does exist in this family for any sample size n greater than one when there is repeated sampling. Then the problem is to maximize $l_{\bar{x}_n}$ (see Section 5.2). But this always has a solution for $n \geq 2$ because then $\bar{x}_n$ lies in the interior of the unit circle with probability one.)

A slight variation on this example causes more of a problem.

Example 4.8 (The Dunce Cap Distribution) Now let λ be the measure in $\mathbb{R}^3$ which is a mixture of the unit circle in the x_1-x_2 plane and an atom somewhere on the x_3 axis, not at the origin. So the convex support is a circular cone (a dunce cap) with an atom at the vertex and a uniform distribution on the rim. Any distribution in the exponential family generated by this measure has some probability on the atom at the vertex and some on the circle. So for all sample sizes it is an event of positive probability that the observation have $n - 1$ of the sample points at the vertex and one sample point on the circle. Then the maximum likelihood estimate does exist in the theory presented here, but is a bit weird.

There is a nonzero normal ϕ at the observation $\bar{x}_n$ which lies on the surface of the cap $\frac{1}{n}$ of the way from the sample point on the circle to the vertex. The hyperplane H_ϕ normal to the sample space at the observation does not have measure zero, so the maximum likelihood estimate does exist, but the restriction $\lambda | H_\phi$ is concentrated at the vertex. Hence so is the maximum likelihood estimate. Thus the maximum likelihood estimate is concentrated at the vertex, but the data are not.

Though these examples give some impetus to think about other topologies, perhaps substituting convergence in distribution for convergence pointwise almost everywhere, it is not clear that any useful theory results. There then appears to be no analog of Theorem 2.6, which is the key to the whole theory. Nor is it clear that there are any useful applications for such a theory. In any case, the current theory is completely satisfactory for all discrete and continuous families.

The reader should be warned that there have been at least two other definitions of "generalized exponential families," one closely related to the definition given here and the other not. The one that is unrelated is the definition of Soler [68] which refers an analogue of exponential families on infinite-dimensional topological vector spaces. The one that is closely related is the definition of Lauritzen [51, p. ff.] that has also been studied by Johansen and others [48]. Lauritzen actually calls his construction

"general," not "generalized" exponential families, but the terms are close enough to cause confusion, especially since the two constructions are partly aimed at the same problem. Lauritzen defines a general exponential family as follows. Let S be a countable semigroup, and let S^* be the set of all homeomorphisms from S to $[0, \infty)$, that is $f \in S^*$ if ξ is a nonnegative real function on S and $\xi(xy) = \xi(x)\xi(y)$ for all x and y in S. If λ is a probability measure on S its Laplace transform is defined to be the function c from S to $[0, \infty]$ defined by

$$c(\xi) = \int \xi \, d\lambda, \qquad \xi \in S^*.$$

Then a family of probability densities with respect to λ is a general exponential family in the sense of Lauritzen if the densities have the form

$$f_\xi(x) = \frac{1}{c(\xi)} \xi(x), \qquad x \in S,$$

where necessarily ξ is restricted to points in

$$\Xi = \{ \, \xi \in S : c(\xi) < +\infty \, \}.$$

(The definition here is, of course, just the "standard" case. As with traditional exponential families it is always possible to take x to be an S-valued random element of some other probability space Ω, but this adds no essential novelty.)

This definition includes families that have little relation to the traditional notion of exponential families. It in fact includes all cases of independent, identically distributed sampling from any discrete family (see also [48]). In the case of a discrete exponential family the two theories produce the same results, although the theory developed here, using the affine structure of an exponential family, gives much stronger results, algorithms for computing estimates, and so forth, that are not available when only the semigroup structure of the sample space is used.

Chapter 5

MONTE CARLO MAXIMUM LIKELIHOOD

5.1 Monte Carlo Likelihood

In many problems of interest the integral

$$c(\theta) = \int e^{\langle x, \theta \rangle} \, d\lambda(x)$$

that defines the Laplace transform of an exponential family is analytically intractable. This makes numerical evaluation of the likelihood impossible and hence numerical evaluation of the maximum likelihood estimate. Many important exponential family models have this property. They arise in almost every situation of dependence more complicated than a simple Gaussian time series, such as Markov random fields used in image processing, in statistical genetics, in sociology, and in artificial intelligence. They also arise in regression with serial dependence in the response. (See Section 6.10 for references).

In many such models it is possible to simulate realizations from a distribution in the family, either independent realizations (classical Monte Carlo) or dependent realizations forming a Markov chain whose equilibrium distribution is the distribution being simulated (the Metropolis algorithm and the Gibbs sampler [54, 30]). This is explained more fully in Chapter 6.

Such a simulation produces an ergodic stochastic process $X_1, X_2, \ldots$ whose invariant distribution is some distribution in the exponential family, say F_ψ. The process is stationary if started in the invariant distribution (that is if X_1 is distributed according to F_ψ). By the ergodic theorem (which reduces to the strong law of large numbers in independence case)

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) \to \int g \, dF_\psi, \qquad \text{a. s.} \tag{5.1}$$

for any integrable function g [17, pp. 113-19]. This is true even if $\int g \, dP_\psi$ is $\pm\infty$, that is, if the positive part of g is integrable but the negative part is not or vice versa.

A simple argument shows that (5.1) holds even if the process is not stationary as long as the initial distribution is dominated by the invariant distribution, that is, as long as X_1 is distributed according to some distribution Q that is dominated by F_ψ. The "almost surely" in (5.1) means that it holds except for a null set N of sample paths. There exists a conditional distribution $P(N|X_1)$ of N conditioned on the σ-field generated by X_1. By definition of conditional probability, the probability of the event N when the process is started in F_ψ is

$$\int P(N|X_1)\, dF_\psi = 0.$$

But this means that $P(N|X_1) = 0$ almost everywhere $[F_\psi]$. By the Markov property the probability of the event N when the process is started in Q is $\int P(N|X_1)\, dQ$. Since Q is dominated by F_ψ, this integral is also zero. Is is not actually necessary that the process X_1, X_2, ... be Markov. It is enough that there be a Markov chain simulation of a general (not necessarily standard) exponential family having X as its natural statistic. That is, ω_1, ω_2, ... is an ergodic Markov chain having some distribution in the (general) exponential family as its invariant distribution. So $X(\omega_1)$, $X(\omega_2)$, ... is an ergodic (but not necessarily Markov) stochastic process whose invariant distribution is the corresponding distribution in the standard family. It is enough that ω_1 be simulated from a distribution dominated by the invariant distribution of the general family. This is the one place in all of the theory discussed in this dissertation where it is necessary to mention a general exponential family.

Observe that

$$c(\theta) = \int e^{\langle x, \theta \rangle}\, d\lambda(x) = c(\psi) \int e^{\langle x, \theta - \psi \rangle}\, dF_\psi(x).$$

So the natural Monte Carlo estimate of $c(\theta)$ is

$$c_n(\theta) = c(\psi) \frac{1}{n} \sum_{i=1}^{n} e^{\langle X_i, \theta - \psi \rangle}.$$

This depends on $c(\psi)$ which we do not know how to calculate, but this turns out not to matter. It will just be an unknown constant that drops out of the (relative) log likelihood.

By the ergodic theorem $c_n(\theta)$ will converge to $c(\theta)$ for any fixed θ for almost all sample paths X_1, X_2, ... This is also true for any countable family θ_i, $i = 1, 2, \ldots$.

For each θ_i there is an exception set N_i of sample paths such that N_i has probability zero and $c_n(\theta_i) \to c(\theta_i)$ for sample paths not in N_i. But $N = \bigcup_{i=1}^{\infty} N_i$ is also a set of probability zero, since

$$\Pr\left(\bigcup_{i=1}^{\infty} N_i\right) \le \sum_{i=1}^{\infty} \Pr(N_i) = 0$$

Thus for almost all sample paths

$$c_n(\theta_i) \to c(\theta_i), \qquad i = 1, 2, \ldots,$$

that is, for each sequence X_1, X_2, ... not in the exception set N of probability zero, the convergence occurs simultaneously for all θ_i.

For an observation x the unconstrained log likelihood is given by

$$\tilde{l}(\theta) = \langle x, \theta \rangle - \log c(\theta)$$

(here the notation has been changed from $\tilde{l}_x$ used in Chapter 2 to $\tilde{l}$ suppressing the explicit dependence on the observation x). The unconstrained Monte Carlo log likelihood is the same with c replaced by c_n

$$\tilde{l}_n(\theta) = \langle x, \theta \rangle - \log c_n(\theta).$$

The constrained versions are, as in (2.8) formed by subtracting the indicator of the natural parameter set Θ.

$$l(\theta) = \tilde{l}(\theta) - \delta_\Theta(\theta)$$

and

$$l_n(\theta) = \tilde{l}_n(\theta) - \delta_\Theta(\theta).$$

The ergodic theorem implies, for almost all sample paths, that $\tilde{l}_n(\theta) \to \tilde{l}(\theta)$ and $l_n(\theta) \to l(\theta)$ simultaneously for all θ in any countable set, which could be chosen to be a dense set, since any finite-dimensional vector space is separable.

This is as much as we can get from the ergodic theorem. What is wanted is a stronger conclusion that the Monte Carlo log likelihood l_n converges (as a function) to the true log likelihood l in some sense that implies convergence of the Monte Carlo maximum likelihood estimate (the argmax of l_n) to the true maximum likelihood estimate (the argmax of l). It turns out that this does follow from convergence pointwise on a dense set by standard facts of variational analysis, but before proving this we take a digression to set up another problem having a similar analysis.

5.2 The Likelihood for Repeated Sampling

Up to now repeated sampling has hardly been mentioned, because no mention has been made of asymptotics. For an independent identically distributed n-sample x_1, x_2, $\ldots$, x_n from an exponential family with respect to a measure λ with Laplace transform c, the probability density for the parameter θ of the n-sample (with respect to the n-fold product of λ) is

$$\prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{c(\theta)^n} \exp\left(\sum_{i=1}^{n} \langle x_i, \theta \rangle\right) = \frac{1}{c(\theta)^n} e^{\langle n\bar{x}_n, \theta \rangle} \tag{5.2}$$

where

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Notice that the density depends on x_1, x_2, $\ldots$, x_n only through $\bar{x}_n$. Thus the density of $n\bar{x}_n$ with respect to the n-fold convolution of λ (written λ^{*n}) is given by the right hand side of (5.2). This means that the Laplace transform of λ^{*n} is c^n. Repeated sampling in an exponential family with natural statistic x and log likelihood l_x produces the exponential family with natural statistic $\bar{x}_n$ and unconstrained log likelihood $n\tilde{l}_{\bar{x}_n}$, since

$$\log\left(\prod_{i=1}^{n} f_\theta(x_i)\right) = \sum_{i=1}^{n} \tilde{l}_{x_i}(\theta) = n\tilde{l}_{\bar{x}_n}(\theta).$$

The repeated sampling produces the "same" maximum likelihood problem with $\bar{x}_n$ substituted for x.

Consider the average log likelihood ratio for an arbitrary parameter point θ versus the true parameter point θ_0

$$\begin{aligned} \tilde{a}_n(\theta) &= \tilde{l}_{\bar{x}_n}(\theta_0) - \tilde{l}_{\bar{x}_n}(\theta) \\ &= \log \frac{c(\theta)}{c(\theta_0)} + \langle \bar{x}_n, \theta_0 - \theta \rangle \end{aligned}$$

and its expectation

$$\tilde{a}(\theta) = E_{\theta_0}\big(\tilde{a}_n(\theta)\big) = E_{\theta_0}\big(\tilde{a}_1(\theta)\big) = E_{\theta_0}\left(\log \frac{f_{\theta_0}}{f_\theta}\right), \tag{5.3}$$

which is the Kullback-Leibler "distance" of f_θ from f_{θ_0}.

Theorem 5.1 *The Kullback-Leibler distance function $\tilde{a}$ defined by (5.3) is a non-negative, lower semicontinuous, proper convex function that is strictly convex in any direction that is not a direction of constancy of the family.*

PROOF. By Jensen's inequality

$$\tilde{a}(\theta) = -\int \left(\log \frac{f_\theta}{f_{\theta_0}}\right) f_{\theta_0} d\lambda \geq -\log\left\{\int \left(\frac{f_\theta}{f_{\theta_0}}\right) f_{\theta_0} d\lambda\right\} = -\log\left(\int f_\theta d\lambda\right) = 0.$$

So $\tilde{a} \geq 0$. That $\tilde{a}$ is convex follows from the fact that it is the average of convex functions, i. e., for each observation x

$$\tilde{a}_1(\theta) = \log \frac{f_{\theta_0}(x)}{f_\theta(x)} = \langle x, \theta_0 - \theta\rangle + \log \frac{c(\theta)}{c(\theta_0)}$$

is a convex function of θ. That $\tilde{a}$ is strictly convex in a direction ϕ if and only if ϕ is a direction of constancy of the family follows from similar considerations. The convexity inequality

$$\tilde{a}\Big(t\theta_1 + (1-t)\theta_2\Big) \leq t\tilde{a}(\theta_1) + (1-t)\tilde{a}(\theta_2)$$

holds with equality (that is, $\tilde{a}$ is not strictly convex in the direction $\theta_1 - \theta_2$) if and only if

$$\tilde{a}_1\Big(t\theta_1 + (1-t)\theta_2\Big) \leq t\tilde{a}_1(\theta_1) + (1-t)\tilde{a}_1(\theta_2)$$

holds with equality almost surely (that is, when $\theta_1 - \theta_2$ is a direction of constancy). That $\tilde{a}$ is proper follows from $\tilde{a} \geq 0$ and $\tilde{a}(\theta_0) = 0$.

By Theorem 7.5 in Rockafellar [62] a proper convex function is lower semicontinuous if and only if it is lower semicontinuous relative to lines, that is, $\tilde{a}$ is lower semicontinuous if and only if for each θ and ϕ the function $h : s \mapsto \tilde{a}(\theta + s\phi)$ is lower semicontinuous on $\mathbb{R}$. If h is identically $+\infty$ or is finite for only one s and $+\infty$ elsewhere, then h is lower semicontinuous. Otherwise, by convexity, h is finite on some open interval W. Then $c(\theta + s\phi)$ must be finite for $s \in W$ (because otherwise $\tilde{a}_1(\theta + s\phi)$ would be $+\infty$ for all x which would imply $\tilde{a}(\theta + s\phi) = +\infty$). Hence the first term on the right hand side in

$$\tilde{a}(\theta + s\phi) = E_{\theta_0}\Big(\langle x, \theta_0 - \theta - s\phi\rangle\Big) + \log \frac{c(\theta + s\phi)}{c(\theta_0)}$$

must be finite also for s in W. But this requires that both $\langle x, \theta_0 - \theta \rangle$ and $\langle x, \phi \rangle$ have finite expectations. So

$$\tilde{a}(\theta + s\phi) = E_{\theta_0}\Big(\langle x, \theta_0 - \theta \rangle\Big) + sE_{\theta_0}\Big(\langle x, \phi \rangle\Big) + \log \frac{c(\theta + s\phi)}{c(\theta_0)}$$

is the sum of a constant term plus a term linear in s plus a term lower semicontinuous in s and is hence lower semicontinuous. □

As in the preceding section for almost sample paths (now independent samples x_1, x_2, ... from F_{θ_0})

$$\tilde{a}_n(\theta) \to \tilde{a}(\theta), \qquad \text{as } n \to \infty$$

holds for all θ in any countable set by the strong law of large numbers. Again what is wanted is a stronger conclusion the average log likelihood $\tilde{a}_n$ converges (as a function) to the Kullback-Leibler distance function $\tilde{a}$ in some sense that implies that the maximum likelihood estimate for sample size n (the argmin of $\tilde{a}_n$) converges to the truth (the argmin of $\tilde{a}$, which contains only points θ that are equivalent to θ_0).

5.3 Likelihood Convergence

In both cases described in the preceding two sections, we have the following setup. There are lower semicontinuous proper convex functions $\tilde{a}_n$ converging pointwise on a dense set to a lower semicontinuous proper convex function $\tilde{a}$ (for the Monte Carlo likelihood case take $\tilde{a}_n = -\tilde{l}_n$ and $\tilde{a} = -\tilde{l}$. Let Θ be the natural parameter set in the family under investigation. Then we want to know about the behavior of the functions

$$a_n = \tilde{a}_n + \delta_\Theta$$

(which is the negative of the constrained Monte Carlo log likelihood for the Monte Carlo problem and the average log likelihood ratio for sample size n in the repeated sampling problem) and

$$a = \tilde{a} + \delta_\Theta$$

(which is the negative of the true log likelihood in the Monte Carlo problem and the Kullback-Leibler distance function in the repeated sampling problem).

A switch has been made here from maximizing a concave function, which is the way the maximum likelihood problem has been described up till now. To minimizing

a convex function (the negative of the log likelihood or average log likelihood). That the Kullback-Leibler distance is usually described as being nonnegative gives some excuse, but the real reason is to make the rest of the chapter easier to understand. The standard literature on optimization theory has a bias toward minimization problems, and staying with maximization would make reference to this literature extremely confusing.

In what sense (beyond the convergence on a dense set that follows from the ergodic theorem) does a_n converge to a? In what sense (if any) does $\operatorname{argmin} a_n$ (the Monte Carlo approximant to the maximum likelihood estimate in the Monte Carlo problem and the maximum likelihood estimate for sample size n in the repeated sampling problem) converge to $\operatorname{argmin} a$ (the actual maximum likelihood estimate in the Monte Carlo problem and the true parameter value(s) in the repeated sampling problem)? Does $\inf a_n$ converge to $\inf a$? For $\alpha > 0$ define the sets

$$S_{n,\alpha} = \{\, \theta \in \Theta : a_n(\theta) \le \inf a_n + \alpha \,\}$$

and

$$S_\alpha = \{\, \theta \in \Theta : a(\theta) \le \inf a + \alpha \,\}.$$

These level sets of the likelihood are likelihood-based confidence sets. $S_{n,\alpha}$ is the set of all hypotheses having log likelihood (either Monte Carlo or n-sample) within α of the maximum. S_α is the set of all hypotheses having log likelihood (Monte Carlo problem) or Kullback-Leibler distance (repeated sampling problem) within α of the maximum. In what sense (if any) does $S_{n,\alpha}$ converge to S_α?

The appropriate sense to describe the convergence of a_n to a is *epiconvergence*. The appropriate sense to describe the convergence of $S_{n,\alpha}$ to S_α is Kuratowski set convergence. Both are described in Section A.13.

A function g is said to be *level-bounded* if the level set

$$\{\, x : g(x) \le \alpha \,\}$$

is bounded for all real α. A sequence $\{g_n\}$ of functions is said to be *equi-level-bounded* if it is minorized by a level-bounded function g ($g_n \ge g$, for all n). The sequence is said to be *eventually* equi-level-bounded if it is eventually minorized by a level-bounded function g (for some n_0, $g_n \ge g$, for all $n \ge n_0$).

Theorem 5.2 *Suppose that* $\{a_n\}$ *and* a *are as in either the Monte Carlo problem or the repeated sampling problem. Suppose that the family has no directions of constancy, that* $\operatorname{int}(\operatorname{dom} a)$ *is nonempty, and that* a *has no directions of recession. Suppose further that the natural parameter set* Θ *is closed and convex. Then for almost all sample paths the following hold simultaneously*

(a) $a_n \xrightarrow{e} a$,

(b) $\{a_n\}$ *is eventually equi-level-bounded.*

(c) a_n *is eventually strictly convex with no directions of recession,*

(d) $\operatorname{argmin} a_n$ *is eventually a singleton and* $\operatorname{argmin} a$ *is a singleton,*

(e) $\inf a_n \to \inf a$,

(f) $\operatorname{argmin} a_n \to \operatorname{argmin} a$, *and*

(g) $S_{n,\alpha} \to S_\alpha, \qquad \alpha > 0.$

REMARK. The conditions that the family have no directions of constancy and that $\operatorname{int}(\operatorname{dom} a)$ be nonempty are satisfied whenever a minimal representation is used. The condition that Θ be actually closed (instead of just closed relative to the effective domain of the Laplace transform c) does no harm if we allow points in Θ that are not in $\operatorname{dom} c$. Although such points θ do not correspond to any f_θ and thus should not, strictly speaking, be part of the parameter set, since $a_n(\theta) = a(\theta) = +\infty$ for such points, they cannot be part of the maximum likelihood estimate or any likelihood-based confidence interval. The condition that a have no directions of recession can be arranged in the Monte Carlo problem by solving the Phase I maximum likelihood problem. It is always true in the repeated sampling problem if the family has been reduced to a minimal representation.

PROOF. For convex functions, pointwise convergence on a dense set implies epiconvergence if the limit is a lower semicontinuous function whose effective domain has a nonempty interior [63, Proposition 3D.10]. This implies (a).

If a sequence of convex functions epiconverges to a limit having no directions of recession, then the sequence is eventually equi-level-bounded [63, Propositions 3C.21 and 3C.22]. This implies (b).

In the repeated sampling problem, the family having no directions of constancy, the log likelihood has no directions of constancy either. In the Monte Carlo problem, a_n may have directions of constancy even though a does not. This can happen, however, only if the Monte Carlo sample X_1, X_2, ... is concentrated on a hyperplane even though the invariant distribution F_ψ is not. This cannot happen for all n, because then for any set A not in the hyperplane such that $F_\psi(A) > 0$

$$\frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) \to F_\psi(A)$$

which implies that some X_i lies in A, hence not in the hyperplane. Since the subspace spanned by the Monte Carlo sample increases with n, if X_1, X_2, ... is not concentrated in a hyperplane for one n is is not for all larger n. This proves (c). Thus $\operatorname{argmin} a_n$ is nonempty [62, Theorem 27.1]. And by strict convexity, the minimum is achieved at a unique point. The same argument applies to $\operatorname{argmin} a$. This is (d).

If an eventually equi-level-bounded sequence $\{a_n\}$ epiconverges to a limit a that is lower semicontinuous and proper, then (e) holds and

$$\limsup_{n \to \infty} \operatorname{argmin} a_n \subset \operatorname{argmin} a$$

[63, Propositions 3C.13 and 3C.15]. Since the sets involved are eventually singletons, set convergence reduces to ordinary convergence of a sequence (defined for all n larger than some n_0). This is (f).

The epiconvergence in (a) and the convergence of the infima in (e) implies for any α that

$$\limsup_{n \to \infty} S_{n,\alpha} \subset S_\alpha$$

and

$$\liminf_{n \to \infty} S_{n,\alpha_n} \supset S_\alpha$$

for some sequence $\alpha_n \to \alpha$ [63, Proposition 3C.11]. Hence

$$\liminf_{n \to \infty} S_{n,\alpha} \supset \liminf_{n \to \infty} S_{n,\alpha_n - \epsilon} \supset S_{\alpha - \epsilon}, \qquad \forall \epsilon > 0,$$

and

$$\liminf_{n\to\infty} S_{n,\alpha} \supset \operatorname{cl}\left(\bigcup_{\epsilon>0} S_{\alpha-\epsilon}\right) \supset S_\alpha$$

because a set lim inf is closed and

$$\bigcup_{\epsilon>0} S_{\alpha-\epsilon} = \{\, \theta : a(\theta) > \inf a + \alpha \,\}$$

has the same closure as S_α by the convexity of a [62, Theorem 7.6]. □

A similar theorem can be proved for nonconvex families (i. e. the natural parameter set Θ is not convex) based on the following lemma.

