Likelihood Inference in Exponential Families and Generic Directions of Recession

Charles J. Geyer

School of Statistics

University of Minnesota

Elizabeth A. Thompson

Department of Statistics

University of Washington

http://www.stat.umn.edu/geyer/gdor/

1

Exponential Families of Distributions

An *exponential family* is a statistical model having log likelihood

$$l(\theta) = \langle y, \theta \rangle - c(\theta)$$

where y is a p-dimensional vector statistic, θ is a p-dimensional vector parameter, and

$$\langle y, \theta \rangle = \sum_{i=1}^{p} y_i \theta_i = y^T \theta = \theta^T y$$

Statistic y and parameter θ that give log likelihood of this form are called *natural*.

Exponential Family Examples

binomial and multinomial distributions

Poisson distribution

geometric and negative binomial distributions

univariate and multivariate normal distributions

exponential and gamma distributions

generalized linear models with above as response distributions

loglinear models for categorical data

aster models (http://www.stat.umn.edu/geyer/aster/)

Exponential Family Theory

Log likelihood for natural parameter is concave.

Can always choose natural parametrization so that log likelihood is strictly concave, in which case maximum likelihood estimate (MLE) is unique if it exists.

Conditions for "usual" asymptotics of maximum likelihood hold if true unknown parameter value is in interior of parameter space.

What is this Talk About?

In exponential families for discrete data, MLE does not always exist in the conventional sense.

When it does not, available software produces nonsense, often with no error or warning. "Usual" asymptotics of MLE are not good approximation to actual sampling distribution. "Usual" hypothesis tests and confidence intervals do not work.

We now have the solution! Old theory (Barndorff-Nielsen, 1978). New software (R contributed package `rcdd`, Geyer