Lemma 5.3 *Suppose $f_n \xrightarrow{e} f$ and $C_n \to C$ with f_n convex, f proper, lower semicontinuous convex,* $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$*, and C_n and C closed. Let $g_n = f_n + \delta_{C_n}$ and $g = f + \delta_C$. Then $g_n \xrightarrow{e} g$.*

Proof. The assumptions imply that f_n actually converges uniformly to f on every compact set that does not meet the boundary of $\operatorname{dom} f$ [63, Proposition 3D.10]. Thus since $\delta_{C_n} \xrightarrow{e} \delta_C$ [63, Proposition 3C.5], g_n epiconverges to g at every point not on the boundary of $\operatorname{dom} f$ [63, Proposition 3D.8]. By the sequential compactness property of epiconvergence [63, Proposition 3C.8] every subsequence of $\{g_n\}$ has a further subsequence that epiconverges everywhere to some limit which agrees with g on the complement of the boundary of $\operatorname{dom} f$. But it must agree on the boundary as well, since both g and the limit of the subsubsequence are lower semicontinuous, hence determined by their values on any dense set, such as the complement of the boundary of $\operatorname{dom} f$. Hence every subsequence of $\{g_n\}$ has a further subsequence that epiconverges to g, which implies $g_n \xrightarrow{e} g$. □

Theorem 5.4 *Suppose that $\{a_n\}$ and a are as in either the Monte Carlo problem or the repeated sampling problem. Suppose that the family has no directions of constancy, that* $\operatorname{int}(\operatorname{dom} \tilde{a})$ *is nonempty, and that $\tilde{a}$ has no directions of recession. Suppose further that the natural parameter set Θ is closed. Then for almost all sample paths the following hold simultaneously*

(a) $a_n \xrightarrow{e} a$,

(b) $\{a_n\}$ *is eventually equi-level-bounded.*

(c) $\inf a_n \to \inf a$,

(d) $\limsup_n \operatorname{argmin} a_n \subset \operatorname{argmin} a$, *and*

(e) $\limsup_n S_{n,\alpha} \subset S_\alpha, \qquad \alpha > 0.$

REMARK. As in Theorem 5.2 the conditions that the family have no directions of constancy and that $\operatorname{int}(\operatorname{dom} \tilde{a})$ be nonempty are satisfied whenever a minimal representation is used. The remark there about Θ being closed applies here as well. The condition that $\tilde{a}$ have no directions of is always true in the repeated sampling problem if the family has been reduced to a minimal representation. In the Monte Carlo problem this condition need not hold and the Phase I algorithm of Chapter 2 does not apply to nonconvex families. The condition will hold, however, when the family is continuous, since then the unconstrained log likelihood has no directions of recession (with probability one).

Notice that dropping the hypothesis that Θ is convex requires that the hypotheses that are applied to the constrained objective function a in Theorem 5.2 apply to the unconstrained objective function $\tilde{a}$ here. Also notice that the conclusions are now weaker. The maximum likelihood estimate need not be unique (a singleton set) in the absence of convexity, and now (d) and (e) only have one side of (f) and (g) in Theorem 5.2. This is somewhat satisfactory for (d) which says that every cluster point of a sequence of minimizers of a_n is a minimizer of a. It is less satisfactory for (e). Even the lim sup of the approximate confidence sets need not cover desired limit S_α. This is, however, the best result that holds in general.

PROOF. That $\tilde{a}_n \xrightarrow{e} \tilde{a}$ follows, as in Theorem 5.2 [63, Proposition 3D.10] as does the eventual equi-level-boundedness of $\{\tilde{a}_n\}$. But this also implies (b) since $\tilde{a}_n \leq a_n$. And (a) follows from Lemma 5.3.

Assertions (c), (d), and (e) follow directly from propositions in Rockafellar and Wets [63, Propositions 3C.11, 3C.13,and 3C.15]. □

Chapter 6

AN APPLICATION TO DNA FINGERPRINTING DATA

6.1 Introduction

The remaining chapters apply the theory developed so far. This chapter arose from one applied problem. Given genetic data on a set of individuals, specifically data from DNA fingerprinting [47], what is the relationship among the individuals? How closely is each related to each of the others? The study of this problem lead to the development of the Phase I algorithm of Chapter 2 and the Monte Carlo maximum likelihood algorithm of Chapter 5, and has revealed difficulties with maximum pseudolikelihood estimates (MPLE) and maximum conditional likelihood estimates (MCLE) that have not previously been noted.

An example of simulated DNA fingerprinting data is shown in Figure 6.1, which

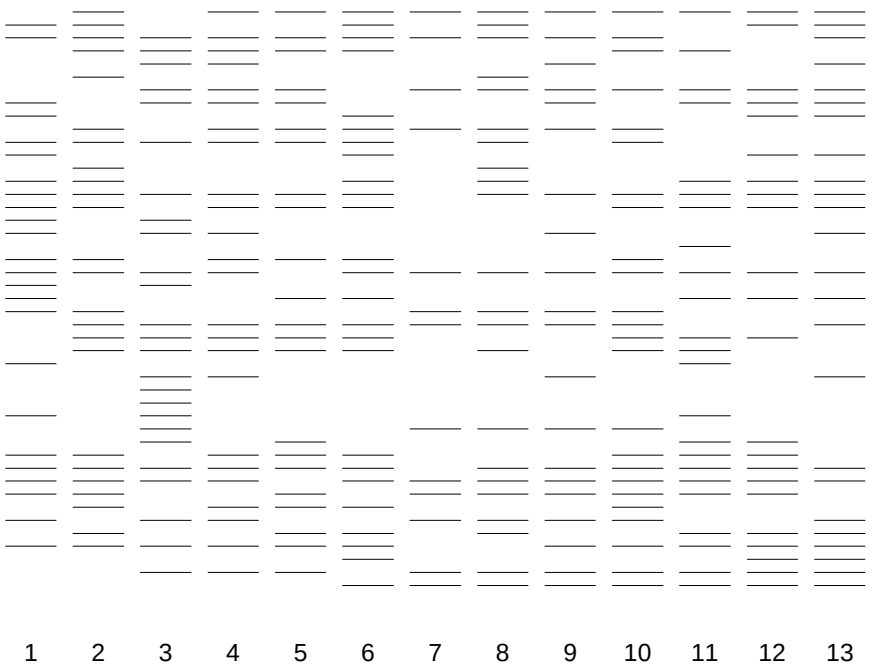


Figure 6.1: Simulated DNA Fingerprint Data for Thirteen Individuals.

represents an electrophoretic gel. Each of the columns or "lanes" is the data for one of thirteen individuals. The black bands represent DNA fragments whose vertical position, the distance traveled from the starting position, indicates the size of the fragment, smaller fragments moving faster. On a real gel the bands would be fuzzy and not exactly aligned horizontally. So one would not be certain which bands in which lanes correspond. We ignore this problem and assume perfect data in which all bands can be identified. That is, we assume we are given a matrix Y of 0's and 1's such that $Y_{ij} = 1$ if individual i has the band in the jth row and $Y_{ij} = 0$ otherwise.

With this simplification, we arrive at the problem of modeling dependent binary data. The stochastic process which produces the data is well understood. It involves known genetic mechanisms (Mendelian segregation, mutation, and recombination) and the molecular genetics of DNA fingerprinting (restriction sites, fragment sizes, probes recognizing specific DNA sequences). A stochastic model can be constructed that incorporates all of these factors; in fact, the data in Figure 6.1 were simulated from such a model. There are, however, far too many parameters in the model to be estimated from the data. Moreover, the parameters of interest, the genealogical relationships of individuals, are discrete, being of the form so-and-so is the second cousin once removed of such-and-such. The collection of all such relationships, the genealogy or pedigree of the individuals, does determine the dependence structure of the data, but in an extremely complex manner. There is little hope of estimating the pedigree from the data.

What is wanted is a simpler model with fewer parameters. Such a model can be derived by reasoning in the opposite direction from the usual statistical argument. Rather than choose a model having some desired properties and then find its sufficient statistics, we start with reasonable sufficient statistics and let them determine the model. The model will be the exponential family induced by these sufficient statististics and some invariant measure on the sample space (here counting measure because the space is discrete and the model should be invariant under relabeling of individuals or bands). The model for DNA fingerprint data produced by this argument is an autologistic model [12].

This procedure of deriving the model from the statistics is unusual, but has a long history. It is essentially the argument used by Maxwell in his derivation of the velocity distribution of an ideal gas. More recent examples of this view have

been the notion of "maximum entropy" [46], and the notion of "extremal families" [51]. Statistical inferences drawn from a model and data do not, of course, depend on the philosophical justification of the model. The validity of such inferences may. Most models, like our model for DNA fingerprint data, are simplifications of reality. Strictly speaking, the model is false, and what can one say about the validity of inferences drawn from such a model? The usual answer is that the inferences are invalid, but the argument of maximum entropy says that, even though the model is incorrect, inferences based on it are the best one can do using only the sufficient statistics chosen to specify the model. Better inferences would require a larger set of sufficient statistics, which requires knowing which statistics to choose and more data to estimate the extra parameters.

The notion of deriving the model from the statistics is not useful unless it is easier to choose among statistics than among models, but there is often a natural choice among statistics. For the DNA fingerprint data, we choose the row and column sums of the data matrix Y (the number of bands exhibited by each individual and the number of individuals exhibiting each band) and the number of shared bands for each pair of individuals. A subset of these statistics, the band counts for individuals and the counts of shared bands is already in common use [53]. The sample band frequencies (column sums) have been omitted from previous analyses even though their importance was recognized, because without a stochastic model to guide inference it is difficult to know what to do with them. These analyses have used a similarity index for each pair of individuals derived solely from the band counts and number of shared bands for those individuals in estimating the relatedness of a pair. This procedure does not work well [53], and the reason is not hard to see. There is information about the relatedness of individuals not only in their similarity to each other but also in their common similarity to other individuals who are close relatives. A way to use all of the information in the data is to estimate the parameters by maximum likelihood.

Before starting to find a maximum likelihood estimate (MLE) one should first discover whether it exists. The commonly used procedure of running an optimization algorithm until "convergence" and then looking for large parameter values that possibly should "go off to infinity" hardly constitutes a method. Determining which parameters tend to infinity, along what path (possibly curved), and what probability distribution is produced in the limit (the MLE in the closure of the family) cannot

be determined simply by inspection of some parameter set that nearly maximizes the likelihood. The limiting distribution will be a distribution in the family conditioned on a face of the convex support of the natural sufficient statistic [7]. The determination of this face, the support of the MLE, is a purely geometric problem, depending only on the geometries of the convex support and parameter space and on the location of the observed data. It is essentially a sequence of linear programming feasibility problems, which we call "phase I" of the maximum likelihood problem. Once the support of the MLE is found, determination of the MLE within the family conditioned on the support (which is phase II) proceeds by using some optimization algorithm, as in standard methods, with the MLE guaranteed to exist. The application of the algorithm for the phase I problem to the DNA fingerprint problem is sketched in Section 6.5. The theory is completely described in Chapter 2.

Maximum likelihood for autologistic models has previously been considered an intractable problem even for simple models with only a few parameters. Any useful model for the DNA fingerprint problem must have at least as many parameters as there are pairs of individuals. For the data in Figure 6.1 with 13 individuals, our model has 136 parameters. Fortunately maximum likelihood is not as difficult as had been thought. The application of the Monte Carlo maximum likelihood algorithm to phase II of the DNA fingerprint problem is is sketched in Section 6.7. The complete theory is described in Chapter 5. It uses the Metropolis algorithm to construct a Markov chain whose equilibrium distribution is a distribution in the model. From one simulated realization of the Markov chain the whole likelihood function is approximated. That is, the likelihood is determined at all parameter values from a simulation at one parameter value. This approximation to the likelihood, the *Monte Carlo likelihood* function is the likelihood of the exponential family generated by the "empirical" distribution of the Monte Carlo sample (not an empirical derived from the data). The parameter value maximizing the Monte Carlo likelihood is taken as the Monte Carlo approximant of the MLE. Its determination is a maximum likelihood problem in an exponential family (the Monte Carlo family, not the actual model) and so is very well behaved and quickly solved. The Monte Carlo is also fast involving only a single realization of a Markov chain. Thus the whole algorithm is fast, taking a few minutes of computer time, even for large constrained models like the DNA fingerprint problem.

Other methods have been proposed for approximating the MLE by Monte Carlo in complex exponential families of this type, though none that would be applicable to constrained problems. The most studied seems to be the stochastic gradient algorithm, also called the Robbins-Munro method [76, 56]. This method uses a very simple optimization algorithm, and would consequently seem to make poor use of the Monte Carlo samples generated by the Metropolis algorithm.

The only method for estimating parameters in autologistic models that has had widespread use is MPLE [13]. Comparing it to MLE in this problem raises serious doubts about its behavior. It so overestimates the dependence parameters that data simulated from the model at the MPLE bear no resemblance to the actual data. We do not know whether this deficiency of MPLE would occur in other problems, in particular those with many fewer parameters, but its behavior in this problem suggests that further comparison with MLE is warranted.

An example of such a comparison is the small simulation study of Section 6.9. The compares MLE and MPLE for the simplest autologistic model, a one-parameter Ising model on a torus. This simulation also shows some properties of MPLE that are troubling.

Since MCLE [2] is at least as hard as MLE, it has not been used for problems of this type. The methods that we propose for MLE would also work for MCLE whenever one is able to prove that the Markov chain constructed by the Metropolis algorithm is irreducible (this would depend on the particular problem), and MCLE would be no more difficult than MLE. In theory one should get better estimates if nuisance parameters are removed by conditioning. This may, however, also remove parameters of interest, which acceptable in an unconstrained model, but is disastrous in a constrained model if the parameters removed are involved in the constraints. The DNA fingerprint problem gives an example of this phenomenon.

Though our methods are illustrated here by only one example, and that a simulated data set, this one example provides a unified demonstration of methods that are applicable to a very broad class of models.

6.2 A Constrained Autologistic Model

Suppose we have DNA fingerprinting data in which all bands can be identified. Let Y_{ij} be the indicator of whether the ith individual has the jth band ($Y_{ij} = 1$ if individual

i has band j, and $Y_{ij} = 0$ otherwise). Suppose there are m individuals and n bands. For sufficient statistics we suggest

$$U_i = \sum_j Y_{ij}, \qquad V_j = \sum_i Y_{ij}, \qquad \text{and} \qquad S_{ik} = \sum_j Y_{ij} Y_{kj}, \tag{6.1}$$

so that U_i is the number of bands exhibited by the ith individual, V_j is the number of individuals exhibiting the jth band, and S_{ik} is the number of bands shared by individuals i and k. Note that $S_{ik} = S_{ki}$ and $S_{ii} = U_i$ so it is enough to use only S_{ik} for $i < k$. The exponential family model induced by these statistics has log likelihood

$$l_Y(\theta) = \sum_i U_i \alpha_i + \sum_j V_j \beta_j + \sum_{i<k} S_{ik} \gamma_{ik} - \psi(\theta) \tag{6.2}$$

where θ denotes the set of all the parameters taken as a vector

$$\theta = (\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n, \gamma_{12}, \ldots, \gamma_{1m}, \gamma_{23}, \ldots, \gamma_{2m}, \ldots, \gamma_{(m-1)m}),$$

and ψ is some function of the parameters alone. Note that only the above-diagonal elements of the matrix γ are used; by convention we define the others by $\gamma_{ik} = \gamma_{ki}$ and $\gamma_{ii} = 0$ for all i and k.

If the statistics are joined in a vector

$$T = (U_1, \ldots, U_m, V_1, \ldots, V_n, S_{12}, \ldots, S_{1m}, S_{23}, \ldots, S_{2m}, \ldots, S_{(m-1)m}) \tag{6.3}$$

in the same way as the parameters, the log likelihood (6.2) can be rewritten as

$$l_Y(\theta) = \langle T, \theta \rangle - \psi(\theta) = \langle t(Y), \theta \rangle - \psi(\theta) \tag{6.4}$$

t denotes the mapping defined by (6.1) and (6.3) that takes the data Y to the natural statistic $T = t(Y)$ and where $\langle T, \theta \rangle = \sum T_l \theta_l$ denotes the inner product of the vectors T and θ. In this form, it is clear that the family of distributions in the model is an exponential family with natural statistic T and natural parameter θ. In the form (6.2) it is seen to be an autologistic model [12].

It is not, however, a minimal canonical exponential family. There is one redundant parameter. Since $\sum U_i = \sum V_j$, one can add a constant to all the α's and subtract the same constant from all the β's without changing the value of the likelihood.

For theoretical purposes we take any value of θ that maximizes the likelihood to be an MLE. For calculation, it is necessary to impose a linear constraint, such as constraining $\sum \beta_j$ to be constant, to force identifiability.

The densities of the family are of the form

$$f_\theta(y) = e^{l_y(\theta)} = \frac{1}{c(\theta)} e^{\langle t(y), \theta \rangle},$$

where $c = e^\psi$. The function c is the *Laplace transform* of the family (and ψ the log Laplace transform). Let $\mathcal{S}$ denote the sample space, the set of all $m \times n$ matrices of 0's and 1's. Then c is given by

$$c(\theta) = \sum_{y \in \mathcal{S}} e^{\langle t(y), \theta \rangle}. \tag{6.5}$$

Many facts about this model can be easily derived from the general theory of exponential families. Since the sum in (6.5) has a finite number of terms, $\psi(\theta) = \log c(\theta)$ is finite for all values of θ. Hence the natural parameter space of the family is the whole space $\mathbb{R}^{m+n+m(m-1)/2}$ of possible values of θ.

Since no term of the log likelihood involves data from different bands (columns of the data matrix), the bands are independent random vectors in this model. This would not be exactly true for real data. In real DNA fingerprint data some bands would be linked (DNA fragments from different loci on the same chromosome and hence positively correlated in their presence or absence in individuals), and some bands would be allelic (fragments from the same locus on a chromosome and hence negatively correlated in their presence or absence). In most applications, however, DNA fingerprint data would not include large families of known genealogy needed to detect linkage and allelism. Hence, in the analysis recommended here, we make the simplifying assumption of independence of bands.

Derivatives of ψ are given by

$$\frac{\partial \psi(\theta)}{\partial \theta_l} = E_\theta(T_l) \tag{6.6}$$

and

$$\frac{\partial^2 \psi(\theta)}{\partial \theta_l \partial \theta_m} = \operatorname{Cov}_\theta(T_l, T_m) \tag{6.7}$$

which apply to any exponential family; see for example [18, pp. 35–36]. Combining (6.6) and (6.7) gives

$$\frac{\partial}{\partial \theta_l} E_\theta(T_l) = \frac{\partial^2 \psi(\theta)}{\partial \theta_l^2} = \mathrm{Var}_\theta(T_l) > 0. \tag{6.8}$$

From (6.8) we see that the expectation of U_i increases as α_i increases, that of V_j as β_j increases, and that of S_{ik} as γ_{ik} increases. An individual is inbred if his parents are related [22]. Thus an inbred individual can receive identical genes from his mother and from his father and hence will, on average, have fewer distinct genes than an individual that is not inbred, and hence fewer bands in his DNA fingerprint. Thus U_i is low when individual i is inbred, so the α parameters are the analogue in this model of inbreeding (large negative values of α corresponding to high inbreeding). There is, of course, no direct relationship between α and inbreeding as measured by the inbreeding coefficient. Since V_j, the number of individuals exhibiting band j, is high when the population frequency of the band is high, β is the analogue of population allele frequency. Here too, there is no direct relationship between β and the real allele frequencies. Since S_{ik}, the number of bands shared by i and k, is high, on average, when i and k are closely related, γ_{ik} is the analogue of relatedness, though again not directly.

The conditional distribution of Y_{ij} and Y_{kj} given the rest of the statistics is an exponential family with log likelihood of the form

$$Y_{ij}Y_{kj}\gamma_{ik} + Y_{ij}\lambda_i + Y_{kj}\lambda_k + \psi'(\gamma_{ik}, \lambda_i, \lambda_k)$$

with log Laplace transform ψ'; natural statistics $Y_{ij}Y_{kj}$, Y_{ij}, and Y_{kj}; and natural parameters γ_{ik}, λ_i, and λ_k, the latter two being functions of θ and the rest of the components of Y. By the Fisher factorization theorem, Y_{ij} and Y_{kj} are conditionally independent given the rest when $\gamma_{ik} = 0$. There are just four points in the conditional sample space with probabilities

$$p_{ll'} = \Pr(Y_{ij} = l \,\&\, Y_{kj} = l' | \text{rest}), \qquad l = 0, 1,\ l' = 0, 1.$$

So the conditional covariance

$$\mathrm{Cov}(Y_{ij}, Y_{kj}|\text{rest}) = E\Big[Y_{ij}Y_{kj} - E(Y_{ij}|\text{rest})E(Y_{kj}|\text{rest}) \;\Big|\; \text{rest}\Big]$$

$$
\begin{aligned}
&= p_{11} - (p_{10} + p_{11})(p_{01} + p_{11}) \\
&= p_{10}p_{01}\left(\frac{p_{11}p_{00}}{p_{10}p_{01}} - 1\right) = p_{10}p_{01}(e^{\gamma_{ik}} - 1)
\end{aligned}
$$

is positive when $\gamma_{ik} > 0$ and negative when $\gamma_{ik} < 0$. Negative covariance makes no sense for genetic data. Unrelated individuals are independent, relatives are positively correlated (more alike than unrelated individuals) conditionally or unconditionally. Thus we add to the model the constraint that all of the γ's be nonnegative, and the parameter space of our model is the subset

$$\Theta = \{\, (\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n, \gamma_{12}, \ldots, \gamma_{(m-1)m}) : \gamma_{ik} \geq 0 \,\} \tag{6.9}$$

of the natural parameter space obtained by imposing this constraint.

6.3 Pseudolikelihood

Exact calculation of the likelihood is difficult except in very small problems because the Laplace transform is analytically intractable. It can only be calculated by doing the explicit sum in (6.5). This has led to the introduction of methods of estimation such as MPLE [13] that avoid the calculation of the Laplace transform.

The conditional probability distribution of Y_{ij} given the rest is a Bernoulli distribution (Y_{ij} taking values in $\{0, 1\}$) with natural parameter

$$\theta_{ij} = \log \frac{\Pr(Y_{ij} = 1|\text{rest})}{\Pr(Y_{ij} = 0|\text{rest})} = \log \frac{\Pr(Y_{ij} = 1 \,\&\, \text{rest})}{\Pr(Y_{ij} = 0 \,\&\, \text{rest})} = \alpha_i + \beta_j + \sum_{k \neq i} \gamma_{ik} Y_{kj} \tag{6.10}$$

(recall that $\gamma_{ik} = \gamma_{ki}$). Hence the log likelihood of this distribution is

$$Y_{ij}\theta_{ij} - \varphi(\theta_{ij})$$

where

$$\varphi(\theta) = 1 + e^{\theta}$$

is the log Laplace transform of the Bernoulli distribution.

The pseudolikelihood is the product of these conditional probabilities, its log, the log pseudolikelihood is

$$\sum_{i,j} \Big(Y_{ij}\theta_{ij} - \varphi(\theta_{ij}) \Big). \tag{6.11}$$

and the MPLE is any value of the parameter θ that maximizes (6.11) subject to the constraint that θ lies in Θ (or by analogy with maximum likelihood the limit of a maximizing sequence).

While the pseudolikelihood for this model is not the true log likelihood for any model, it is computationally identical to the log likelihood for a logistic regression in which the data Y serves both as a response and a predictor. Maximum pseudolikelihood is in effect a logistic regression of each component of Y on the rest of the components with which it is correlated.

Where Y_{ij} appears explicitly in (6.11) it plays the role of "response." Where it appears implicitly though the dependence of θ_{ij} on Y_{kj}, $k \neq i$, it plays the role of "predictor." More explicitly, substituting (6.10) in (6.11) and writing X for Y where it occurs in (6.10) so X is the "predictor" and Y is the "response," the log pseudolikelihood is

$$\begin{aligned} l_{Y,X}(\theta) &= \sum_{i,j} \Bigl(Y_{ij}(\alpha_i + \beta_j + \sum_{k \neq i} \gamma_{ik} X_{kj}) - \varphi(\theta_{ij}) \Bigr) \\ &= \sum_i \alpha_i U_i + \sum_j \beta_j V_j + \sum_{i,k} \gamma_{ik} \sum_j Y_{ij} X_{kj} - \sum_{i,j} \varphi(\theta_{ij}) \\ &= \sum_i \alpha_i U_i + \sum_j \beta_j V_j + \sum_{i<k} \gamma_{ik} \sum_j (Y_{ij} X_{kj} + Y_{kj} X_{ij}) - \sum_{i,j} \varphi(\theta_{ij}). \end{aligned} \tag{6.12}$$

All of the properties of MPLE are those of MLE in a model with this log likelihood where Y is random and X fixed. This is, of course, just an artifice, X and Y are actually two identical copies of the data, but this fiction allows us to derive facts about MPLE from the theory of maximum likelihood in exponential families.

6.4 Conditional Likelihood

While it is necessary to make some adjustment for the population frequencies of the bands, which maximum likelihood does by estimating the β's, in most cases the estimated β's are of minor interest; they are "nuisance" parameters. This suggests that conditional maximum likelihood [2] might be a good idea. The MCLE is any value of the parameters α and γ that maximizes subject to the constraint $\gamma \geq 0$ the conditional log likelihood

$$l_{Y|V}(\alpha, \gamma) = \log f_{\alpha,\gamma}(Y|V) \tag{6.13}$$

which, because of the conditioning on the observed band frequencies V, does not depend on the parameter β. (Or, as with the other methods, the MCLE is the limit of a maximizing sequence, if the supremum is not attained.) As is the case with log likelihood, the conditional log likelihood can only be calculated up to a normalizing constant, which can be calculated exactly only by summing over the whole sample space (now just the set of Y having the observed column sums $V_1, \ldots, V_n$). For large data sets, the estimates must be calculated by Monte Carlo.

With the MCLE, however, a new problem arises. The identity

$$\sum_j V_j^2 = \sum_j \sum_i \sum_k Y_{ij} Y_{kj} = \sum_i \sum_j Y_{ij}^2 + 2 \sum_{i<k} \sum_j Y_{ij} Y_{kj} = \sum_i U_i + 2 \sum_{i<k} S_{ik}$$

being nonlinear does not cause a lack of identifiability in the maximum likelihood problem. When we condition on V, this imposes a linear constraint on the remaining statistics U and S. Together with the constraint $\sum U_i = \sum V_j$ that we inherit from the maximum likelihood problem, this gives the two constraints $\sum U_i$ is constant and $\sum S_{ik}$ is constant. The first is not a problem, but the second is a serious one. It means that an arbitrary constant can be added to all of the γ_{ik} without changing the probabilities in the conditional model. Thus the constraints $\gamma_{ik} \geq 0$ no longer make sense, and only unconstrained estimation is possible. This is unavoidable if we condition on V.

6.5 The Phase I Problem

The first task (often ignored) in any maximum likelihood problem is to discover whether a maximum exists, that is, whether the likelihood achieves its supremum at some point of the parameter space. If the MLE fails to exist in this traditional sense, the MLE in the closure of the exponential family may. This MLE is a probability distribution (*not* a parameter point) that is the limit of distributions for a sequence of parameter points along which the likelihood goes to its supremum. For a convex exponential family, one in which the natural parameter set (6.9) is convex (as in the DNA fingerprint model), this theory of maximum likelihood in the closure has been understood for quite some time, but has had no practical application, perhaps because no effective algorithm for calculating the MLE in the closure had been devised. The theory for the case of a full exponential family is found in Barndorff-Nielsen [7] and

Brown [18]. The theory for convex exponential families seems to have been known as well [7, p. 164] but does not seem to have appeared in print. An algorithm for calculating the MLE in the closure of a convex family is sketched here and explained in full in Chapter 2. Note that the methods explained here cover not only MLE but also MPLE and MCLE since both are just maximum likelihood in some model.

The MLE fails to exist (the supremum of the likelihood is not attained) if and only if there exists a direction of recession of the log likelihood that is not a direction of constancy. A *direction of recession* is a nonzero vector ϕ satisfying the following two conditions

(i) $\theta + s\phi \in \Theta$, for all $s \geq 0$ and for all $\theta \in \Theta$.

(ii) $s \mapsto l_Y(\theta + s\phi)$ is a nondecreasing function for some $\theta \in \Theta$ (in which case it is nondecreasing for all $\theta \in \Theta$ because l_Y is a concave function).

When (i) holds we say that ϕ is a direction of recession of the parameter set Θ; when (ii) holds we say the ϕ is a direction of recession of the unconstrained log likelihood. A direction of recession is a *direction of constancy* if $s \mapsto l_Y(\theta + s\phi)$ is a constant function (this also will be true for all θ if it is true for any θ).

A nonzero vector ϕ is a direction of recession of the parameter set Θ given by (6.9) if and only if the γ-components of ϕ are nonnegative (the γ-components being the components corresponding to the components of θ that are the γ's). A vector ϕ satisfies (ii) if and only if it is a normal to the convex support of the natural sufficient statistic at the observation T, that is if

$$\langle t(y) - T, \phi \rangle \leq 0, \qquad y \in \mathcal{S}. \tag{6.14}$$

Note that if ϕ is a direction of constancy so that $-\phi$ is also a direction of recession, (6.14) is satisfied with equality and the natural statistic is actually concentrated on a hyperplane normal to ϕ.

Thus the problem of finding a direction of recession boils down to the question: Is there a nonzero vector which is a direction of recession of the parameter set and which satisfies (6.14)? This is a linear programming feasibility problem, since the constraints (6.14) are linear and there are a finite number of them, the sample space being finite.

If there exists a direction of recession ϕ that is not a direction of constancy, then no parameter point maximizes the likelihood, but $s \mapsto f_{\theta+s\phi}(Y)$ is nondecreasing and for any $\theta \in \Theta$ and any $y \in \mathcal{S}$

$$f_{\theta+s\phi}(y) \to f_\theta(y|D) \qquad \text{as } s \to \infty$$

where

$$D = \{\, y \in \mathcal{S} : \langle t(y) - T, \phi \rangle \le 0 \,\}$$

is the preimage of the face of the convex support normal to ϕ. Hence every density f_θ in the original family has smaller likelihood than some density in the conditional family

$$\mathcal{F}_D = \{\, f_\theta(\,\cdot\,|D) : \theta \in \Theta \,\}. \tag{6.15}$$

Every distribution of this family is the limit of a sequence of distributions in the original family so is part of the closure of the original family. Thus what has been established by discovering the direction of recession ϕ is that the MLE lies in the closure of $\mathcal{F}_D$.

We now have another convex exponential family $\mathcal{F}_D$ in which to find an MLE. This family has a different sample space: D replaces $\mathcal{S}$. It also has a different parameter space: Θ is replaced by $\{\, \theta + s\phi : \theta \in \Theta,\ s \in I\!R \,\}$ since ϕ is a direction of constancy of $\mathcal{F}_D$. There is no guarantee that the MLE exists in this new family, but if not we iterate the process finding a sequence of directions of recession and faces of the convex support. This must eventually stop because each iteration decreases the dimension of the face by at least one. The final iteration reaches a family $\mathcal{F}_{\hat{D}}$ in which the MLE lies. This ends phase I. The problem has now been reduced to finding the MLE $\hat{\theta}$ in $\mathcal{F}_{\hat{D}}$, a problem whose solution is guaranteed to exist by the phase I algorithm.

The vectors $t(y) - T$, $y \in \mathcal{S}$ are called *tangent vectors* to the convex support at T. Note that if all of the tangent vectors can be written as sums (or, more generally as nonnegative combinations) of vectors in some subset of the tangent vectors, then only the subset need be used in the phase I calculation.

An example of the use of this principle is the phase I calculation in the MPLE problem. Looking at (6.12) one sees that the natural sufficient statistic is a linear function of Y (X being fixed), and hence every tangent vector is the sum of tangent vectors in which only one component Y_{ij} is changed. So the linear programming

problem to find ϕ needs to have only mn constraints, much fewer than the 2^{mn} constraints that results from using all of the tangent vectors. Problems this size are easily solved by available linear programming software.

For the MLE problem, all of the tangent vectors are sums of vectors in which only one band (column of Y) is changed, but this still leaves a subset of $n2^m$ tangent vectors. This is far too many for linear programming. A further simplification is needed. Comparing (6.2) and (6.12) one sees that the tangent vectors in which only one component is changed are the same for the MLE and MPLE problems, hence tangent vectors in the MPLE problem are a subset of those for the MLE problem. Thus any direction of recession of the MLE problem has a nonpositive inner product with all the tangents of the MPLE problem and is a direction of recession of the MPLE problem as well. So the MLE problem can be solved by checking whether MPLE direction of recession is also an MLE direction of recession. This check will be easy if the direction to be checked has only a few nonzero components.

This illustrates a general principle. If there are too many tangent vectors, solve (6.14) for just a subset and try to check that the solution is normal to the whole set. If not, increase the subset by adding some of the tangents that had positive inner products with the solution to the subset and try again. It must be admitted that there is no guarantee that one can always solve the phase I problem. Cases could presumably be constructed in which geometry can be arbitrarily complex. Most cases, however, will be simple, and the algorithm will find the solution without exceeding the capacities of available software.

6.6 Avoiding Phase I: Name Tag Bands

When there is a direction of recession of the likelihood, the MLE in the closure of the family is an unfamiliar concept to those who think of estimates only as points in $\mathbb{R}^n$, not as probability distributions. It is also true that an estimate concentrated on a proper subset of the sample space is unsuitable for some purposes, such as a parametric bootstrap. There is a time honored tradition in statistics, dating to Laplace, of avoiding such estimates by adding imaginary data to the data actually observed.

The obvious way to implement such an approach here is to invent new bands of known population frequency, one for each individual in the data set, and each

appearing in just one individual. We call these "name tag" bands, because they are a unique identifier for each individual, like a name. We assume the name tag bands all have the same population frequency, which we set by giving all of them a β of zero. This fixes a reference level for the β's so we do not impose a constraint $\sum \beta_i =$ a constant. For comparability with the name tag estimates we choose the constraint imposed in the MLE and MPLE problems to be $\sum \beta_i = 1.25$. This constant makes the α's and β's have the same average for the MLE and MPLE as for the name tag estimates.

The use of name tags has some resemblance to Bayesian inference using a conjugate prior. The name tag MLE is the mode of the posterior for a specific prior that is proportional to the part of the likelihood contributed by the name tag bands. There is, however, no real option in the choice of "prior." There is no analytic form for the likelihood that makes sense for noninteger values of the data. So we must invent an integer amount of data, and one name tag band per individual seems to be enough. The choice of "prior" is forced. Moreover, we do not advocate the use of name tag bands when there are no directions of recession. If the name tag bands really represented prior knowledge, they should be used whether or not there is a direction of recession.

The name tag bands are enough to prevent directions of recession no matter what the real data may be, unless there are bands which appear in all individuals or none, and such bands may as well be removed from the data set since they contribute no information about relatedness. We show this in three steps.

(Step 1) Consider the tangent vector obtained by changing the data for individual i in "his" name tag band from 1 to 0. This gives a tangent vector with only the component $U_i = -1$ nonzero. Applying (6.14) this imposes the constraint $-\alpha_i \le 0$. (Here again we use the notation that α, β, and γ refer to components of ϕ not θ).

(Step 2) Consider tangent vectors obtained by changing the data for individual k in the name tag band for individual i from a 0 to a 1. This gives a tangent vector with only the components $U_k = +1$ and $S_{ik} = +1$ nonzero. This imposes the constraints $\gamma_{ik} + \alpha_k \le 0$ for all i and k, or equivalently $\gamma_{ik} \le -\alpha_k \le 0$. But since $\gamma_{ik} \ge 0$, this can only happen if all the components of γ and α are 0.

(Step 3) Now the condition that ϕ be a direction of recession can be written $\sum(V_j - v_j)\beta_j \le 0$ with probability 1 (where v denotes the observed value of V). Because of

the independence of the bands and because $V_j = v_j$ with positive probability, this implies $(V_j - v_j)\beta_j \le 0$ with probability 1 for each j, i. e., either $\beta_j = 0$ or $\beta_j < 0$ and $V_j - v_j \ge 0$ with probability 1, which can happen only if $v_j = 0$, or $\beta_j > 0$ and $V_j - v_j \le 0$ with probability 1, which can happen only if $v_j = m$.

6.7 The Phase II Problem

6.7.1 Maximum Monte Carlo Likelihood Estimates

Although exact calculation of probabilities or the likelihood is not feasible for autologistic models (except for small data sets), Monte Carlo calculations are feasible using the Metropolis algorithm (explained in the next section) or the Gibbs sampler. Either produces a stationary ergodic process T_1, T_2, ... having the same state space as the data and equilibrium distribution F_ψ, the distribution of the natural statistic T for some parameter value ψ, which may be chosen freely. How does one use the process T_1, T_2, ... to compute the MLE? The idea is to maximize the Monte Carlo approximant to the likelihood. The log Laplace transform (6.5) can be rewritten as

$$c(\theta) = c(\psi) \sum_{y \in \mathcal{S}} e^{\langle t(y), \theta - \psi \rangle} f_\psi(y) = c(\psi) \int e^{\langle t, \theta - \psi \rangle} \, dF_\psi(t),$$

which suggests the Monte Carlo approximant

$$c_N(\theta) = c(\psi) \frac{1}{N} \sum_{i=1}^{N} e^{\langle T_i, \theta - \psi \rangle}$$

of the Laplace transform, since T_1, T_2, ... are (dependent) samples with (asymptotically) marginal distribution F_ψ. (We do not know the value of $c(\psi)$, but this turns out not to matter.) More precisely, $c_N(\theta)$ converges to $c(\theta)$ almost surely as $N \to \infty$ by the ergodic theorem. Because c is a convex function, this is true for all θ, that is, except for a set of sample paths T_1, T_2, ... of probability zero, $c_N(\theta)$ converges to $c(\theta)$ for all θ along each sample path. Hence so does the Monte Carlo log likelihood

$$l_{N,Y}(\theta) = \langle T, \theta \rangle - \log c_N(\theta) = \langle T, \theta \rangle - \log \left(\frac{1}{N} \sum_{i=1}^{N} e^{\langle T_i, \theta - \psi \rangle} \right) + \log c(\psi)$$

converge to the true log likelihood l_Y given by (6.4). For each N let $\hat{\theta}_N$ be a point that maximizes $l_{N,Y}$ if a maximum exists. (Note that the value of $c(\psi)$ does not affect

the shape of the likelihood surface, only its vertical position, and so does not affect the value of $\hat{\theta}_N$.) Because of the strict convexity of c and the absence of directions of recession of l_Y (both of which are guaranteed by the phase I algorithm), there is a unique point $\hat{\theta}$ that maximizes l_Y. This and the convergence of c_N to c are sufficient to guarantee that $\hat{\theta}_N$ exists for all but a finite number of N and $\hat{\theta}_N$ converges to $\hat{\theta}$ almost surely.

One cannot actually do an infinite amount of work in generating the Monte Carlo samples. One just chooses some large N and takes $\hat{\theta}_N$ to be the MLE. The solution $\hat{\theta}_N$ is the point where the gradient of $l_{N,Y}$ is zero in an unconstrained problem or where the projection of the gradient on the set of feasible directions is zero for a constrained problem, so $\hat{\theta}_N$ will be a good estimate of $\hat{\theta}$ only if the gradient of $l_{N,Y}$ is close to the gradient of l_Y, the gradient of $l_{N,Y}$ being $T - \mu_N(\theta)$, where

$$\mu_N(\theta) = E_{N,\theta}(T) = \frac{1}{N} \frac{1}{c_N(\theta)} \sum_{i=1}^{N} T_i e^{\langle T_i, \theta - \psi \rangle} \tag{6.16}$$

is the Monte Carlo estimate of expectation of T for the parameter θ. Notice that $\mu_N(\theta)$ is just a weighted average $\sum T_i w_i(\theta)$ of the T_i with weights

$$w_i(\theta) = \frac{e^{\langle T_i, \theta - \psi \rangle}}{\sum_{k=1}^{N} e^{\langle T_k, \theta - \psi \rangle}}.$$

When $\theta = \psi$ the weights are all $\frac{1}{N}$. When θ is far from ψ there will be only a few large weights, only a few terms will make a significant contribution to the weighted average, and the Monte Carlo maximum likelihood procedure will not produce a good approximation to the MLE. To avoid this problem it is necessary to iterate the procedure. After finding an estimate, a new Monte Carlo sample is generated using the estimate as the parameter ψ for the Metropolis algorithm. But there is no need to iterate "until convergence." It is only necessary to get ψ somewhere in the vicinity of $\hat{\theta}$ so that the weights are roughly equal. A few cheap Monte Carlo runs with small sample sizes followed by one long run with N as large as can be afforded should suffice.

A convenient property of this algorithm is that each Monte Carlo likelihood is the likelihood of an exponential family, the family generated by the empirical distribution of $T_1, \ldots, T_N$. So all of the theory developed above applies to maximizing the Monte Carlo log likelihood $l_{N,Y}$. In particular, each problem has its own phase I. Though

the real log likelihood, being the result of a phase I calculation, has no directions of recession, $l_{N,Y}$ will have one if $T_1, \ldots, T_N$ is not a large enough sample so that the set $\{T_i - T\}$ of tangent vectors of the Monte Carlo problem does not generate the tangent cone of the real problem. A phase I calculation is needed to see that N is large enough so that this has not happened.

There are several methods that have been suggested for obtaining approximate maximum likelihood estimates from the Metropolis algorithm or the Gibbs sampler. The method that seems to have received the most study is the stochastic gradient algorithm, also called the Robbins-Munro method [76, 56]. This method uses a very simple optimization algorithm, and consequently requires large Monte Carlo samples. A more complicated algorithm would be to use Newton's method, calculating the gradient and Hessian of the log likelihood by Monte Carlo. This was apparently first suggested by Penttinen [60, p. 65]. The method discussed here is an algorithm that attempts to make the maximum use of the Monte Carlo samples by approximating the whole likelihood function.

6.7.2 The Metropolis Algorithm

Monte Carlo calculations for models in which it is not possible to calculate the probabilities of states, depend on the Metropolis algorithm [54] or its close relative, the Gibbs sampler [30]. Good general introductions to this area of Monte Carlo calculation are found in the textbooks of Ripley [59, pp. 113–116] and Binder and Heermann [15]. The basic idea of both of these methods is to simulate an ergodic Markov chain with stationary transition probabilities $Y_1, Y_2, \ldots$ whose equilibrium distribution is one of the distributions in the model, say for the parameter point ψ. Then, writing $T_i = t(Y_i)$, the stochastic process $T_1, T_2, \ldots$ will also be stationary and ergodic and for any function h

$$\hat{h}_N = \frac{1}{N} \sum_{i=1}^{N} h(T_i) \to E_\psi h(T) \tag{6.17}$$

for almost all sample paths of the process by the ergodic theorem. Since the sample space is discrete, the the starting point Y_1 can be any point of the sample space.

If adjacent values of the process $T_1, T_2, \ldots$ are highly correlated, as is true of both the Metropolis algorithm and the Gibbs sampler, it is preferable to only sample the chain at regularly spaced intervals with spacing Δ making a new process $T_\Delta, T_{2\Delta}, \ldots$

with the same equilibrium distribution but less correlation. There is, however, no point in making the spacing Δ extremely large in an attempt to reduce the serial correlation to zero. There is a balance to be struck between wasting too much time calculating the function h for nearly identical arguments and wasting too much time calculating steps of the Markov chain. For a discussion of how to choose Δ, see Binder and Heermann [15, pp. 24 and 33–35].

There are many ways of constructing a Markov chain with a specified equilibrium distribution. The most commonly used are the Metropolis and Gibbs sampler schemes, but there is continuously variable class of methods [37] that ranges from Metropolis at one extreme to the Gibbs sampler at the other. For autologistic models like our model for DNA fingerprint data Peskun [58] shows that the Metropolis algorithm is the best, so that is what we describe here.

The Metropolis algorithm for generating the Markov chain works as follows. Select one of the components of the state vector at random (for the DNA fingerprints, this is a random individual i and a random band j). Calculate the odds ratio r comparing the state (0 or 1) this variable is in to the opposite state. If $r \geq 1$ (the opposite state is more probable, conditioning on the rest of the variables, than the current state), change the state of the variable. If $r < 1$, change with probability r. If the current state has $Y_{ij} = 0$, the odds ratio r is $\exp(\theta_{ij})$, where θ_{ij} is given by (6.10), otherwise it is $\exp(-\theta_{ij})$.

Both the Metropolis algorithm and the Gibbs sampler can be used to construct Markov chains having an arbitrary specified equilibrium distribution on an arbitrary sample space. They do not, however, guarantee that the chain is irreducible (that any state can be reached starting at any other state) so that the process is ergodic. A proof of irreducibility is required in each case. For the chain just described, the proof is trivial. The random choice of individuals and bands might visit them in sequence, and at each step the random choice of whether to change that variable or not might always result in a choice to change.

The situation becomes more difficult when a distribution concentrated on a subset of the sample space is simulated. The subset might be the support of the MLE found by the phase I algorithm, or it might be the support of the conditional distribution being used for MCLE. In either case there is no guarantee that changing one variable Y_{ij} at a time produces an irreducible chain. A proof must be devised for each separate

case. It may be necessary to consider a different Markov chain scheme. For example, changing one variable at a time cannot simulate the distribution of the model that conditions on the column sums of the data matrix. It is necessary to change at least two individuals to preserve the constraint. Considering a random pair of individuals in a random column and swapping their values according to the odds ratio of the swapped and current states does, however, produce an ergodic chain.

6.7.3 Exact Estimates

If the number of individuals (rows of Y) is relatively small, exact calculation of the log likelihood and its derivatives are possible, calculating expectations by sums over the whole sample space. This is not practical for most problems, nor is it necessary. The Monte Carlo methods in Section 6.7.1 are preferable except for very small problems. We regard exact calculations mainly as a check on the accuracy of the Monte Carlo.

Even when doing exact calculations, Monte Carlo should be used to find an accurate starting point for the exact calculation so that perhaps only one or two iterations of Newton's method (in constrained problems, sequential quadratic programming) are necessary to obtain the required accuracy. These iterations are performed as follows. The gradient and Hessian of the log likelihood are calculated exactly. These are the score

$$s_l(\theta) = \frac{\partial l_Y(\theta)}{\partial \theta_l} = T_l - \frac{\partial \psi(\theta)}{\partial \theta_l} = T_l - E_\theta(T_l) \tag{6.18}$$

and the negative of the Fisher information

$$-I_{lm}(\theta) = \frac{\partial^2 l_Y(\theta)}{\partial \theta_l \partial \theta_m} = -\frac{\partial^2 \psi(\theta)}{\partial \theta_l \partial \theta_m} = -\operatorname{Cov}_\theta(T_l, T_m). \tag{6.19}$$

Then the local quadratic approximant to the log likelihood is given by the first two terms of the Taylor series

$$l_Y(\theta + \delta) = l_Y(\theta) + s(\theta)^T \delta - \frac{1}{2} \delta^T I(\theta) \delta, \tag{6.20}$$

and the next iterate is the point $\theta + \delta$ at which this function achieves its maximum subject to the constraint that $\theta + \delta$ lie in Θ.

The task of calculating exact expectations is made easier by the fact that the bands (columns of Y) are independent random variables. Let $T^{(j)}$ denote the contribution

to the natural sufficient statistic made by the jth band, i. e.,

$$T^{(j)} = (Y_{1j}, \ldots, Y_{mj}, 0, \ldots, 0, V_j, 0, \ldots, 0, Y_{1j}Y_{2j}, \ldots, Y_{(m-1)j}Y_{mj}).$$

Then then $T^{(1)}$, $\ldots$, $T^{(n)}$ are independent so

$$E_\theta(T) = \sum_{j=1}^{n} E_\theta(T^{(j)}) \qquad \text{and} \qquad \operatorname{Var}_\theta(T) = \sum_{j=1}^{n} \operatorname{Var}_\theta(T^{(j)})$$

where $\operatorname{Var}_\theta(T)$ denotes the matrix with entries $\operatorname{Cov}_\theta(T_l, T_m)$. This reduces the work from order 2^{mn} to order $n2^m$, but this still grows exponentially in the size of the problem.

6.8 Numerical Results

To do computations using the methods we have described, three kinds of optimization software are needed, linear programming for phase I, nonlinear programming for the Monte Carlo phase II calculations, and quadratic programming for exact phase II calculations. We used the packages NPSOL [32] for nonlinear programming, LSSOL [31] for quadratic programming, and MINOS [57] for linear programming. All three are available from the Office of Technology licensing, Stanford University. Slightly modified versions of the first two are also found in the NAG Fortran library; NPSOL is routine E04UCF and LSSOL is routine E04NCF. MINOS is a nonlinear programming package that could have been used for all of the computations, but NPSOL and LSSOL are easier to use. What is required for phase I is a linear programming package that handles large, sparse problems with many linear dependencies among the constraints. Among the packages we had, only MINOS could do this.

6.8.1 Phase I

In phase I we start out by doing the MPLE problem. For the data shown in Figure 6.1 the nonzero components of the directions of recession are shown in Table 6.1. Having found these directions, they are easy to check for the MLE problem, since they only involve four components of the data in each band. The check that the inner product with tangent vectors is always nonpositive need involve only $n2^4$ tangent vectors where $n = 45$ is the number of bands. The check shows that the MLE and MPLE problems have the same directions of recession.

Table 6.1: Directions of Recession of the MPLE and MLE problems for the data shown in Figure 6.1. Only nonzero components are shown. Vectors ϕ_1, ϕ_2, and ϕ_3 are successive directions of recession found in iterations of the phase I algorithm. The four vectors make an orthogonal basis for the space spanned by these components.

	α_7	$\gamma_{7,8}$	$\gamma_{7,9}$	$\gamma_{7,10}$
ϕ_1	-1	0	$+1$	0
ϕ_2	-1	$+1$	-1	$+1$
ϕ_3	0	$+1$	0	-1
ϕ_4	$+1$	$+1$	$+1$	$+1$

The first direction ϕ_1 corresponds to a support of the MPLE or MLE characterized by the condition

$$Y_{7,j} = 1 \quad \text{only if} \quad Y_{9,j} = 1.$$

In conjunction with ϕ_1 the second direction ϕ_2 corresponds to a support characterized by

$$Y_{7,j} = 1 \quad \text{only if} \quad Y_{8,j} = Y_{9,j} = Y_{10,j} = 1. \tag{6.21}$$

With this condition imposed, the third direction ϕ_3 is a direction of constancy, so it implies no additional restriction of the sample space. The additional vector ϕ_4 is the orthogonal complement of the first three in the space spanned by these four components.

For some purposes it does no harm to think of the parameters involved in the directions to have "gone off to infinity," α_7 to $-\infty$ and the three γ's to $+\infty$, but they do so in such a way that the sum of the four variables (that is, $\langle \phi_4, \theta \rangle$) stays finite. It controls the odds ratio of states with $Y_{7,j} = 1$ to states with $Y_{7,j} = 0$.

The result of phase I is that the support of the MLE is the subset of the sample space satisfying (6.21). The three vectors ϕ_1, ϕ_2, and ϕ_3 are all normal to hyperplanes containing the support and hence are directions of constancy in the phase II problem. To get identifiability of parameters in phase II we impose the constraints $\langle \phi_i, \theta \rangle = 0$, $i = 1, 2, 3$. This results in estimates satisfying $\alpha_7 = \gamma_{7,8} = \gamma_{7,9} = \gamma_{7,10}$, but imposing these equalities has effect on the probabilities of the (phase II) model (only the sum $\alpha_7 + \gamma_{7,8} + \gamma_{7,9} + \gamma_{7,10}$ affects probabilities).

Table 6.2: Maximum Likelihood Estimate (α's and γ's). Only the upper triangle of the γ matrix is shown ($\gamma_{1,2}$ in the upper left corner, $\gamma_{12,13}$ in the lower right). Dashes show exact zeros. The three parameters in the middle of the array with values 0.4 ($\gamma_{7,8}$, $\gamma_{7,9}$, and $\gamma_{7,10}$) have gone to $+\infty$ and that α_7 has gone to $-\infty$, but the the sum of the four parameters (1.7) is a finite parameter.

Alpha

0.0	−2.1	−0.5	−2.1	−2.2	−1.4	0.4	−1.2	−2.3	−2.2	−1.4	−2.8	−1.3

Gamma

—	—	—	—	—	—	—	—	—	—	0.2	0.2
	—	—	0.4	2.0	—	0.9	—	1.1	—	—	—
		1.1	—	—	—	—	0.5	—	—	—	—
			1.3	—	—	—	0.5	1.8	—	—	—
				0.5	—	—	—	0.9	0.9	—	—
					—	—	—	—	—	0.9	—
						0.4	0.4	0.4	—	—	—
							—	—	—	—	—
								—	—	—	1.3
									—	—	—
										2.2	—
											1.8

6.8.2 Phase II

Having found directions of recession in phase I, we now must prove that the Markov chain of the Metropolis algorithm is irreducible when we condition on the event (6.21). This is trivial: if the Markov chain is in a state with $Y_{7,j} = Y_{8,j} = Y_{9,j} = Y_{10,j} = 1$, it can go to any other state satisfying (6.21) flipping only one component of Y at a time by flipping $Y_{7,j}$ first (and can get back to this state from any other by the reverse process). So the chain is ergodic and can be used to calculate Monte Carlo estimates.

The estimated γ's in the MLE for the data in Figure 6.1 are shown in Table 6.2, which shows the exact estimates. The Monte Carlo estimates were calculated using a Monte Carlo sample size of 2000 and a Monte Carlo step size of 13 steps per band (each element of the simulated data matrix Y_i was visited once on average between samples and changed or not according to the Metropolis update). This produces a moderate sample size with moderate serial dependence. The parameter values of the Monte Carlo estimate differ from those of the exact estimate by a maximum of

0.066; half of the parameters differ by less than 0.01. Only two of the values shown in Table 6.2 would change if the Monte Carlo estimate were shown instead and then by only one in the last place.

This inaccuracy in the Monte Carlo should be compared with the sampling error. The pseudoinverse of the Fisher information having the same null eigenvectors has diagonal elements corresponding to the γ's with square roots (that is, approximate standard errors of the estimates) ranging from 0.7 to 1.7, about the size of the estimates themselves. (The approximate standard errors for the α's and β's are a bit smaller, 0.3 to 1.1.) Only the eight estimated γ's greater than 1.0 are more than one (approximate) standard error away from zero, and only the two greater than 2.0 are more than two standard errors away. This does not count the three infinite γ's. Their inverse Fisher information is infinite, so the approximation breaks down, but if infinity is considered significantly larger than zero, just five of the 78 dependence parameters are "significantly" nonzero. The inaccuracy from the Monte Carlo is an order of magnitude lower than the sampling error.

The Monte Carlo sample took 1 minute and 25 seconds to generate on a SPARCstation 330 (1-2 megaflops); maximizing the Monte Carlo likelihood took 1 minute and 15 seconds, NPSOL doing 33 evaluations of the objective function and its gradient (the log likelihood and score). This is for the final Monte Carlo calculation using a point somewhere in the vicinity of the MLE as the parameter for the Metropolis algorithm and the starting point for the optimization. It is fast enough so that a moderate amount of bootstrapping would be feasible if desired. The optimization part of the calculation is easily vectorized so that it would run 100-1000 times faster on a vector processor. The Metropolis algorithm is not so easily parallelized, but some study has been given to the feasibility of parallelizing the Gibbs sampler [30]. For comparison, each Newton iteration of the exact calculation took a little over an hour, and each individual added to the data set would double this time.

6.8.3 MLE versus MPLE

Figure 6.2 shows the MLE and MPLE estimates. It is clear from the figure that something is seriously wrong with one or the other. One of them has the β's backwards, estimating high frequencies for the low frequency bands and vice versa. A glance at the data shows that the MLE is correct. From the plot for the γ's we see

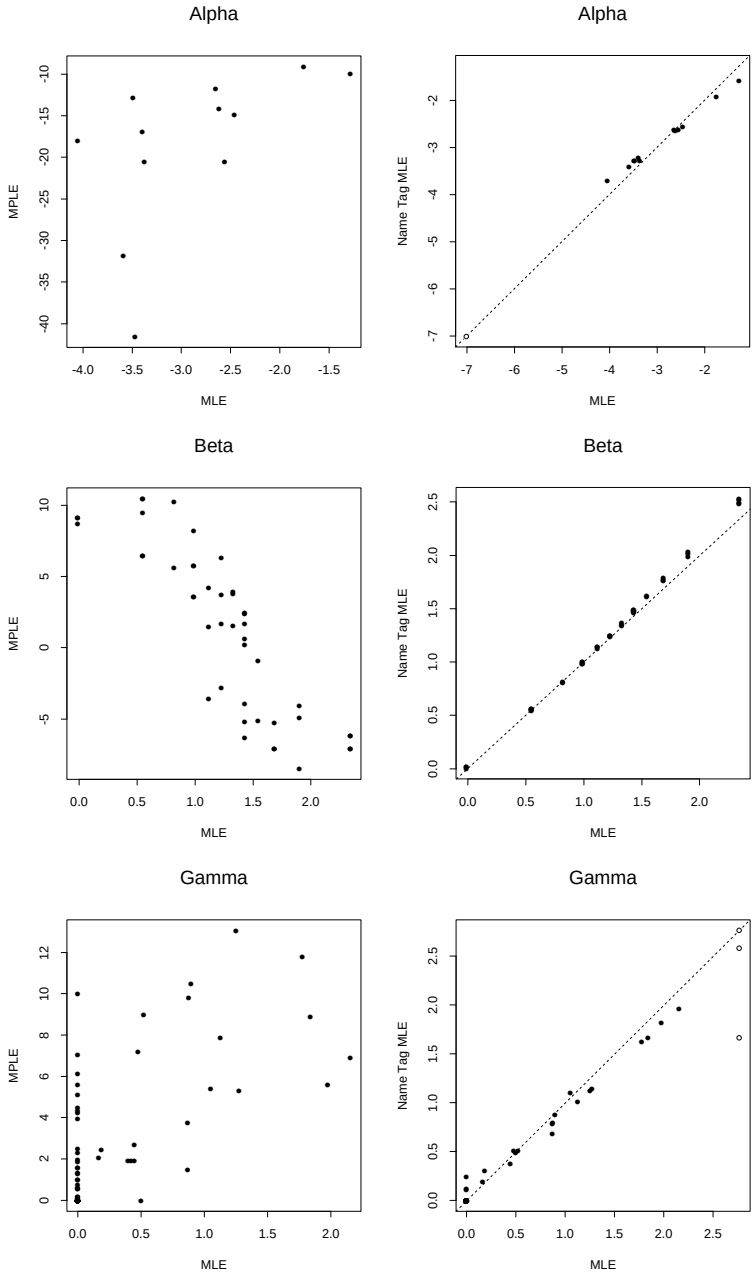


Figure 6.2: Comparison of estimates. Left column, comparison of MPLE and MLE (parameters "at infinity" not shown). Right column, comparison of name tag MLE and MLE (parameters "at infinity" for the MLE are shown as hollow dots whose y-coordinate is the name tag MLE value and x-coordinate is near the left edge for the parameter α_7 at $-\infty$ and near the right edge for the three parameters $\gamma_{7,8}$, $\gamma_{7,9}$, and $\gamma_{7,10}$ at $+\infty$. Dotted lines are the locus of equal x and y values.

that the MPLE estimates much higher values of the dependence parameters than the MLE. Not only does the MPLE have values as high as 10 when the MLE estimates the largest (finite) parameters to be no larger than 2, some of the very large MPLE values are estimated to be zero by MLE.

Moreover, the MPLE induces probabilities that are completely unreasonable. Under the MPLE, the bands have all of the individuals identical (all zeros or all ones) with very high probability. The probability (obtained by exact calculation) that even one of the 45 bands has both zeros and ones is 0.0057. In order that this joint probability be so low, the probability for any single band is very small, at most 0.00104. For all but two of the bands, one of the events all zeros or all ones has probability greater than 0.9989. For the other two, band 35 has all zeros with probability 0.205, all ones with probability 0.794 and anything else with probability 0.00049, and the corresponding numbers for band 38 are 0.533, 0.466, and 0.00104. No bands in the data have all zeros or ones because that would result in an infinite β for that band, and such bands might as well be removed from the data set before the analysis. Because it gets the β's upside down, the MPLE is predicting all ones for bands that have observed frequencies ranging from 1 to 7 out of 13 and is predicting all zeros for bands that have observed frequencies from 5 to 12 out of 13. A global measure of how badly the MPLE fits the data is the log of the likelihood ratio of the MLE and MPLE, which is -3324, a very bad fit.

Another comparison of MLE and MPLE is made in Section 6.9.

6.8.4 Name Tags

Also shown in Figure 6.2 is the comparison of the name tag MLE (NTMLE) with the (real) MLE. Except for the four infinite parameters in the MLE, which have only moderate values in the NTMLE, all of the estimates are quite close, absolute differences being less than 0.35 for the α's, less than 0.25 for the γ's, and 0.20 for the β's.

The log likelihood ratio comparing the MLE and NTMLE is -3.50. It consists of two pieces. If θ is the MLE, D the support of the MLE, ϕ the NTMLE, and Y the observation, the log likelihood ratio is

$$\log \frac{f_\phi(Y)}{f_\theta(Y|D)} = \log \frac{f_\phi(Y|D)}{f_\theta(Y|D)} + \log \Pr_\phi(D),$$

the first term on right hand side being the log likelihood ratio comparing the MLE and NTMLE for the maximum likelihood problem conditioning both on the support D and the second being the probability given the NTMLE ϕ of the data lying in D. The first term, which measures the closeness of the parameters that remain finite is rather small, -0.79, indicating close agreement. The second term, which measures the NTMLE's "distance from infinity," is rather large, -2.71, indicating, at least at first glance, that keeping the parameters finite using name tags makes an appreciable difference in the estimated probabilities. This ratio -2.71 corresponds to $\Pr_\phi(D) = 0.067$; the MLE puts all of its mass on D and the NTMLE less than 7 per cent there. This probability is, however, the product of the probabilities that each of the 45 bands satisfy the condition (6.21). In order that the product be as large as 6.7 percent, the individual bands must have high probability. All are larger than 0.907, and the geometric mean is $0.067^{1/45} = 0.942$. Thus the difference between the MLE and the NTMLE is less than at first appears.

6.8.5 Maximum Conditional Likelihood

The MCLE problem also has directions of recession. Because it is impossible to impose the constraints, directions of recession are possible that send γ's off to $-\infty$. This actually occurs in the data in Figure 6.1. Table 6.3 shows the first direction of recession found by the phase I algorithm. No fewer than 14 of the 78 γ's become $-\infty$. Finite negative γ's would perhaps be acceptable estimates, since MCLE only estimates the differences between the γ's and an inestimable constant. Infinite negative γ's make no sense. So at this point we gave up on MCLE for this data set.

The behavior of the phase I problem is interesting, however, because of what it tells us about the effect of the constraints. If the constraints had not been imposed in the MLE problem, it would have had similar directions of recession. The constraints are what make reasonable parameter estimates possible.

6.8.6 Means and Covariances

In the MLE shown in Table 6.2 (or the NTMLE in shown in Figure 6.2) the dependence parameters (γ's) that are zero do *not* mean that the individuals involved are independent (not genetically related). These parameters only "measure" conditional dependence. In order to interpret the estimates we need some information about the

Table 6.3: Direction of recession for the maximum conditional likelihood problem arising when one conditions on the band totals (V_j).

Alpha Components												
18	−3	18	−18	18	−5	−26	17	4	−26	18	−11	−4
Gamma Components												
	0	0	0	0	0	0	0	0	0	0	0	0
		−23	23	0	23	−23	22	1	21	0	0	−23
			29	0	−6	−14	1	13	0	0	0	0
				0	6	−6	−23	7	29	0	−22	22
					0	0	0	0	0	0	0	0
						1	0	−23	22	−22	22	23
							29	29	29	0	0	−1
								0	1	0	−22	22
									15	−22	22	1
										22	0	−22
											29	−7
												29

actual probabilities induced by the model or about moments of statistics. These can be easily calculated by Monte Carlo using the same sample from which the MLE was calculated.

Figure 6.3 shows the expectation of Y_{ij} as a function of β_j for each of the 13 individuals. The expectations for the actual bands in the data lie somewhere along the curve, but these points are less interesting than the shape of the whole curve. Individuals 1 and 3, which, as can be seen from the estimates, have only low relatedness to the others, are clearly atypical, having shallower curves. The other individuals have steeper curves because of their dependence. Individual 7 whose α is $-\infty$ in the MLE has very few bands, his curve lying below the rest, except for bands with population frequencies above 70 per cent, where the curve for 7 crosses the curve for 1. The other individuals (8, 9, and 10) with infinite parameters also have slightly atypical curves, though not as much as might be expected from their association with infinite parameter values.

The (unconditional) dependence structure of the data is shown most clearly by looking the covariances of Y_{ij} and Y_{kj} for a typical band, one with expected population frequency of about 50 per cent. From Figure 6.3 this is seen to be about $\beta = 1.25$. These covariances are shown in Table 6.4. Both the estimates of the natural

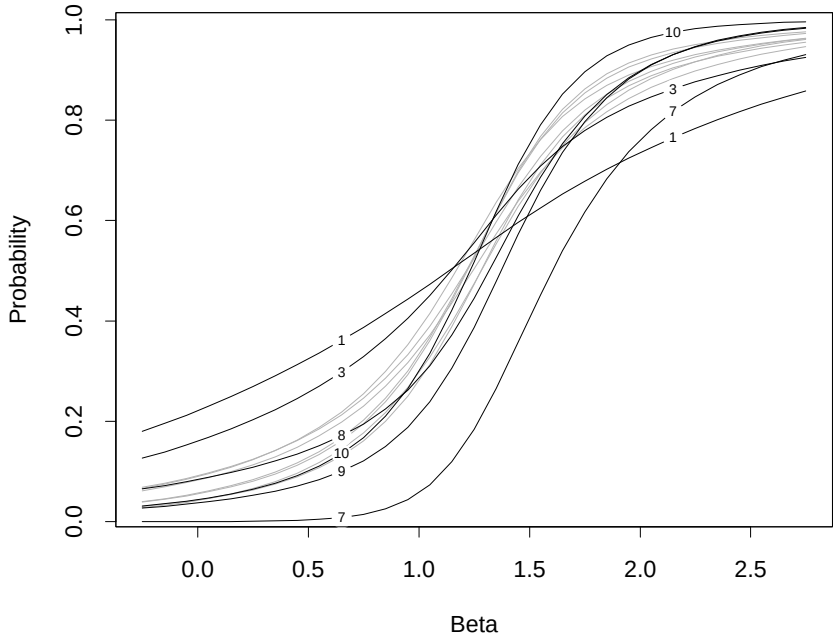


Figure 6.3: Individual versus Population Band Frequencies. Curves are plots of $\Pr(Y_{ij} = 1)$ as a function of β_j with the α's and γ's fixed at their values in the MLE. Black curves are six atypical individuals: 1 and 3 have unusually low relatedness to the rest, and 7–10 are involved in γ's "at infinity." Light grey curves are the other seven individuals.

Table 6.4: Unconditional Covariance. Covariance of Y_{ij} and Y_{kj} for a typical band ($\beta_j = 1.25$).

.003	.001	.002	.003	.004	.003	.002	.005	.003	.008	.015	.016
	.020	.057	.070	.120	.049	.072	.043	.094	.029	.037	.025
		.070	.031	.013	.027	.020	.044	.041	.010	.011	.015
			.103	.039	.055	.045	.066	.127	.029	.026	.027
				.060	.041	.038	.043	.098	.061	.042	.027
					.028	.038	.029	.054	.038	.060	.031
						.102	.113	.089	.018	.024	.037
							.078	.073	.016	.021	.028
								.079	.025	.039	.074
									.031	.030	.032
										.125	.056
											.109

parameters (Table 6.2) and the corresponding covariances (Table 6.4) show the same pattern for very large dependencies. The largest values are in the same locations in both. The small dependencies are quite different. All of the covariances are positive, though the majority of the γ's are exactly zero. A few of the covariances corresponding to zero γ's are fairly large, notably those between individuals 8 and 9 (.078) and between 9 and 10 (.079).

The unconditional covariance is perhaps the closest analogue in our autologistic model of the coefficient of relatedness. In quantitative genetics the covariance structure of the additive component of variance is proportional to this coefficient. By analogy it seems reasonable that the covariance structure (in the model) measures approximately the same thing, though here we have binary, not quantitative, traits.

6.9 A Simulation Study Comparing MLE and MPLE

In order to understand better why MPLE does so badly here, we looked at a much simpler model with only one parameter, and Ising model without external field on a toroidal lattice. This is, like the model for DNA fingerprint data, an autologistic model.

The data form an $n \times n$ matrix Z_{ij}, with each Z_{ij} having a value -1 or $+1$. Each index i and j takes values in $1, \ldots, n$ considered as a cyclic set ($n + 1 = 1$). This "pasting together" of index values 0 and n in both indices makes the index set a toroidal lattice, the set of intersections of a grid on a torus.

Two variables are considered to be "neighbors" if one of their indices differs by one, and the other is the same (counting 1 and n as differing by one). For example, if n=8, the neighbors of Z_{55} are Z_{45}, Z_{65}, Z_{54}, and Z_{56}, and the neighbors of Z_{11} are Z_{81}, Z_{21}, Z_{18}, and Z_{12}.

The sufficient statistic, a scalar, is the sum of like-valued neighbors minus the sum of unlike-valued neighbors

$$t(Z) = \sum_{i=1}^{n} \sum_{j=1}^{n} Z_{ij}(Z_{i(j+1)} + Z_{(i+1)j})$$

(with the convention $n + 1 = 1$ still in force.) The probability density of Z is

$$f_\theta(z) = \frac{1}{c(\theta)} e^{t(z)\theta},$$

and the log likelihood is

$$l_z(\theta) = t(z)\theta - \log c(\theta).$$

Even in this one-parameter model, it is not possible to calculate analytically the Laplace transform $c(\theta)$ of the log likelihood. It is easy, however, to generate samples of the process using the Metropolis algorithm, just as in the more complicated DNA fingerprint model. This makes it possible to calculate c by Monte Carlo, or, what is much more useful in this one-dimensional case, the derivative of $\log c$

$$\mu(\theta) = \frac{d}{d\theta} \log c(\theta) = E_\theta(Z),$$

also called the mean value parameter for the model. Given samples $Z_1, \ldots, Z_N$ from the Metropolis algorithm and $T_i = t(Z_i)$, the Monte Carlo estimate of the mean value parameter $\mu(\theta)$ as a function of θ is given by (6.16). This curve, calculated for a 32×32 torus using a sample of size 1000 spaced one Monte Carlo step apart (the Metropolis algorithm was run with a random scan, each lattice being visited once on average between samples) is shown in Figure 6.4.

Finding the MLE corresponding to an observation T given this curve is simple. One just finds the abscissa $\hat{\theta}$ corresponding to the ordinate $t(Z)$, that is

$$\hat{\theta}(Z) = \mu^{-1}(t(Z)).$$

Finding the MPLE is quite a bit more difficult. One needs to save for each sample not just the sufficient statistic T but the number of Z_{ij} that had 0, 1, 2, 3, or 4 neighbors +1 and how many having each of the 5 neighbor patterns were themselves +1 (the MPLE is a function of a nine-dimensional statistic whereas the MLE is a function of the one-dimensional statistic t). Then calculating the MPLE for each sample requires a logistic regression. This shows that the "calculational simplicity" of the MPLE only applies to calculating one MPLE for a single data set. When doing a simulation study of the performance of the estimators (or a parametric bootstrap) the MLE can be actually easier to calculate.

The MLE's and MPLE's for a sample of size 300 produced by a different run of the Metropolis algorithm from the one that produced the 1000 samples determining the mean value curve (Figure 6.4) are shown in Figure 6.5. It is necessary that the curve be determined by a larger sample so that none of the 300 points for which

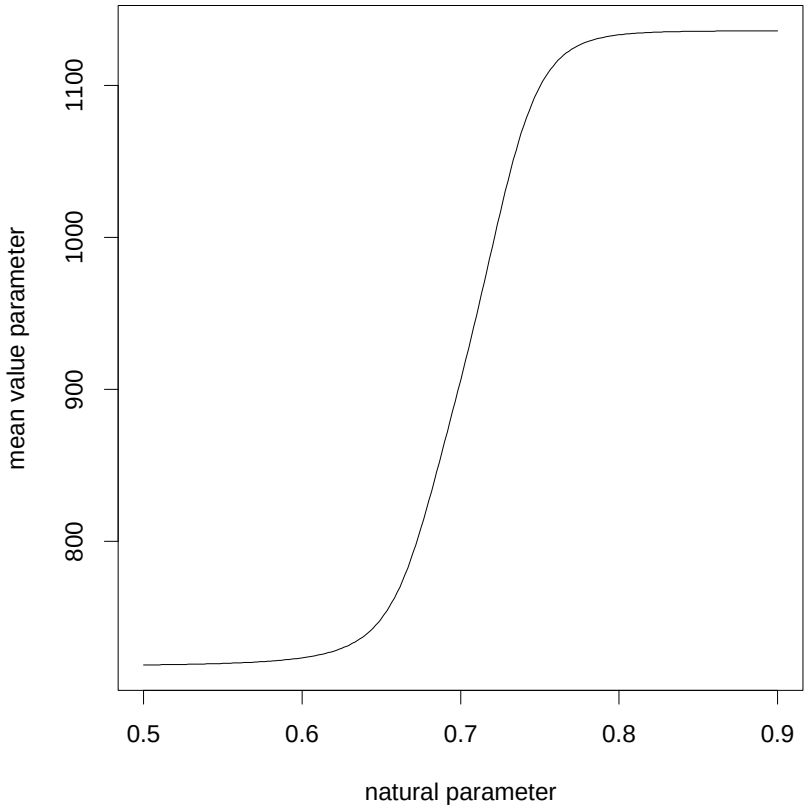


Figure 6.4: Mean Value Parameter as a Function of the Natural Parameter for an Ising Model on a 32×32 Torus.

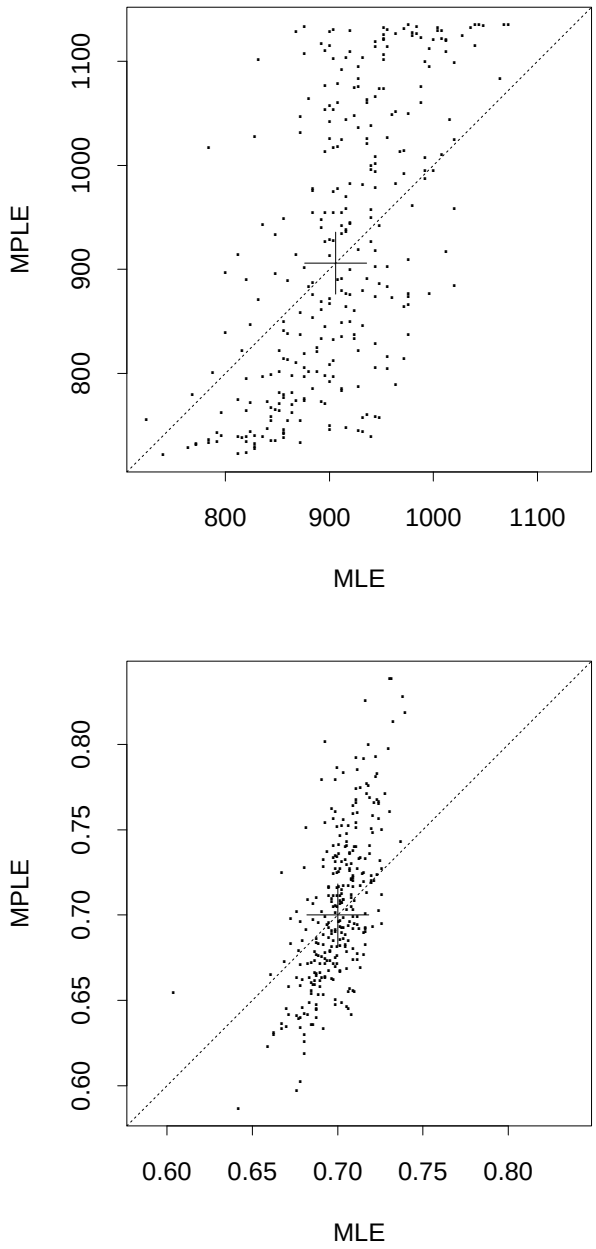


Figure 6.5: Comparison of MLE and MPLE for an Ising Model on a 32×32 Torus. Top panel, scatterplot of estimates of the mean value parameter. Bottom panel, scatterplot of estimates of the natural parameter. Crosses in the center mark the true parameter values, 906.02 for the mean value parameter and 0.7 for the natural parameter. Dotted lines are the lines of equality of estimators. There are 300 estimates in each plot.

MLE are determined lies on the boundary of the "Monte Carlo sample space" and all of the Monte Carlo MLE's exist. In the actual case used here the 300 samples determining MLE's ranged from 724 to 1072 while the 1000 samples determining the curve ranged from 716 to 1136. The estimates can be compared in two ways, on the mean value scale (top panel in the figure) or on the natural scale (bottom panel). The mean values corresponding to MPLE were obtained using the mean value curve of Figure 6.4.

Another view of the same samples is given by Figure 6.6, which shows density estimates of the sampling distributions of the MLE and the MPLE on both scales. As is evident from either figure, the MLE outperforms the MPLE. This is perhaps not surprising. No one ever expected that MPLE would be as efficient as MLE. In the one case where efficiency calculations have been made [14], for Gaussian lattice processes, MPLE performs poorly in some cases. What is interesting, though is the way MPLE performs relative to MLE. When one looks only at the natural parameter scale MLE and MPLE seem about the same, MPLE just has 2.7 times the variance, but both are unimodal and the lumpiness of the MPLE density may be an artifact of undersmoothing (both densities were smoothed with the same bandwidth to make honest comparisons, hence the MPLE density is undersmoothed and the MLE density oversmoothed).

But parameters have meaning only as indices of probability distributions. In particular, Euclidean distance between parameter values has no meaning at all. Probabilities and expectations do have direct meaning in a real problem (this notion is further discussed in Section 6.8.6). By this line of argument, the significant comparison is made on the mean value scale. Here, much as in the DNA fingerprint example, MPLE often produces estimates outside of the range of reasonableness. From the top panel in Figure 6.5 it can be seen that data for the MLE is near the true parameter value can have MPLE's that are at the extremes of the distribution. (A caution is appropriate here. Recall that these are the extremes of the Monte Carlo distribution, the true distribution has a somewhat wider range). The upper extreme corresponds to an Ising model that is "frozen." Almost all realizations from a distribution with such a parameter value would have very high values of the sufficient statistic. The points at the top of the plot in the center (very high MPLE values associated with middling MLE values) would be extremely improbable if the MPLE were correct.

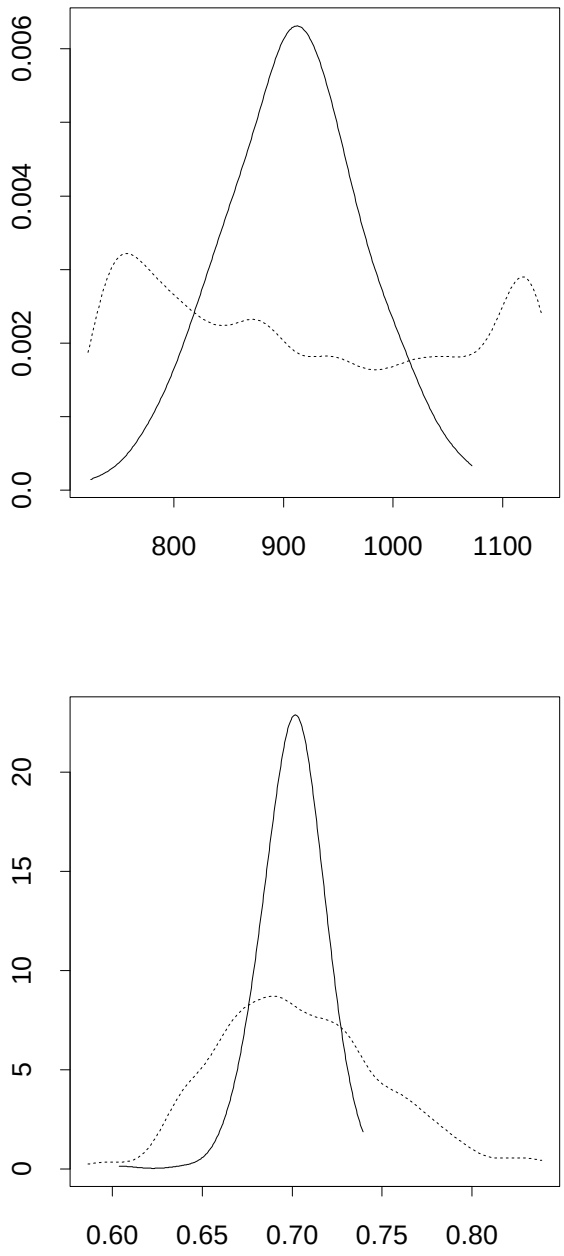


Figure 6.6: Comparison of MLE and MPLE for an Ising Model on a 32×32 Torus. Top panel, density estimate for the mean value parameter. Bottom panel, density estimate for the natural parameter. Solid lines are the density of the MLE, dotted lines the density of the MPLE. Densities were estimated using a gaussian kernel smoother with bandwidth 60 for the mean value parameter and bandwidth 0.025 for the natural parameter.

This is qualitatively the same as the situation for the MPLE and the MLE in the DNA fingerprint problem.

6.10 Discussion

The methods discussed above have many applications besides the one discussed here. The phase I problem arises for any discrete exponential family parametrized by a convex set in the natural parameter space. This includes all of classical discrete multivariate analysis (contingency tables, log-linear models) and many generalized linear models. Monte Carlo maximum likelihood has an even wider range of potential applications, essentially to any area of statistics dealing with dependent data. If some exponential family model seems appropriate, the Monte Carlo methods described here can be used. This includes time series, spatial statistics (lattice processes and point processes), many areas of genetics [67], conditional likelihood inference (conditioning induces dependence even when the unconditional model has independent data, see Section 6.4), in models for social interaction [29, 70], and in expert systems [52]. Also included would be logistic regression with dependent observations [21] or other regressions with dependent observations.

For each of the algorithms, phase I and Monte Carlo maximum likelihood, there is one place where the analysis can become difficult. For the phase I problem this is simply checking whether a vector is or is not a direction of recession, that is, verifying whether (6.14) holds. No algorithm can find a solution faster than it can check a potential solution. For arbitrarily complex geometry of the sample space, the phase I problem can become arbitrarily hard. For most problems, however, it seems that the phase I problem can be solved by available software. Similarly for the Monte Carlo maximum likelihood, finding a Markov chain constructed by the Metropolis algorithm or the Gibbs sampler for which a proof of ergodicity can be constructed may be arbitrarily hard given arbitrarily complex geometry of the sample space. For many problems of interest, however, suitable Markov chains are already well known. For many others, those with sample spaces having simple geometries, construction of ergodic Markov chains would not be difficult.

Chapter 7

CONSTRAINED MAXIMUM LIKELIHOOD EXEMPLIFIED BY CONVEX LOGISTIC REGRESSION

7.1 Introduction

This chapter brings together three ideas: maximum likelihood, nonlinear constrained optimization, and the parametric bootstrap. The first and third are familiar to most statisticians, the second to only a few. Maximum likelihood is by far the most common method of obtaining point estimates, used in least-squares linear regression, ANOVA, and log-linear models, as well as in most generalized linear models—those in which there is a true likelihood rather than just a quasi-likelihood. In many situations in which maximum likelihood is applied there are natural constraints on the parameters, such as nonnegativity constraints or order restrictions, but, except in simple models, these constraints are often ignored for lack of methods for dealing with them. But methods for solving almost all such problems are well known in the field of optimization theory. Software packages such as MINOS [57], first released in 1980, can solve large-scale optimization problems with a smooth objective function and linear or smooth nonlinear equality or inequality constraints.

The availability of estimates in such problems leads to the problem of assessing uncertainty of the estimates, for example to calculating the likelihood ratio test for model comparison. In general constrained maximum likelihood problems the asymptotic distribution of the likelihood ratio statistic is known but cannot be calculated analytically, only by Monte Carlo. So even asymptotic tests are bootstrap tests. This is an example of how consideration of problems with general constraints can change one's attitudes about the practice of statistics. In unconstrained problems bootstrapping is not essential, one can argue that comparing the likelihood ratio to its asymptotic chi-squared distribution (a Wilks test) is good enough. In general constrained problems no such nice asymptotic result obtains. P-values for the likelihood ratio test can be obtained only by the (parametric) bootstrap. This fact makes these methods moderately computer intensive but, given the wide availability of fast

workstations, not excessively so.

7.2 Convex Regression and Constrained Maximum Likelihood

Univariate convex least squares regression is the problem of finding fitted values that minimize the error sum of squares subject to the constraint that the graph of the regression function be convex. If n pairs of predictor and response values (X_i, Y_i) are observed, indexed so that $X_1 \leq X_2 \leq \cdots \leq X_n$, then the problem is to find the value of θ that minimizes the error sum of squares

$$\sum_{i=1}^{n} (Y_i - \theta_i)^2 \tag{7.1}$$

subject to the constraints

$$\theta_i \leq \frac{X_{i+1} - X_i}{X_{i+1} - X_{i-1}} \theta_{i-1} + \frac{X_i - X_{i-1}}{X_{i+1} - X_{i-1}} \theta_{i+1}, \qquad i = 2, \ldots, n-1. \tag{7.2}$$

These constraints require the fitted value θ_i associated with X_i to lie below the line segment connecting (X_{i-1}, θ_{i-1}) and (X_{i+1}, θ_{i+1}). An expression equivalent to (7.2) is

$$\frac{\theta_i - \theta_{i-1}}{X_i - X_{i-1}} \leq \frac{\theta_{i+1} - \theta_i}{X_{i+1} - X_i}, \qquad i = 2, \ldots, n-1, \tag{7.3}$$

which says that the slope of the line segment connecting (X_{i-1}, θ_{i-1}) and (X_i, θ_i) is less than that connecting (X_i, θ_i) and (X_{i+1}, θ_{i+1}). In the problems considered here, the predictor values are always equispaced, in which case (7.2) simplifies to

$$\theta_i \leq \tfrac{1}{2}(\theta_{i-1} + \theta_{i+1}), \qquad i = 2, \ldots, n-1. \tag{7.4}$$

The constraints (7.2) or (7.3) are the discrete versions of the usual characterizations of a convex function f, that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

for all x and y and $0 \leq \lambda \leq 1$ or (for twice differentiable f) that $f''(x) \geq 0$ for all x. Reversing the inequalities gives the definition of a concave function. So any method for convex regression also does concave regression.

Least squares convex regression is a quadratic programming problem, since the objective function (7.1) is quadratic and the constraints are linear (in θ). It can be solved by any quadratic programming algorithm. This problem was apparently first discussed in the statistics literature by Hildreth [38]. He discussed several models in economics in which theory dictates that the regression function be concave and also gave a reasonably efficient algorithm for computing constrained least squares estimates. Other algorithms for the least squares problem have been given by Wu [75], Dykstra [26], and Miller and Sofer [55].

The convex least squares regression problem is a maximum likelihood problem, the error sum of squares (7.1) being proportional to the log likelihood for normal errors. This suggests looking at the general convex regression problem in which any one-dimensional exponential family is substituted for the normal errors. The example of this chapter is convex logistic regression, but the same methods would apply to any such problem. For logistic (binomial) regression the likelihood is

$$\sum_{i=1}^{n} Y_i \log p(\theta_i) + (N_i - Y_i) \log(1 - p(\theta_i)) \tag{7.5}$$

where p is the inverse logit function

$$p(\theta) = \frac{e^\theta}{1 + e^\theta},$$

and Y_i is the binomial response and N_i the sample size for the binomial. The optimization problem is to maximize the likelihood (7.5) subject to the constraints (7.2). Also considered is the problem of isotonic convex regression in which the additional constraint $\theta_1 \le \theta_2$ is added, which with (7.3) implies $\theta_i \le \theta_{i+1}$ for all i so the regression function is nondecreasing.

Since θ is the natural parameter of the binomial exponential family, the likelihood (7.5) is a strictly concave function. Since the constraints are all linear inequality constraints, the set of feasible parameter values, being the intersection of closed half spaces, is a closed convex set. Hence a unique maximum likelihood estimate always exists (perhaps a point in the closure of the exponential family).

General methods of solving large-scale smooth optimization problems have been known since the mid 1970's and have been implemented in software packages such as MINOS and NPSOL that are available from the Systems Optimization Laboratory

at Stanford University [57, 32]. (NPSOL is also available as the function E04UCF in the NAG Fortran Routine Library). These packages are designed to solve any smooth optimization problem with linear and nonlinear equality and inequality constraints. The problems discussed in this chapter were solved by MINOS without difficulty. Good general references on optimization theory are the textbooks by Gill, Murray, and Wright [33], which explains the operation of MINOS, and Fletcher [28].

Very similar to convex regression is the notion of constraining the second differences of the fitted curve. A convex curve is one having nonnegative second differences. Placing bounds on the second differences, i. e., replacing (7.3) by

$$-\lambda_1 \le \frac{\theta_{i+1} - \theta_i}{X_{i+1} - X_i} - \frac{\theta_i - \theta_{i-1}}{X_i - X_{i-1}} \le \lambda_2, \qquad i = 2, \ldots, n-1, \tag{7.6}$$

where λ_1 and λ_2 are nonnegative constants (the smoothing parameters) produces a smoother that operates by maximum likelihood. As with spline smoothing, the maximum smoothing ($\lambda_1 = \lambda_2 = 0$) produces a linear regression, and no smoothing ($\lambda_1 = \lambda_2 = \infty$) fits the data exactly. For intermediate values of the smoothing parameters, the methods behave differently. As with other smoothing methods, the smoothing parameters are not chosen by the method itself (maximum likelihood), but must be taken as constants determined *a priori* or chosen by some other method, such as cross-validation. One advantage of smoothing by constrained maximum likelihood is that it is immediately applicable to any regression problem for which a likelihood can be specified and computed. This illustrates the great flexibility of constrained maximum likelihood. The class of models that come under this rubric is very large.

7.3 The Down's Syndrome Data

Down's syndrome is a genetic disorder caused by trisomy or partial trisomy of chromosome 21 (an extra chromosome 21 or a part of chromosome 21 translocated to another chromosome). The incidence of Down's syndrome is highly dependent on the age of the mother, rising sharply after age thirty. In the 1960's and 70's, as prenatal diagnosis of Down's syndrome via amniocentesis was becoming available, four large scale studies taking data from the British Colombia (B. C.) Health Surveillance Registry for 1961–70 [72], from ascertainment of Down's syndrome cases in Massachusetts (Mass.) for 1950–53 and 1958–66 [43], from vital statistics for upstate

New York (N. Y.) for 1963–74 [41], and from ascertainment of Down's syndrome cases in Sweden for 1971 [44] were done to determine the age-specific incidence of Down's syndrome. The data are shown in Table 7.1.

Two changes have been made to the numbers reported by the original authors. In the B. C. study, mothers age 17 and below and those 46 and above were pooled. To make the data comparable to the other three studies, the distribution of births in these age classes was estimated from the other three studies and the pooled births distributed accordingly. The observed cases were also distributed assuming a constant incidence for age 17 and below and the best-fitting convex regression for age 46 and above (this moves only five cases at the low end and three at the high end). The other change corrects the number of births to mothers age 50 and above in the Mass. study from 25 to 3 to accord with the rest of the data. The authors of the study would presumably not disagree, since one said in a later paper [42], "reports of pregnancies or livebirths to women 50 years and over have been found usually to be erroneous." A plot of number of births in the Mass. study versus age (not shown) convincingly demonstrates that 3 or 4 births is right and that 25 is certainly erroneous.

The data from the four studies are plotted in Figure 7.1. Observed frequencies

Table 7.1: Incidence of Down's syndrome or trisomy 21 by maternal age. Data from five studies of the age-specific incidence of Down's syndrome or pure trisomy 21: (*B. C.*) data from the British Colombia Health Surveillance Registry for 1961–70 [72], (*Mass.*) data from ascertainment of Down's syndrome cases in Massachusetts for 1950–53 and 1958–66 [43], (*N. Y.*) data from vital statistics for upstate New York for 1963–74 [41], (*Sweden*) data from ascertainment of Down's syndrome cases in Sweden for 1971 [44], (*prenatal*) data on fetuses tested by amniocentesis and reported several studies (totals from Hook, et al. [42]). In each column n is the number at risk and x the number observed. For the first four columns n is the number of live births during the study period to mothers of the specified age. (Some of the data for the *B. C.* and *Mass.* differ from that reported by the original authors. See text.) For the *prenatal* column n is the number of fetuses tested prenatally in the reported studies. For the first four columns x is the number of live births recorded as having Down's syndrome phenotype. For the *prenatal* column x is the number of fetuses diagnosed as trisomy 21 (Table next page).

Table 7.1 (cont.)

maternal	*B. C.*		*Mass.*		*N. Y.*		*Sweden*		*prenatal*	
age	x	n	x	n	x	n	x	n	x	n
15	1	1145	1	1364	1	5142	0	383		
16	4	3599	2	3959	4	12524	1	1979		
17	11	8811	10	9848	7	27701	3	5265		
18	15	13675	9	19632	14	51057	10	9212		
19	16	18752	24	32687	16	80075	4	13433		
20	22	22005	26	44376	18	100524	10	17267		
21	16	23896	30	51875	24	115428	16	21133		
22	12	24667	37	54748	32	123769	19	24584		
23	17	24807	46	55757	29	129108	23	26862		
24	22	23986	34	54335	42	128648	19	27747		
25	15	22860	31	52898	40	124718	22	27525		
26	14	21450	30	50181	42	115869	23	25016		
27	27	19202	38	47562	45	105611	16	21694		
28	14	17450	46	44739	33	92588	18	18623		
29	9	15685	42	41901	27	80967	17	15888		
30	12	13954	30	38106	26	70899	11	13835		
31	12	11987	33	34408	25	60403	21	11541		
32	18	10983	39	32116	33	52938	10	9578		
33	13	9825	47	28767	26	45447	16	7861		
34	11	8483	52	25867	30	40212	14	6672		
35	23	7448	54	22947	42	35225	11	5413	44	13525
36	13	6628	70	19605	37	29887	12	4648	55	11694
37	17	5780	72	16707	39	25235	18	3917	54	9450
38	15	4834	68	14006	45	21280	16	3129	70	7062
39	30	3961	50	10986	48	17172	16	2415	57	5298
40	31	2952	96	8586	50	13119	22	1805	65	3883
41	33	2276	74	5729	45	9162	17	1343	44	2323
42	20	1589	51	3961	48	6636	15	845	29	1404
43	16	1018	44	2357	24	4095	6	527	26	732
44	22	596	23	1248	10	2083	13	360	11	433
45	11	327	20	638	15	1111	9	161	7	161
46	4	152	9	258	10	514	7	82	3	67
47	2	60	7	103	3	183	1	35	4	29
48	1	26	2	41	1	65	2	19	3	9
49	0	9	0	13	1	22	0	7	1	5
50	0	2	0	3	1	5	0	2		

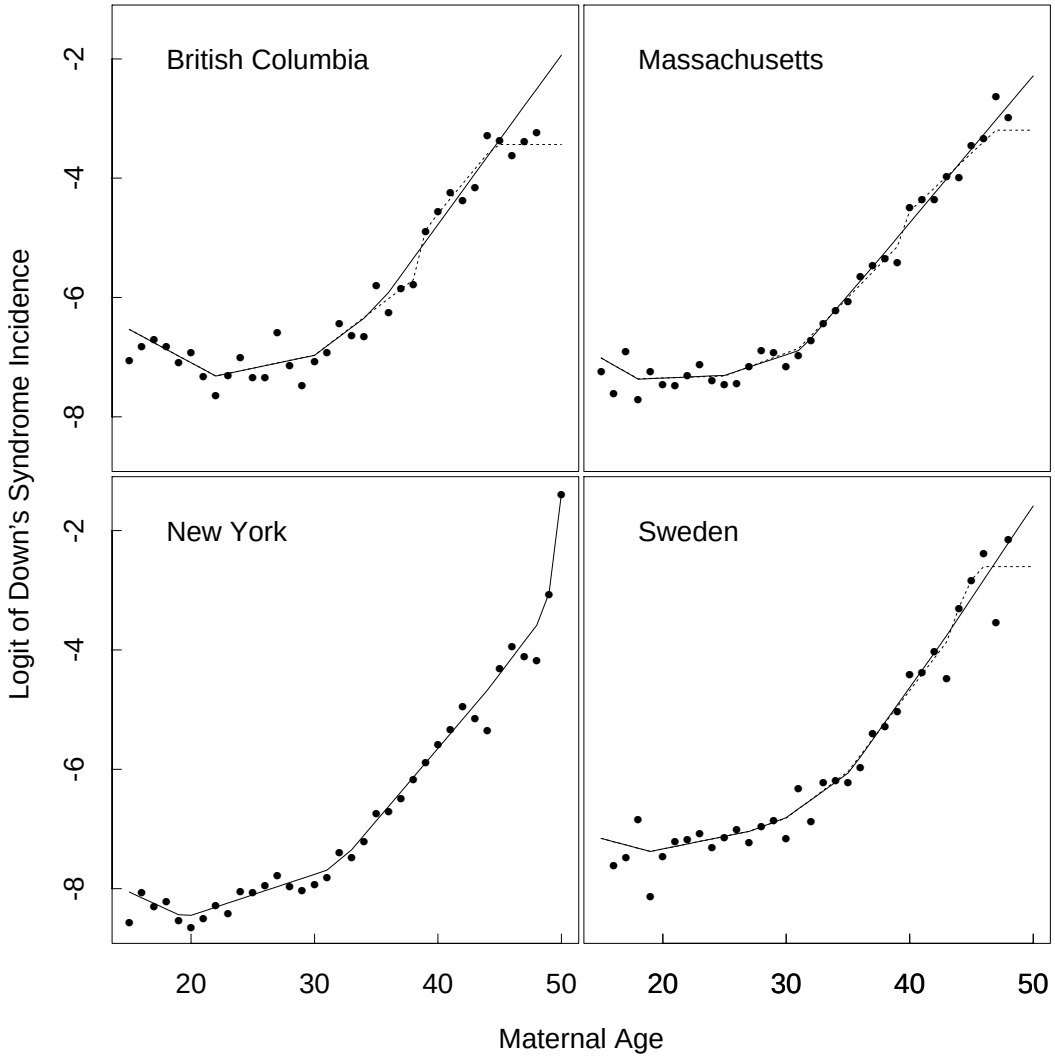


Figure 7.1: Convex and Convex-Concave Regressions. Convex regression for each of the four studies (solid line) and best regression curve that is convex for low ages, concave for high, and nondecreasing on the concave part. Dots show observed rates.

are plotted as points (except for zero frequencies, which are at $-\infty$ on the logit scale of the figure). The solid lines are the maximum likelihood convex regressions. The "curve" here is something of a fiction. Since the data are binned, there is actually only an estimate for each age class. These have been connected with straight lines, so that the linear sections show which constraints are binding (where the convexity hypothesis is determining the shape of the curve). It is not intended that the curve be taken as an estimate of the true curve as a continuous function of age. It could not be, since the uneven distribution of births within the one-year age classes has not been taken into account.

There is no particular biological justification for a convex regression curve. Here convexity is just a very general hypothesis that, perhaps excepting the uppermost end of the curve above age 40, is almost certainly correct. The regression curve obviously continues to bend up through most of its range. The small local oscillations evident to the eye, particularly in the Mass. and N. Y. data, seem random (this was not tested) and would be difficult to explain biologically.

The assumption of convexity is quite unlike assuming a specific functional form, such as a polynomial, for the regression curve. It is more like using a smoother. Like a smoother, convex regression estimates the local conditional expectation of the response given the predictor. It does so, however, without enforcing smoothness. A conventional smoother rounds off the peaks (local maxima) and valleys (local minima) of the regression curve and thus has local bias at such points. Convex regression eliminates peaks while not touching the valleys. If the truth is convex, then convex regression has no local bias. The analogy between convex regression and smoothing is clarified by the preceding discussion about smoothing by constrained maximum likelihood. In that context convex regression is a "one-sided" smoother, the case $\lambda_1 = 0$, $\lambda_2 = \infty$ in (7.6).

Though convex regression seems reasonable, its validity should be checked if possible. The solid curves in Figure 7.1 certainly appear to fit the data well, but since three of the four curves are almost straight after age 35, one might wonder if they would bend down if allowed to do so. It has been suggested that the age-specific incidence might decrease in slope at high ages. This is discussed by Hook, et al. [42] who present evidence against a decrease (column 5 in Table 7.1) involving data (trisomy 21 diagnosed prenatally) similar though not exactly comparable to that analyzed here.

The dotted curves in Figure 7.1 show one alternative to convexity. They are convex up to a transition age chosen by maximum likelihood and concave thereafter. After the transition age (but not before) they are also constrained to be nondecreasing, since the true curve almost certainly does not decrease with age at high ages. Except right at the end, convex-concave curves stay near the convex curves. The curve (N. Y.) having the largest sample size does not bend down at all. If there is any departure from convexity it occurs above age 45, where there is not enough data to test whether the departure is real. In the flat regions of the curves there are 18 cases in the B. C. study, 9 cases in the Mass. study, and 10 cases in the Swedish study. Whether the convex-concave curves fit significantly better was not tested, partly because they did not turn out to be scientifically interesting and partly because their significance or lack thereof depends strongly on whether the number of mothers age 50 and above in the Mass. study is taken to be 3 or 25.

Thus we may take the convex fits in Figure 7.1 as an appropriate "big model" and attempt to find more parsimonious models that fit as well. The next smaller model pools the four studies. Three of the four studies measure the incidence with nearly complete ascertainment. In the N. Y. study ascertainment was very incomplete. Only Down's syndrome cases recorded on birth certificates were ascertained, and completeness of ascertainment was estimated by the authors of the study to be 37.5% (compared to a maximum likelihood estimate from pooling the studies of 39.8%, less than one standard deviation of their estimate away). There is no reason to believe that there was any age-specific bias in ascertainment; therefore it is reasonable to take as a model that the three studies except for N. Y. have the same curve with incidence $p(\theta_i)$ in age class i and the N. Y. study has the same curve shifted by a factor r (the probability of ascertainment) that does not depend on age, giving an incidence $r \cdot p(\theta_i)$ in age class i.

There is another obvious candidate for a pooled model, to take the incidence in the N. Y. to be $p(\theta_i + \rho)$, where ρ is a parameter to be estimated. This model says that the N. Y. curve is parallel to the others on the logit scale rather than on the log scale. There is actually very little difference in the estimates for the two models so there is no statistical reason to prefer one rather than the other.

From a mathematical point of view the second model is simpler. The first is a curved exponential family and the second is a flat one. So only in the second is there

a guarantee that there be only one local maximum of the likelihood (though since the estimates are almost the same one would expect this to be true for both). The second model is a (constrained) generalized linear model and the first a generalized linear model (again constrained) with missing data, the true number of cases, both ascertained and not, in the N. Y. study. Hence if we were doing unconstrained logistic regression, the first model could be fitted using a standard logistic regression routine, but the second would require a specialized method such as the EM algorithm [23] using a logistic regression routine for the M step. When doing constrained regression, however, these considerations are less important. MINOS fits both models with equal ease.

From a scientific point of view the first model is much better. Its parameter r is the scientifically interpretable one, the probability that a case of Down's syndrome is recorded on a birth certificate. The parameter ρ of the second model corresponds to ascertainment that has a particular form of age dependence, which has been built into the model purely for reasons of mathematical convenience against the scientific expectation that the ascertainment bias is not age dependent. Thus we choose the first model. The fits for this model are the solid lines in Figure 7.2, which are compared to the four fits for the separate studies.

The bootstrap distribution for the likelihood ratio test (explained in Section 7.4) of this pooled model versus the model with four separate curves is shown in Figure 7.3. The null hypothesis (of pooling) cannot be rejected ($P = 0.21$) from these data. Hence we go to the next simpler hypothesis, that the regression curve is isotone. In all four studies the fitted curve starts off with incidence decreasing with age, which is hard to explain biologically and may be just an "edge effect." The curve, constrained to be convex, can bend up but not down at each edge. It often will do so, even when the truth is isotone. The isotone and nonisotone curves for the pooled model are compared in Figure 7.4. Again the null hypothesis (of isotonicity) cannot be rejected ($P = 0.17$). The bootstrap distribution for this test is also shown in Figure 7.3. Since it is biologically reasonable that the true regression curve is isotone and convex, and since the data give very weak evidence to the contrary, we may take as the best estimate the isotonic convex regression curve from pooling all four data sets, which is the solid line in Figure 7.4. To give some idea of the variability of this estimate, 198 (parametric) bootstrap replications of the isotonic convex line have been plotted

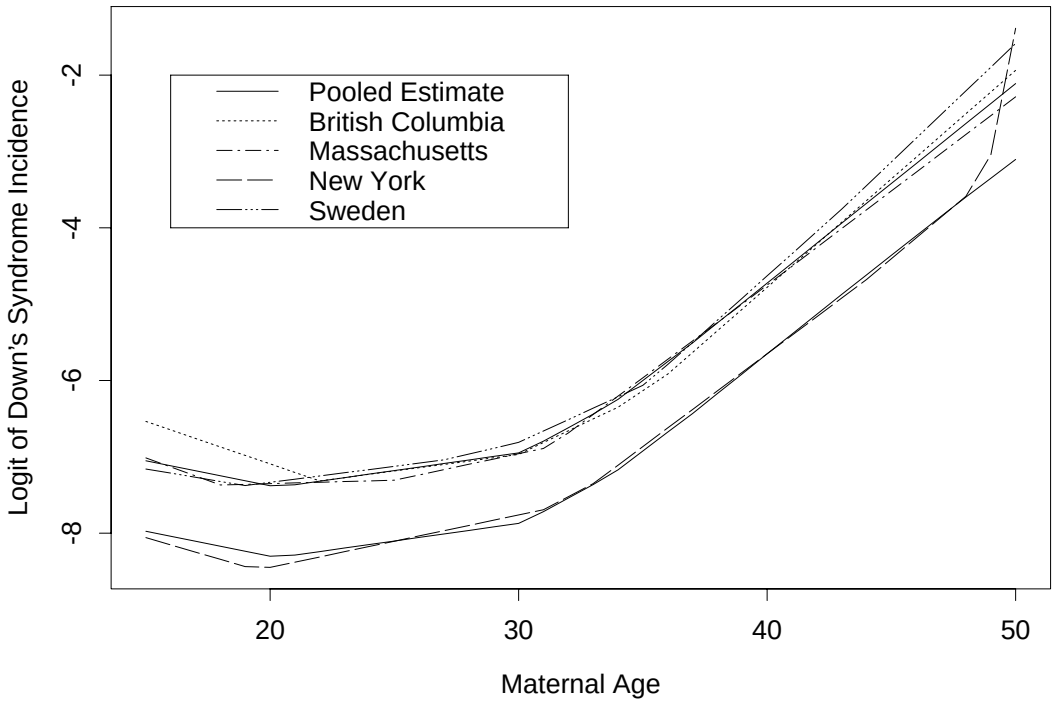


Figure 7.2: Convex Regressions, Separate and Pooled. Convex regressions for each of the four studies and the maximum likelihood estimate determined from pooling the four (solid lines); the upper solid line is the pooled estimate for the three studies excepting N. Y., the lower for N. Y.

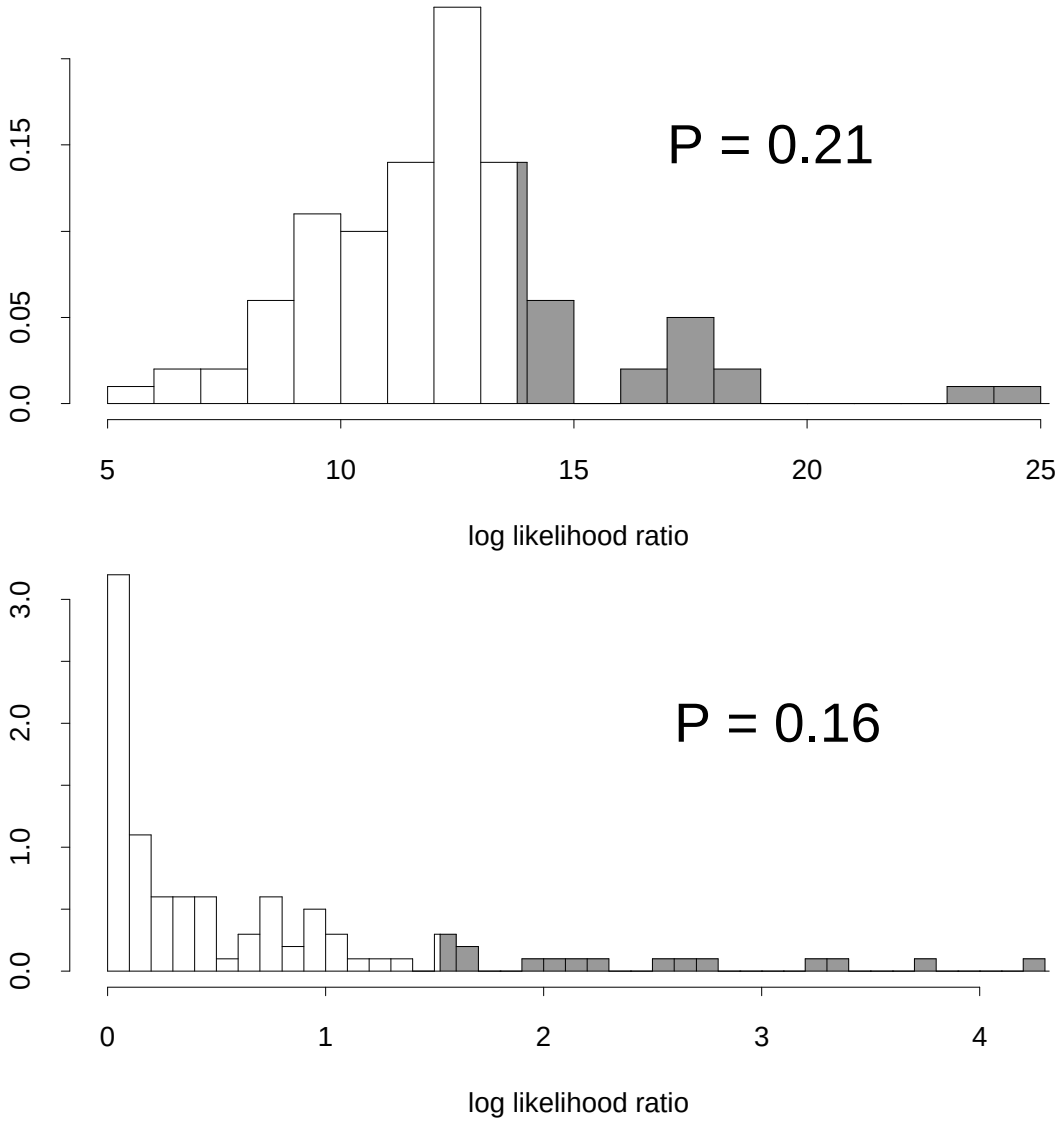


Figure 7.3: Histograms of the Bootstrap Distributions of Likelihood Ratio Tests. Top, test of pooling. Bottom, test of isotonicity. The tail areas more extreme than the observed likelihood ratio are shaded.

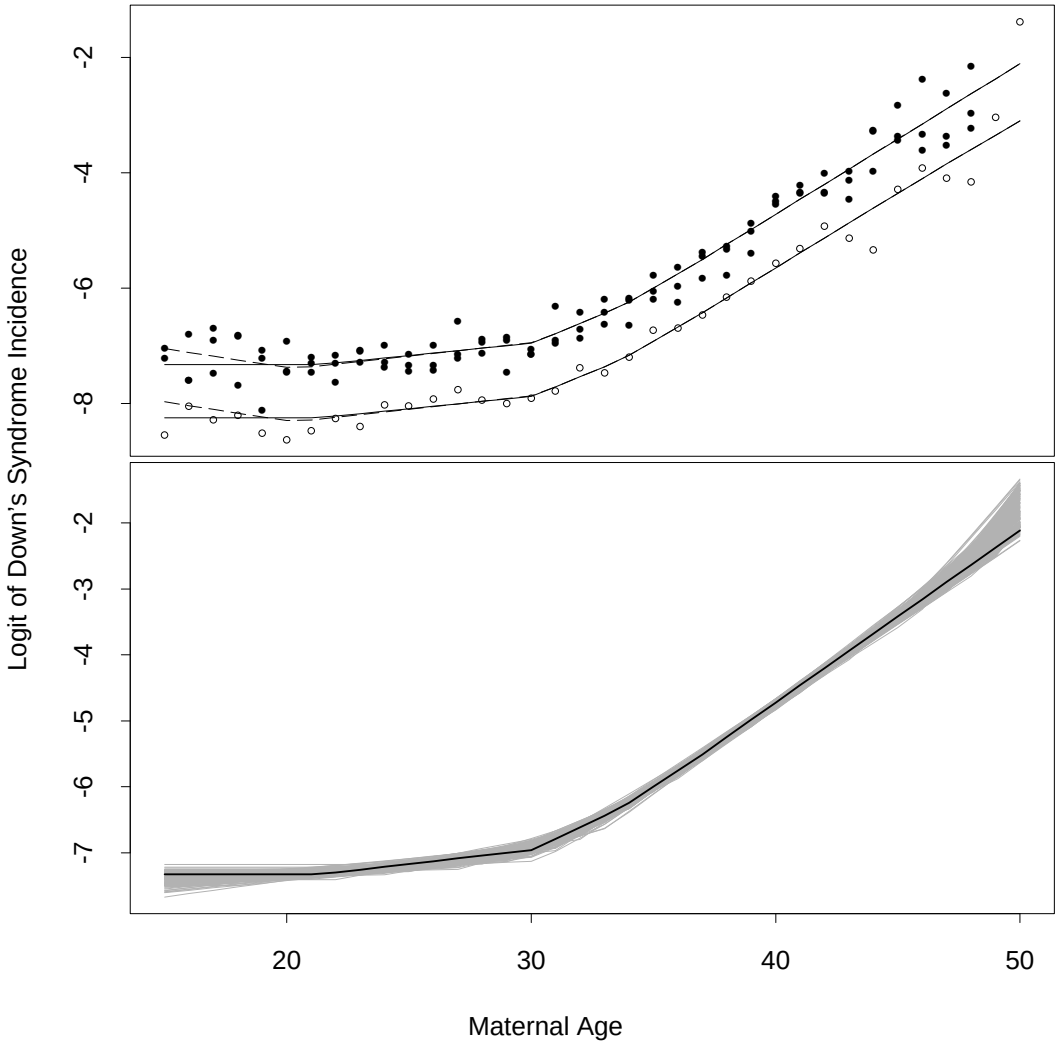


Figure 7.4: Isotone and Non-isotone Regressions. Top panel, comparison of the isotonic convex (solid line) and general convex (dashed line) regression curves. Hollow dots are the observed rates for the N. Y. study; solid dots are the observed rates for the other three studies. Bottom panel, the isotonic convex estimate (solid line) together with the overlapping plots of the isotonic convex estimates for 198 bootstrap replications of the data (gray region).

in the lower panel of Figure 7.4.

7.4 The Bootstrap, Iterated Bootstrap, and Asymptotics for the Likelihood Ratio

The likelihood ratio test is the natural procedure for testing hypotheses involving maximum likelihood estimates. Each estimate maximizes an objective function, the likelihood, subject to constraints, and the natural measure of "better fit" is the increase in the objective function when constraints are relaxed or parameters added.

For models as complex as those considered here, the parametric bootstrap [27] seems to be the only method of calculating P-values. Although the asymptotic distribution of the likelihood ratio statistic is known [20, 66], it can only be calculated by Monte Carlo in constrained problems without special symmetries, including this one. Thus even use of asymptotics requires a parametric bootstrap, albeit a somewhat simplified one.

Given a test statistic T with distribution function $H(\cdot, \theta)$, we want to know $P = 1 - H(T, \theta)$, the probability given θ of T having a value larger than the observed value when θ is the true value of the parameter. We take this to be the P-value of the test. The conventional (Neyman-Pearson) definition of a hypothesis test would require us to take the supremum of this quantity over all values of θ in the (compound) null hypothesis, but this would be extremely difficult to estimate. Moreover, as is the case in the tests described in this chapter, this would be inappropriate when a conclusion to "accept" the null hypothesis is drawn from the failure of the P-value to be anywhere near significance. Taking the supremum makes it easier to reach this conclusion, not taking the supremum is conservative here.

For a fixed value of θ the value of P can be calculated to any desired precision by Monte Carlo integration. The bootstrap idea is to substitute an estimate $\hat{\theta}$ for the unknown true value of θ giving $\tilde{P} = 1 - H(T, \hat{\theta})$. This formula is a compact notation for the following procedure. Estimate $\hat{\theta}$. Simulate from the probability distribution indexed by $\hat{\theta}$ independent samples $Y_1, \ldots, Y_n$. Fit the large and small models and calculate the value of the test statistic for each. Now, assuming that $\hat{\theta}$ is the truth, the actual data Y has the same distribution as the simulations Y_i so

$$\tilde{P} = \frac{\#\{\, T(Y_i) \geq T(Y) : i = 1, \ldots, n \,\} + 1}{n + 1} \tag{7.7}$$

gives the fraction of the $n+1$ independent identically distributed values of T that exceed the observed value. (Here # denotes the cardinality of a set.) Including the observed data in the numerator and denominator of (7.7) is standard procedure for Monte Carlo tests [59, p. 171]. The P-values in Figure 7.3 were calculated according to (7.7).

If the distribution $H(\cdot, \theta)$ of T does not actually depend on θ, rejecting the null hypothesis when $\tilde{P} \leq \alpha$ gives a correct (randomized) test of size α when α is a multiple of $1/n$. Increasing the number n of bootstrap replications while keeping the data fixed will make the value of $\tilde{P}$ depend less on the randomness in the bootstrap samples Y_i and increase the power of the test. For very large n the estimate $\tilde{P}$ will be almost exactly $1 - H(T, \hat{\theta})$, but, unless $H(\cdot, \theta)$ does not depend on θ, it will not be the "true" value P. There is no point in increasing n indefinitely.

To get closer to the true answer it is necessary to iterate the bootstrap [11]. Observe that although our original intention was to conduct a hypothesis test based on the test statistic T, the bootstrap heuristic has led us to doing something else. We now accept the alternative when $\tilde{P}$ is low, that is when $\tilde{T} = 1 - \tilde{P}$ is high. So we are actually conducting a test based on the statistic $\tilde{T}$. The bootstrap idea is just as applicable to $\tilde{T}$ as to T. For fixed θ the distribution of $\tilde{H}(\cdot, \theta)$ of $\tilde{T}$ can be calculated by Monte Carlo, and the bootstrap estimate of the correct P-value is

$$\tilde{\tilde{P}} = 1 - \tilde{H}(\tilde{T}, \hat{\theta}).$$

This formula is a (very) compact notation for the following procedure. Simulate independent samples $Z_1, \ldots, Z_m$ from the probability distribution indexed by $\hat{\theta}$. For each i obtain an estimate $\hat{\theta}_i$ in the null hypothesis corresponding to the data Z_i. Simulate independent samples $Z_{i1}, \ldots, Z_{in}$ from the probability distribution indexed by $\hat{\theta}_i$. Then calculate for each i

$$\tilde{P}_i = \frac{\#\{\, T(Z_{ij}) \geq T(Z_i) : j = 1, \ldots, n \,\} + 1}{n+1}. \tag{7.8}$$

Then

$$\tilde{\tilde{P}} = \frac{\#\{\, \tilde{P}_i \leq \tilde{P} : i = 1, \ldots, m \,\} + 1}{m+1}, \tag{7.9}$$

where $\tilde{P}$ is defined by the one-stage bootstrap (7.7). This differs slightly from the procedure recommended by Beran, although not in any way that would make a difference for large m and n. In Beran's procedure only one sample is taken from the

distribution indexed by $\hat{\theta}$ which plays the role of both the Y_i and Z_i, and the denominators in (7.7), (7.8), and (7.9) are n and m rather than $n+1$ and $m+1$. Beran recommends $m = n = 1000$ instead of the $m = n = 99$ used here. At those bootstrap sample sizes the difference in formulas would be negligible. The procedure recommended here seems best when one is willing to use a smaller sample size. Note that the amount of variability in $\tilde{\tilde{P}}$ is not enough to lead one to think that either of the tests was actually significant. The binomial sampling standard deviation for $\tilde{\tilde{P}} = .17$ is less than .04.

Beran [11] has shown that, given suitable regularity assumptions about the dependence of $H(\cdot, \theta)$ on θ and on the continuity of the asymptotic distribution of T, the distribution $\tilde{H}(\cdot, \theta)$ will be less dependent on θ (more nearly pivotal) than $H(\cdot, \theta)$. Hence Beran's term "prepivoting" for the substitution of $\tilde{T}$ for T. Beran has shown (under the same regularity assumptions) that if the errors in rejection probability for the test based on T go to zero at a rate of $n^{-k/2}$, then those for the test based on $\tilde{T}$ have rate $n^{-(k+1)/2}$. Beran's regularity conditions do not apply to constrained maximum likelihood (the asymptotic distribution of T is not continuous in θ), but whether or not Beran's argument can be extended to cover the constrained case, the iterated bootstrap still makes sense because it is just the ordinary parametric bootstrap applied to the test statistic $\tilde{T}$.

Perhaps more interesting for practical data analysis is the fact that the calculation of the iterated bootstrap is in effect a simulation study of the behavior of the one-stage bootstrap. If the distribution of T is independent of θ, the distribution of the one-stage bootstrap P-values $\tilde{P}_1, \ldots, \tilde{P}_m$ will be uniform on $[0, 1]$. In this case the iterated bootstrap is unnecessary. If the distribution of T is *almost* independent of θ, the distribution of the $\tilde{P}_i$ will be *almost* uniform, and this will be demonstrated by a quantile-quantile (QQ) plot like Figure 7.5. Then one can have confidence that the one-stage bootstrap provides reasonable answers and the iterated bootstrap improves these answers. When the QQ plot shows strong nonuniformity, then one knows that the dependence of the distribution of T on θ is strong, a one-stage bootstrap is not accurate, and one can only hope that a two-stage bootstrap gives a strong enough correction. To gain more confidence in the answers one must do another round of iteration or reformulate the problem. Thus the iterated bootstrap is both a correction and a diagnostic for the one-stage bootstrap. Figure 7.5 is a QQ plot of

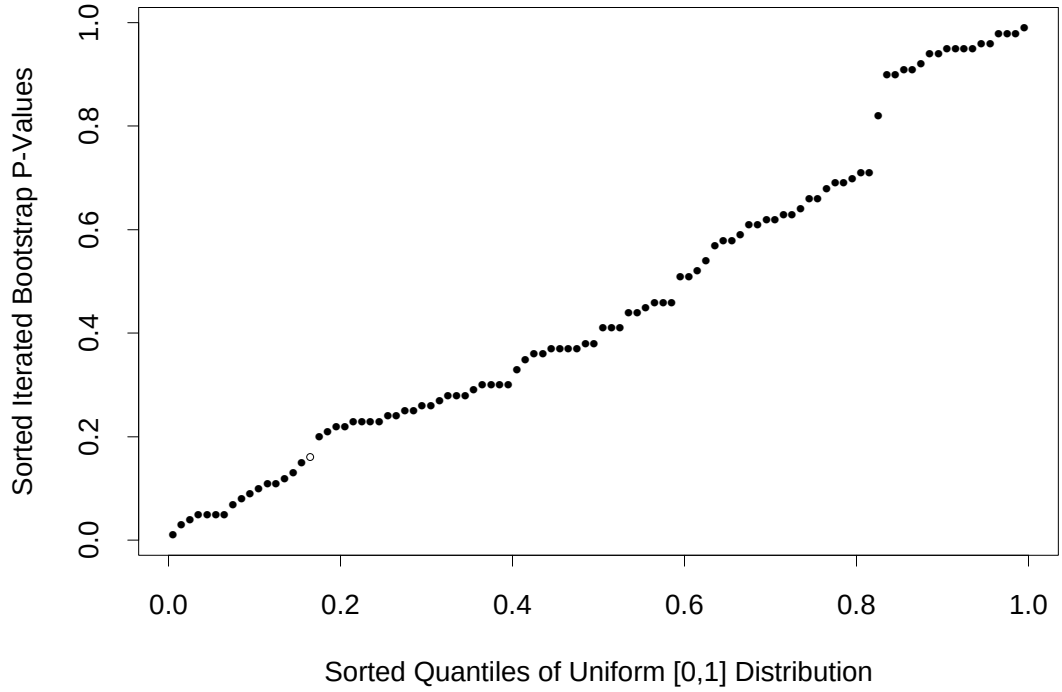


Figure 7.5: QQ Plot of an Iterated Bootstrap. Quantile-quantile plot of the distribution of bootstrap P-values for computed for 99 data sets independently identically distributed from the maximum likelihood distribution in the isotonic convex model (filled dots) and the bootstrap P-value for the actual data (hollow dot).

the distribution of the one-stage bootstrap P-values for the test of isotonicity. The iterated bootstrap only makes a small adjustment in the estimated P-value, from $P = .16$ for the one-stage to $P = .17$ for the two-stage. The QQ plot does show noticeable non-uniformity. Larger P-values would have received stronger correction: $P = .4$ would have gone to $P = .5$. The large jump in the curve at the abscissa .825 is real. The shape of the curve above the jump seems to be determined largely by rounding error. Here very small values of T are being compared. If all of the bootstrap log likelihood ratios, the $T(Z_{ij})$ and $T(Z_i)$ in (7.8), were rounded to three significant figures, thus making small differences exactly zero, the curve would jump to 1.0 at .805. This results from $\tilde{T}$ being zero or nearly zero with a probability of about .20 as can be seen from the histogram in Figure 7.3.

Another important point to notice is that the argument given here for bootstrapping $\tilde{P}$ does not depend in any way on the fact that $\tilde{P}$ is itself a bootstrap estimate. Any other reasonable test statistic could be substituted for $\tilde{P}$. Beran suggests substituting a Bartlett-corrected test statistic when this is easy to calculate (not so for constrained maximum likelihood). Another obvious statistic one might substitute is a P-value obtained from the asymptotic distribution of the likelihood ratio test.

This asymptotic distribution is given by Chernoff [20] and Self and Liang [66]. Under suitable regularity conditions, which are satisfied for logistic regression, the log likelihood ratio converges in distribution to the log likelihood ratio for the *asymptotic problem* defined as follows. Change the sampling distribution of the data from the actual distribution (here binomial) to the multivariate normal distribution with mean at the true parameter value θ and Fisher information equal to the expected Fisher information (at θ) of the actual sampling distribution. Change the null and alternative hypotheses by dropping all constraints that are not active at θ (this replaces the hypotheses by their tangent cones at θ). For this asymptotic result to be valid for convex regression we must imagine the number of of predictor values (age classes for the Down's syndrome data) to remain fixed while the number of observations goes to infinity. It would not be valid if the there were a distinct predictor value for each response value.

To take a simple example, suppose that we want to test whether the B. C. data is isotonic convex (null hypothesis) or just convex (alternative). The asymptotic

problem has a sampling distribution that is normal with mean θ_i at age i and variance

$$\sigma_i^2 = \frac{1}{N_i p(\theta_i)(1 - p(\theta_i))}.$$

The hypotheses for the asymptotic problem are formed by dropping the constraints $\theta_i \leq \frac{1}{2}(\theta_{i-1} + \theta_{i+1})$ for i corresponding to ages 24, 26, 30, 31, 35, and 37 (these are satisfied with strict inequality when θ is the fit for the null model). The remaining constraints divide into two groups, one involving only ages 16–30 and the other involving only 31–50. Since each term of the likelihood involves only one variable, these two groups are decoupled. In the asymptotic model only the estimates for ages 16–30 change when the constraint $\theta_1 \leq \theta_2$ is added; the estimates for ages 31–50 stay the same. Hence the variables for ages 31–50 can be dropped from the asymptotic model; they do not change the likelihood ratio. This leaves a model with 15 variables and 11 or 12 constraints (versus 36 variables and 34 or 35 constraints originally).

A variety of examples showing the types of asymptotic distributions the likelihood ratio statistic can exhibit are given by Self and Liang [66]. There is no closed form solution for the general constrained problem. The distribution for the asymptotic problem can, however, be bootstrapped, just like the actual model. The amount of work done by the computer in such a bootstrap is quite a bit smaller, since normally distributed random variables are roughly ten times easier to simulate than binomial ones and the optimization problem with 15 variables and 12 constraints will also be five to ten times easier than one with 36 variables and 35 constraints, both because of the reduction in the number of variables and because the objective function for the asymptotic problem is quadratic so quadratic programming rather than a general nonlinear optimization algorithm can be used.

For a one-stage bootstrap the asymptotic problem is unappealing. Not only is it necessary to code two different problems—the actual problem and its asymptotic version—but one can't be sure how much to trust the asymptotics. For the iterated bootstrap, however, using the asymptotics for the first stage and the actual problem for the second stage would greatly reduce the work without changing the validity of the result. Doing a one-stage bootstrap for the asymptotic problem is analogous to obtaining a P-value in an unconstrained problem by comparing the deviance to its asymptotic chi-squared distribution (a Wilks test). Doing an iterated bootstrap with asymptotic first stage is analogous to bootstrapping the Wilks test.

S. Self (personal communication, May 1989) has pointed out that between the asymptotic problem and the actual problem there is what might be called the *semi-asymptotic* problem. Use the normal distribution from the asymptotic problem with the constraints from the actual problem. This has no asymptotic justification, but as an approximation to the actual problem should be more accurate than the fully asymptotic problem. It does, however, lose some of the speed-up in bootstrapping gained from going to the asymptotic problem, that due to the reduction in the number of variables.

It should perhaps also be mentioned that similar bootstrap methods are available for obtaining confidence intervals [10, 25, 39]. Confidence intervals and, more generally, confidence regions are more problematical than tests. It is rarely clear what sort of confidence region is of real scientific interest in a multiparameter problem. Furthermore, one is tempted by the illusion of accuracy in the form of the confidence statement to do much more work than for a test. Perhaps simpler methods will suffice, if one has no particular use for an accurate interval. In conducting the iterated bootstrap test of isotonicity, 198 bootstrap samples from the distribution indexed by the isotonic convex estimate were simulated. The corresponding estimates have been plotted (in gray) in Figure 7.4. The region they cover is not, of course, a confidence band for the curve in the strict Neyman-Pearson sense (or in any other sense for that matter). It only says something about the sampling distribution of the estimate at one parameter point, the estimate $\hat{\theta}$. To be a confidence region, it would have to say something about other parameter points, which would be the case only if $\hat{\theta} - \theta$ (the vector of residuals) were a pivotal quantity (having a sampling distribution not depending on θ) which is not the case. The shaded region does, however, give one a useful picture of the variability of the estimates.

7.5 Computing

All of the data analysis for this chapter was done with MINOS which requires that subroutines to calculate the log likelihood and its gradient (the score) be coded in FORTRAN. The binomial random variates for bootstrapping the Down's syndrome data came from an implementation of the recursive binomial generator on p. 537 of Devroye [24] with the design constant t taken to be 10 and the assignment $i = \lfloor (n+1)p \rfloor$ corrected to $i = \lfloor np + 1 \rfloor$. The simple method used for small sample sizes

was Devroye's second waiting time method on page 525. The beta random variates used in the binomial generator were generated from gamma variates (p. 432) and the gamma variates were produced by the constant time algorithm of Cheng and Feast [19] as described in Ripley [59, pp. 90 and 231].

7.6 Discussion

The notion of constrained maximum likelihood is not new. Isotonic regression, which dates back to the 1950's, is maximum likelihood subject to the isotonicity constraint when the response has a distribution in any one-dimensional exponential family [61, p. 34]. Isotonic regression is performed by a comparatively simple algorithm (pool adjacent violators), which does not generalize to more complicated problems. There has been a paper [4] about additive isotonic multivariate models in which pool adjacent violators combined with the so-called "backfitting" algorithm (optimizing over one coordinate at a time) was used for finding estimates. It is not clear that this problem is not better solved by standard methods of quadratic programming as recommended by Robertson et al. [61]. Even in the least-squares case, convex regression requires a much more complicated algorithm (quadratic programming), see Miller and Sofer [55].

Non-Gaussian convex regression requires algorithms, such as MINOS and NPSOL, that are able to do arbitrary smooth nonlinear optimization problems. At this level of difficulty there is an extremely large class of problems, which includes any problem in which the objective function (not necessarily a likelihood, perhaps a quasi-, partial, or pseudolikelihood) and its gradient (the score) and the constraints and their gradients can be easily evaluated. Convex and smooth one-dimensional regression have already been discussed in Section 7.2. Multidimensional convex and smooth regression are conceptually no more difficult, though the number of constraints per data point becomes very large perhaps requiring more advanced algorithms. The class of constrained generalized linear models is no more difficult than the special case (isotonic convex logistic regression) considered here.

Maximum likelihood estimates alone are not enough. Some measure of sampling variability is required. In the theory of generalized linear models, Wilks' theorem gives the asymptotic (chi-squared) distribution of the deviance. Even for the simplest constrained problems, this asymptotic distribution becomes a mixture of chi-squared

distributions but the mixing weights can sometimes be calculated explicitly (see, for example, Robertson, et al. [61, ch. 2] or Self and Liang [66]). In general constrained problems the distribution of the likelihood ratio statistic and even its asymptotic distribution can only be calculated by the parametric bootstrap. The computer time to do such calculations (twenty minutes of CPU time for an ordinary bootstrap and fifteen and a half hours for the iterated bootstrap on a VAX 3500) is extensive but insignificant compared to the thousands of man-hours it took to collect these data.

The methods demonstrated in this chapter provide a sound recipe for data analysis in constrained maximum likelihood problems. Once one has chosen a constrained model, the analysis proceeds along lines familiar from standard procedures for unconstrained models. One obtains maximum likelihood estimates of the parameters for the model or for a nested family of models from MINOS or NPSOL. Then one obtains likelihood ratio tests for model comparison from the parametric bootstrap, perhaps an iterated bootstrap. The computations take a little longer for constrained models, but the statistical principles are the same as for unconstrained ones.

BIBLIOGRAPHY

[1] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71:1–10, 1984.

[2] E. B. Andersen. *Conditional Inference and Models for Measuring.* Mentalhygiejnisk Forlag, Copenhagen, 1973.

[3] H. Attouch. *Variational Convergence for Functions and Operators.* Pitman, 1984.

[4] P. Bacchetti. Additive isotonic models. *Journal of the American Statistical Association* 84:289–294, 1989.

[5] R. R. Bahadur. Sufficiency and statistical decision functions. *Annals of Mathematical Statistics* 25:423–462, 1954.

[6] Ole Barndorff-Nielsen. *Exponential Families: Exact Theory.* Various Publications Series, Number 19, Matematisk Institut, Aarhus Universitet, 1970.

[7] Ole Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* John Wiley, 1978.

[8] Ole Barndorff-Nielsen. Exponential families. In *Encyclopedia of Statistical Sciences.* John Wiley, 1982.

[9] Jean-René Barra. *Notions Fondamentales de Statistique Mathématique.* Bordas Dunod Gauthier-Villars, Paris, 1971. English translation as *Mathematical Basis of Statistics,* Academic Press, 1981.

[10] R. Beran. Prepivoting to reduce level error of confidence sets. *Biometrika* 74:457–68, 1987.

[11] Rudolf Beran. Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* 83:687–697, 1988.

[12] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* 36:192–236, 1974.

[13] J. Besag. Statistical analysis of non-lattice data. *Statistician* 24:179–195, 1975.

[14] J. Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* 64:616–618, 1977.

[15] K. Binder and D. W. Heermann. *Monte Carlo Methods in Statistical Physics*. Springer-Verlag, 1988.

[16] Nicholas Bourbaki. *Algèbre*. Book II of *Éléments de Mathématique*. New edition, Chapters 1–3. Hermann, Paris, 1970. English translation, Addison-Wesley, 1974.

[17] Leo Breiman. *Probability*. Addison-Wesley, 1968.

[18] Lawrence D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Lecture Notes–Monograph Series. Institute of Mathematical Statistics, Hayward, California, 1986.

[19] R. C. H. Cheng and G. M. Feast. Some simple gamma variate generators. *Applied Statistics* 28:290–295, 1979.

[20] H. Chernoff. On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* 25:573–578, 1954.

[21] M. A. Connolly and K.-Y. Liang. Conditional logistic regression models for correlated binary data. *Biometrika* 75:501–506, 1988.

[22] James F. Crow and Motoo Kimura. *An Introduction to Population Genetics Theory*. Harper & Row, 1970.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* 39:1–38, 1977.

[24] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[25] T. J. DiCiccio and J. P. Romano. A review of bootstrap confidence intervals (with discussion). *Journal of the Royal Statistical Society, Series B* 50:338–370, 1988.

[26] R. L. Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association* 78:837–842, 1983.

[27] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7:1–26, 1979.

[28] R. Fletcher. *Practical Methods of Optimization*. Second edition. John Wiley, 1987.

[29] Ove Frank and David Strauss. Markov graphs. *Journal of the American Statistical Association* 81:832–842, 1986.

[30] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–741, 1984.

[31] Philip E. Gill, Sven J. Hammarling, Walter Murray, Michael A. Saunders, and Margaret H. Wright. *User's Guide for LSSOL (Version 1.0): A Fortran Package for Constrained Linear Least-Squares and Convex Quadratic Programming*. Technical Report SOL 86-1, Department of Operations Research, Stanford University, January 1986.

[32] Philip E. Gill, Walter Murray, Michael A. Saunders, and Margaret H. Wright. *User's Guide for NPSOL (Version 4.0): A Fortran Package for Nonlinear Programming*. Technical Report SOL 86-2, Department of Operations Research, Stanford University, January 1986.

[33] Philip E. Gill, Walter Murray, and Margaret E. Wright. *Practical Optimization.* Academic Press, 1981.

[34] Roger Godement. *Cours d'Algèbre.* Hermann, Paris, 1963. English translation as *Algebra,* Hermann, 1968.

[35] Shelby J. Haberman. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Annals of Statistics* 1:617–632, 1973.

[36] Shelby J. Haberman. Maximum likelihood estimates in exponential response models. *Annals of Statistics* 5:815–841, 1977.

[37] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109, 1970.

[38] Clifford Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association* 49:598–619, 1954.

[39] D. V. Hinkley. Bootstrap methods (with discussion). *Journal of the Royal Statistical Society, Series B* 50:321–337 and 355–370.

[40] Kenneth Hoffman and Ray Kunze. *Linear Algebra.* Second edition. Prentice-Hall, 1971.

[41] Ernest B. Hook and Geraldine M. Chambers. Estimated rates of Down syndrome by one year maternal age intervals for mothers aged 20–49 in a New York state study—implications of the risk figures for genetic counseling and cost-benefit analysis of prenatal diagnosis programs. in *Birth Defects: Original Article Series* 13(3A):123–141, 1977.

[42] Ernest B. Hook, Philip K. Cross, and Ronald R. Regal The frequency of 47,+21, 47,+18, and 47,+13 at the uppermost extremes of maternal ages: Results on 56,094 fetuses studied prenatally and comparisons with data on livebirths. *Human Genetics* 68:211–220, 1984.

[43] Ernest B. Hook and Jacqueline J. Fabia. Frequency of Down syndrome in live-births by single-year maternal age interval: Results of a Massachusetts study. *Teratology* 17:223–228, 1978.

[44] Ernest B. Hook and Agneta Lindsjö. Down syndrome in live births by single year maternal age interval in a Swedish study: Comparison with results from a New York state study. *American Journal of Human Genetics* 30:19–27, 1978.

[45] Martin Jacobsen. *Discrete Exponential Families: Deciding When the Maximum Likelihood Estimator Exists and is Unique*. Preprint Number 1. Institute of Mathematical Statistics, University of Copenhagen, 1988.

[46] E. T. Jaynes. Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus, editors, pp. 15–118. MIT Press, 1978.

[47] Alec J. Jeffreys, Victoria Wilson, Swee Lay Thein, David J. Weatherall, and Bruce A. J. Ponder. DNA "fingerprints" and segregation analysis of multiple markers in human pedigrees. *American Journal of Human Genetics* 39:11–24, 1986.

[48] S. Johansen. Homomorphisms and general exponential families. In *Recent Developments in Statistics*, J. R. Barra, F. Brodeau, G. Romier, and B. van Cutsem, editors, pp. 489–499. North-Holland, 1977.

[49] V. L. Klee, Jr. The Structure of Semispaces. *Mathematica Scandanavica* 4:54–61, 1956.

[50] V. L. Klee, Jr. Convex Sets in Linear Spaces. *Duke Mathematical Journal* 18:443-466, 1951.

[51] Steffen L. Lauritzen. *Extremal Families and Systems of Sufficient Statistics*. Springer-Verlag, 1988.

[52] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B* 50:157–224, 1988.

[53] Michael Lynch. Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution*, 5:584–599, 1988.

[54] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087–1092, 1953.

[55] D. R. Miller and A. Sofer. Least-squares regression under convexity and higher-order difference constraints with application to software reliability. In *Advances in Order Restricted Statistical Inference*, R. Dykstra, T. Robertson, and F. T. Wright, editors, Springer-Verlag, 1986.

[56] R. A. Moyeed and A. J. Baddeley. Stochastic approximation of the MLE for a spatial point pattern. Report BS-R8926. Department of Operations Research, Statistics, and System Theory, Centrum voor Wiskunde en Informatica, Amsterdam, October 1989.

[57] Bruce A. Murtagh and Michael A. Saunders. *MINOS 5.1 User's Guide*. Technical Report SOL 83-20R, Department of Operations Research, Stanford University, December 1983, Revised January 1987.

[58] P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika* 60:607–612, 1973.

[59] B. D. Ripley. *Stochastic Simulation*. Wiley, 1987.

[60] B. D. Ripley. *Statistical Inference for Spatial Processes*. Cambridge University Press, 1988.

[61] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. John Wiley, 1988.

[62] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[63] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer-Verlag, forthcoming.

[64] Walter Rudin. *Functional Analysis*. McGraw-Hill, 1973.

[65] Walter Rudin. *Real and Complex Analysis*. Third edition. McGraw-Hill, 1987.

[66] Steven G. Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82:605–610, 1987.

[67] Nuala A. Sheehan. *Image Processing Procedures Applied to the Estimation of Genotypes on Pedigrees*. Technical Report No. 176, Department of Statistics, University of Washington, 1989.

[68] J.-L. Soler. Infinite dimensional exponential type statistical spaces (generalized exponential families). In *Recent Developments in Statistics*, J. R. Barra, F. Brodeau, G. Romier, and B. van Cutsem, editors, pp. 269–284. North-Holland, 1977.

[69] Lynn Arthur Steen and J. Arthur Seebach, Jr. *Counterexamples in Topology*. Second edition. Springer-Verlag, 1978.

[70] David Strauss. A general class of models for interaction. *SIAM Review* 28:513–527, 1986.

[71] Karl R. Stromberg. *Introduction to Classical Real Analysis*. Wadsworth, 1981.

[72] Benjamin K. Trimble and Patricia A. Baird. Maternal age and Down syndrome: Age-specific incidence rates by single-year intervals. *American Journal of Medical Genetics* 2:1–5, 1978.

[73] Albert Verbeek. *The Compactification of Generalized Linear Models*. Prepublication Number 4. Interuniversity Center for Sociological Theory and Methodology, Utrecht, 1989.

[74] R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika 63:27–32, 1976.

[75] Chien-Fu Wu. Some algorithms for concave and isotonic regression. *Studies in the Management Sciences* 19:105–116, 1982.

[76] L. Younes. Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincare. Probabilites et statistiques* 24:269–294, 1988.

Appendix A

MATHEMATICAL BACKGROUND

A.1 Vector Spaces

It is assumed that the reader is familiar with the definition of an abstract vector space found in standard texts on linear algebra, such as Hoffman and Kunze [40]. Let E be a finite-dimensional real vector space. A real linear function on E is called a *linear functional*. The *dual space* of E, denoted E^*, is the space of all linear functionals on E [40, p. 98].

The dual space E^* is a real vector space of the same dimension as E. For f in E^* and x in E the evaluation $f(x)$ is usually written as $\langle f, x\rangle$ or $\langle x, f\rangle$. This emphasizes the the fact that not only is $x \mapsto \langle x, f\rangle$ a linear functional on E (just by definition) but also $f \mapsto \langle x, f\rangle$ is a linear functional on E^*. This is a trivial consequence of the definition of vector addition and scalar multiplication in E^*. Since $f + g$ is the linear functional $x \mapsto f(x) + g(x)$ and αf is the linear functional $x \mapsto \alpha f(x)$, the identity

$$\langle x, \alpha f + \beta g\rangle = \alpha f(x) + \beta g(x) = \alpha\langle x, f\rangle + \beta\langle x, g\rangle$$

holds as an immediate consequence. This is summarized in the statement that the map $(x, f) \mapsto \langle x, f\rangle$ is a *bilinear form* on $E \times E^*$ (a function linear in both arguments [40, p. 166]), the *canonical* bilinear form that places the spaces E and E^* in duality.

Since E and E^* are finite-dimensional, something stronger is true. Every linear functional on E^* is of the form $f \mapsto \langle x, f\rangle$ for some x in E. This means that the dual E^{**} of E^* is E itself, or more precisely the map that associates each point x of E with the linear functional $f \mapsto \langle x, f\rangle$ is an isomorphism from E to E^{**} [40, p. 107].

A.2 Affine Spaces

An affine space is the mathematical structure that generalizes a Euclidean space. It has the geometric properties of a vector space but no distinguished point (an origin).

It is the proper setting for topics like exponential families, in which the origin plays no essential role in the theory.

Let T be a real vector space. An *affine space* associated with T consists of a set E and a binary operation from $T \times E$ to E denoted $(t, x) \mapsto t + x$ satisfying the following three axioms [16, 34].

(A1) For every s and t in T and for every x in E

$$s + (t + x) = (s + t) + x.$$

(A2) For every x in E

$$0 + x = x.$$

(A3) For every x and y in E there exists a unique t in T such that

$$t + y = x.$$

This t is called the *difference* of x and y and is denoted by

$$t = x - y.$$

Any such space E is called a *real affine space* and T is called the space of *translations* of E. The elements of E are called *points.* The translation of a point $x \in E$ by a vector $t \in T$ is denoted by $x + t$ as well as $t + x$.

If one chooses an arbitrary point x_0 of E to serve as an origin, then the map $x \mapsto x - x_0$ is a bijection from E to T that preserves lines and hence all of the geometric properties of the space. A line in E is the set of points $x + \lambda t$, $\lambda \in \mathbb{R}$ for some x in E and nonzero t in T. This maps to the set $x - x_0 + \lambda t$, $\lambda \in \mathbb{R}$, which is a line in T. Thus when a point x_0 of E is distinguished as the "origin" E "becomes" a vector space.

The Euclidean topology of an affine space is the topology carried from its translation space by this bijection. It is the topology generated by the open balls, which are sets of the form

$$x + \epsilon B = \{\, x + \epsilon t : t \in B \,\}$$

where B is the open unit ball of the translation space, x is the center of the ball and ϵ the radius.

Similarly any vector space V has a natural identification with an affine space. Identify the translation space of the affine space with V itself and define the addition of points and translations by vector addition in V which makes sense because both the affine space and its translation space are identified with V. This "forgets" the origin of V in the sense that when V is though of as being the set of points of the affine space the origin plays no distinguished role (though when V is thought of as being the translation space of the affine space, the origin of V is the origin of the translation space).

A.3 Affine and Convex Combinations

It is elementary to show that if $\lambda_1, \ldots, \lambda_n$ is a set of real numbers that sum to 1, and $x_0, x_1, \ldots, x_n$ is a set of points in an affine space E then the point

$$x = x_0 + \sum_{i=1}^{n} \lambda_i (x_i - x_0) \tag{A.1}$$

does not depend on x_0. In this case x is called an *affine combination* of the points $x_1, \ldots, x_n$ and the notation

$$x = \sum_{i=1}^{n} \lambda_i x_i \tag{A.2}$$

is used as an equivalent to (A.1). If, in addition, the real numbers $\lambda_1, \ldots, \lambda_n$ are all nonnegative, (A.1) or (A.2) is called a *convex combination* of $x_1, \ldots, x_n$.

Note that though (A.2) looks like a linear combination (i. e., what the formula would mean if the points $x_1, \ldots, x_n$ were in a vector space and the scalars $\lambda_1, \ldots, \lambda_n$ were arbitrary, being defined by ordinary vector addition and scalar multiplication) it is not. The notion of a "linear combination" has no meaning in an affine space. We must be careful to only use the notation (A.2) when the scalars $\lambda_1, \ldots, \lambda_n$ sum to 1 so that the meaning of (A.2) can be given by reference to (A.1). Of course if the affine space under discussion actually "is" a vector space, then (A.2) agrees with its usual interpretation, since it may be defined by taking $x_0 = 0$ in (A.1).

A.4 Affine and Convex Sets

A subset A of an affine space E is *affine* if it contains every affine combination of its points. By induction this is true if A contains every affine combination of each pair of points, i. e., if

$$\lambda x + (1-\lambda)y \in A, \qquad x, y \in A,\ \lambda \in I\!R.$$

The set A is itself an affine space whose translation space is the subspace $\{\, x - y : x, y \in A \,\}$ of the translation space of E. For this reason, an affine subset A of an affine space E is called an affine subspace of E. This usage should not be confused with the usage "subspace" to denote a vector subspace of a vector space, since a "vector subspace" of an affine space is not a meaningful concept. Affine sets have also been called "flats," "affine varieties," or "linear varieties." None of these synonyms will be used in this dissertation.

A set C of an affine space is *convex* if it contains every convex combination of points in C. Again, it follows by induction that this is true if it only holds for convex combinations of pairs of points, i. e., if

$$\lambda x + (1-\lambda)y \in A, \qquad x, y \in A,\ \lambda \in [0,1].$$

It follows directly from the definitions that the intersection of a family of affine sets is affine and that the union of a nested family of affine sets is affine. This is also true when "affine" is replaced by "convex."

The smallest affine set containing a set S of E is called the *affine hull* of S and is denoted by $\operatorname{aff} S$. The smallest convex set containing S is called the *convex hull* of S and is denoted by $\operatorname{con} S$. Since an affine combination of affine combinations is an affine combination, $\operatorname{aff} S$ can also be characterized as the set of all affine combinations of points of S, and similarly $\operatorname{con} S$ is the set of all convex combinations of of points of S.

Again by induction it follows that if S is convex then $\operatorname{aff} S$ is the set of of all affine combinations of *pairs* of points of S. If x is a point of $\operatorname{aff} S$, it is some affine combination of points in S. Any points with zero coefficients in the affine combination may be deleted and of the rest, if more that two remain, there will be two, say x_1 and x_2 with coefficients of the same sign. But then

$$\lambda_1 x_1 + \lambda_2 x_2 = (\lambda_1 + \lambda_2) z$$

where

$$z = \frac{\lambda_1}{\lambda_1 + \lambda_2} x_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} x_2$$

is a convex combination of points in S and hence lies in S. So x is an affine combination of z and $x_3, \ldots, x_n$, and the number of points in the affine combination can always be reduced until only two remain.

A.5 Affine Functions and the Affine Dual

A function f from a real affine space E to $\mathbb{R}$ is *affine* if it preserves affine combinations, that is, if

$$f\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) = \sum_{i=1}^{n} \lambda_i f(x_i)$$

whenever $\sum_{i=1}^{n} \lambda_i x_i$ is an affine combination of points of E. By induction, it follows that f is affine if

$$f\Big(\lambda x + (1-\lambda) y\Big) = \lambda f(x) + (1-\lambda) f(y), \qquad x, y \in E,\ \lambda \in \mathbb{R}. \tag{A.3}$$

The set $A(E)$ of all real affine functions on an affine space E is called the *affine dual* of E. If E is a finite-dimensional affine space (its dimension being that of its translation space), then $A(E)$ is a vector space of dimension one greater than the dimension of E. Thus the relation of an affine space to its dual is not so simple as the relation of dual vector spaces. The dual of an affine space is not just an affine space but a vector space as well, the zero function being its natural origin, and the dimensions of the spaces are not the same. Nevertheless, the notion of the affine dual is an important one. It simplifies several topics in the theory of exponential families, most notably the notion of the closure of an exponential family.

A.6 The Compactified Real Line

The set $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ is called the *compactified* or *extended* real number system. With the obvious ordering $-\infty < x < +\infty$, $x \in \mathbb{R}$ it is order complete: every nonempty subset has an infimum and supremum. Hence it is compact in the order topology, which is the topology generated by the semi-infinite open intervals of the form $[-\infty, x)$ and $(x, +\infty]$ for x in $\mathbb{R}$. The notation $\overline{\mathbb{R}}$ is natural since $\mathbb{R}$ is a dense subset of $\overline{\mathbb{R}}$ in this topology.

Arithmetic operations in $\overline{I\!R}$ are defined in the obvious way for most arguments. If $x \neq -\infty$, then $x + \infty = \infty$. The operation $\infty - \infty$ is undefined. If $x > 0$, then $x \cdot \infty = \infty$, if $x < 0$, then $x \cdot \infty = -\infty$. But $0 \cdot \infty = 0$. For more on the topology and arithmetic of $\overline{I\!R}$ see Stromberg [71, pp. 27–28] or Rockafellar [62, pp. 23–24].

A.7 Convex Functions

If E and F are affine spaces with translation spaces T and U, then $E \times F$ is an affine space with translation space $T \times U$ when addition and subtraction are defined coordinatewise:

$$(x, y) + (t, u) = (x + t, y + u), \qquad x \in E,\ y \in F,\ t \in T,\ u \in U$$

and

$$(w, y) - (x, z) = (w - x, y - z), \qquad w, x \in E,\ y, z \in F.$$

Following Rockafellar [62] we say that a function f from a real affine space E to $\overline{I\!R}$ is *convex* if its *epigraph*

$$\operatorname{epi} f = \Big\{ (x, t) : x \in C,\ t \in I\!R,\ f(x) \leq t \Big\}$$

is a convex subset of $E \times I\!R$. A function f is *concave* if $-f$ is convex, that is if its *hypograph*

$$\operatorname{hypo} f = \Big\{ (x, t) : x \in C,\ t \in I\!R,\ f(x) \geq t \Big\}$$

is a convex subset of $E \times I\!R$.

For any function f from some set S to $\overline{I\!R}$, the set

$$\operatorname{dom} f = \{ x \in S : f(x) < +\infty \} \tag{A.4}$$

is called the *effective domain* of f. This notion is useful mainly when applied to convex functions (or more generally to functions that are the objective function in a minimization problem), but is defined for general functions so that reference to $\operatorname{dom} f$ can made before the convexity of f is established. There is a dual notion, $\operatorname{dom}(-f)$, which is useful mainly when applied to concave functions (or objective functions in maximization problems). To distinguish the two (A.4) is referred to as the effective

domain in the *epigraphical* sense and the notation $\operatorname{dom}^e f$ is used as an alternative to $\operatorname{dom} f$ when necessary for clarity, and the dual notion is referred to as the effective domain in the *hypographical* sense and the notation

$$\operatorname{dom}^h f = \{\, x \in S : f(x) > -\infty \,\}$$

is used (this notation is taken from Rockafellar and Wets [63]).

Extended-real-valued convex functions can also be characterized by an inequality like that used to characterize real convex functions (this is taken as the definition of convexity in Rockafellar and Wets [63] and the characterization in terms of epigraphs derived as a consequence). A function f is *convex* if and only if

$$f\Big(\lambda x + (1-\lambda)y\Big) \leq \lambda f(x) + (1-\lambda) f(y), \qquad x, y \in \operatorname{dom} f,\ \lambda \in (0,1) \tag{A.5}$$

Equation (A.5) is referred to as the *convexity inequality.* A function f is *strictly* convex if (A.5) is satisfied with strict inequality whenever $x \neq y$. For $z \neq 0$ a convex function f is said to be strictly convex in the direction z if (A.5) with strict inequality whenever $x \neq y$ and $x - y = \lambda z$ for some $\lambda \in \mathbb{R}$.

A convex function is said to be *proper* [62, p. 24] if it is nowhere $-\infty$ and not identically $+\infty$. Otherwise, it is *improper.* A concave function f is proper if $-f$ is a proper convex function.

A convex function is *closed* [62, p. 51–52] if it is proper and lower semi-continuous or if it is improper and constant (i. e., either identically $+\infty$ or identically $-\infty$). A concave function is closed if $-f$ is a closed convex function. This property is extremely important in optimization problems, such as maximum likelihood. Only a closed convex or concave function behaves well.

A.8 Level Sets

For any function f on an affine space E, sets of the form

$$\{\, x \in E : f(x) < t \,\} \qquad \text{and} \qquad \{\, x \in E : f(x) \leq t \,\}, \tag{A.6}$$

for t in $\overline{\mathbb{R}}$ are called the *level sets* of f in the *epigraphical* sense, and sets of the form

$$\{\, x \in E : f(x) > t \,\} \qquad \text{and} \qquad \{\, x \in E : f(x) \geq t \,\}, \tag{A.7}$$

are called the *level sets* of f in the *hypographical* sense.

It follows directly from the convexity inequality and the fact that the union of a nested family of convex sets is convex that the level sets in the epigraphical sense of a convex function are convex sets (Theorem 4.6 in Rockafellar [62]), and hence the level sets in the hypographical sense of a concave function are also convex.

If f is a closed convex function, all of the level sets of the form $\{\, x \in E : f(x) \leq t \,\}$ for t in $\mathbb{R}$ are closed [62, p. 51]. Similarly, if f is a closed concave function, the level sets of the same form with the inequality reversed are closed.

A.9 Relative Interior

The *relative interior* of a convex set is its interior relative to its affine hull [62, Section 6]. The most important fact about relative interiors is that every nonempty convex set (in a finite-dimensional space) has a nonempty relative interior [62, Theorem 6.2]. Thus it is always permissible to assume a point in the relative interior.

A.10 Support Functions

For any set C in a vector space E let σ_C denote the *support function* of the set C [62, Section 13], the function from E^* to $\overline{\mathbb{R}}$ defined by

$$\sigma_C(\theta) = \sup\{\, \langle x, \theta \rangle : x \in C \,\}, \qquad \theta \in E^*.$$

If the set C is closed and convex is actually determined by its support function [62, Theorem 13.1]

$$C = \{\, x \in E : \langle x, \theta \rangle \leq \sigma_C(\theta),\ \forall \theta \in E^* \,\}. \tag{A.8}$$

The relative interior of C is also determined by σ_C [62, Theorem 13.1]

$$\text{r-int}\, C = \{\, x \in E : \langle x, \theta \rangle < \sigma_C(\theta) \text{ or } -\sigma_C(-\theta) = \sigma_C(\theta),\ \forall \theta \in E^* \,\} \tag{A.9}$$

A.11 Faces of a Convex Set

A subset D of a convex set C is a *face* of C if it is a convex extreme set of C. More specifically, D is a face of C if

(a) D is a convex subset of C.

(b) $x, y \in C$, $z \in (x, y)$, and $z \in D$ imply $x, y \in D$.

The following lemma, a slight variation of a theorem given by Rockafellar [62, p. 163], is the characterization of faces of convex sets that is most convenient for working with generalized affine functions.

Lemma A.1 *A subset D of a convex set C is a face of C if and only if*

(a) *D is a convex set,*

(b) *$C \setminus D$ is a convex set, and*

(c) *$D \supset C \cap \operatorname{aff} D$.*

Proof. First suppose (a), (b), and (c), and let x and y be points of C such that some point z in (x, y) lies in D. By (b) $x, y \notin D$ would imply $z \notin D$. Hence one of the points x and y lies in D, but then both lie in $\operatorname{aff} D$, and hence by (c) both lie in D. Thus (a), (b), and (c) imply that D is a face of C.

Conversely, suppose that D is a face of C. Then (a) is true by definition, and (b) is true because $x, y \in C \setminus D$ implies $[x, y] \subset C \setminus D$ because if any point of (x, y) lay in D then this would imply $x, y \in D$. To prove (c) suppose that x is a point in the intersection of C and $\operatorname{aff} D$. Then because D is convex, x is the affine combination of two points y and z in D, say $x = ty + (1 - t)z$ for some $t \in \mathbb{R}$. If $0 \leq t \leq 1$ then $x \in D$ because D is convex. Otherwise, assume without loss of generality $y \in (x, z)$. It then follows that $x \in D$ because $y \in D$ and $x, z \in C$. □

A.12 Directions of Recession of Sets and Functions

Given an arbitrary set C, the *recession cone* of C, denoted $\operatorname{rc} C$ is defined to be $\{0\}$ if $C = \emptyset$ and otherwise the set of points x such that there exists a sequence $\{x_n\}$ in C and a sequence $\{\lambda_n\}$ of positive real numbers decreasing to zero such that $\lambda_n x_n \to x$. This definition is taken from Rockafellar and Wets [63, Section 1E]. (This definition is for sets in vector spaces. For a set C in an affine space, the recession cone is a subset of the translation space, the zero subspace if C is empty, and otherwise the set of translations z such that $\lambda_n(x_n - x) \to z$ where $\{x_n\}$ and $\{\lambda_n\}$ are as defined above and x is an arbitrary point.)

The recession cone is always (as the name suggests) a cone and is always closed. The nonzero elements of $\operatorname{rc} C$ are referred to as *directions of recession* of C.

The following theorem is a fundamental property of directions of recession. It is from Theorems 2C.18 and 2C.19 in Rockafellar and Wets [63].

Theorem A.2 *If C is a convex set, $\operatorname{rc} C$ is a closed convex cone, and for any $y \in \operatorname{rc} C$*

$$x + \lambda y \in \text{r-int}\, C, \qquad \lambda \geq 0,\ x \in \text{r-int}\, C.$$

Furthermore, if C is closed,

$$x + \lambda y \in C, \qquad \lambda \geq 0,\ x \in C$$

holds if and only if $y \in \operatorname{rc} C$.

The *recession function* $\operatorname{rc} f$ of a proper convex function f is given by the formula [62, p. 66] (the notation $\operatorname{rc} f$ is that of Rockafellar and Wets [63])

$$(\operatorname{rc} f)(y) = \lim_{s \to \infty} \frac{f(x + sy) - f(x)}{s} = \sup_{s > 0} \frac{f(x + sy) - f(x)}{s}$$

where x is an arbitrary point of r-int($\operatorname{dom} f$). If f is also lower semicontinuous, then x may be an arbitrary point of $\operatorname{dom} f$. (Note that if the domain of f is an affine space, then the domain of $\operatorname{rc} f$ is the translation space).

The *directions of recession* of a convex function f are the nonzero vectors y such that $(\operatorname{rc} f)(y) \leq 0$ [62, p. 69]. These directions have the following important property, which is Theorem 8.6 in Rockafellar [62].

Theorem A.3 *For a lower semicontinuous convex function f, a nonzero vector y is a direction of recession of f if and only if the function*

$$s \mapsto f(x + sy)$$

is nondecreasing for every x. This is true for every x if

$$\lim_{s \to \infty} f(x + sy) > -\infty$$

for even one x in $\operatorname{dom} f$.

A.13 Set Convergence and Epiconvergence

For a sequence of sets $\{C_n\}$ in a finite-dimensional vector space E, the limit inferior and limit superior of the sequence are defined as follows [63, Definition 3A.1]

$$\liminf_{n\to\infty} C_n = \{\, x \in E : \exists\{x_n\},\ x_n \to x \text{ with } x_n \text{ eventually in } C_n \,\},$$

where "eventually" means for all n greater than some n_0, and

$$\limsup_{n\to\infty} C_n = \{\, x \in E : \exists\{x_n\},\ x_n \to x \text{ with } x_n \text{ frequently in } C_n \,\},$$

where "frequently" means for some $m \geq n$ for all n. A way of saying this in words is that the lim inf is the set of all all points x such that $\{C_n\}$ eventually meets each neighborhood of x, and that the lim sup is the set of all points x such that $\{C_n\}$ frequently meets each neighborhood of x.

If the lim inf and the lim sup are the same, the common value is the limit

$$\lim_{n\to\infty} C_n = \liminf_{n\to\infty} C_n = \limsup_{n\to\infty} C_n.$$

The limit consists of points x such that $\{C_n\}$ eventually meets each neighborhood of x, and the complement of the limit has no points x such that $\{C_n\}$ frequently meets every neighborhood of x.

The lim inf, lim sup, and lim (if it exists) of a sequence of sets are always closed sets [63, Proposition 3A.4] and

$$\liminf_{n\to\infty} C_n \subset \limsup_{n\to\infty} C_n.$$

A sequence of functions $\{f_n\}$ *epiconverges* to a function f if the epigraphs converge in the sense of set convergence, that is, if

$$\lim_{n\to\infty} \operatorname{epi} f_n = \operatorname{epi} f$$

in which case one also writes

$$\operatorname*{e\text{-}lim}_{n\to\infty} f = f$$

or

$$f_n \xrightarrow{e} f.$$

The epi limits inferior and superior are defined similarly by the formulas

$$\operatorname{epi}\Big(\operatorname*{e-lim\,inf}_{n\to\infty} f_n\Big) = \limsup_{n\to\infty}\Big(\operatorname{epi} f_n\Big)$$

and

$$\operatorname{epi}\Big(\operatorname*{e-lim\,sup}_{n\to\infty} f_n\Big) = \liminf_{n\to\infty}\Big(\operatorname{epi} f_n\Big).$$

The reason for reversing the lim inf and lim sup in the definition is so that the relation

$$\operatorname*{e-lim\,inf}_{n\to\infty} f_n \le \operatorname*{e-lim\,sup}_{n\to\infty} f_n$$

will hold (the epigraph of $\text{e-lim inf}_n f_n$ is contained in the epigraph of $\text{e-lim sup}_n f_n$ and hence represents a function that is greater.) The epigraphs of the e-lim inf, e-lim sup and e-lim (if it exists) are all closed sets, hence the functions they represent are all lower semicontinuous.

An alternative characterization of epiconvergence is the following [63, Proposition 3C.2]: $f_n \xrightarrow{e} f$ if and only if for every x both of the following hold

(a) $\forall x_n \to x, \quad \liminf_n f_n(x_n) \ge f(x)$, and

(b) $\exists x_n \to x, \quad \limsup_n f_n(x_n) \le f(x)$.

This is the appropriate notion of convergence of functions for studying minimization problems. The dual notion obtained by substituting hypographs for epigraphs is called hypoconvergence and is the appropriate notion of convergence for studying maximization problems. Together the two notions are referred to as variational convergence. All of the theorems about variational convergence in standard reference works [62, 3] are stated in terms of epiconvergence. The relevant theorems about hypoconvergence are derived by turning the problem upside down (replacing the objective function by its negative so hypographs become epigraphs and replacing maximization with minimization).

VITA

Charles James Geyer III was born April 29, 1946 in Front Royal, Virginia. He grew up in Berwyn, Pennsylvania, a suburb of Philadelphia and graduated from Conestoga High School in Berwyn in 1964. He then went to Hampden-Sydney College, Hampden-Sydney Virginia, left school for service in the U. S. Army in 1967–70, returned and graduated with a B. S. in physics in 1972. After working as an editor and software developer with the W. B. Saunders publishing company, he entered graduate study in statistics at the University of Washington in 1986, received an M. S. in statistics in 1987 and Ph. D in 1990. He will be at the Department of Statistics, University of Chicago on a National Science Foundation Postdoctoral Fellowship in the 1990-91. In the fall of 1991 he will begin as an Assistant Professor in the Department of Theoretical Statistics, University of Minnesota.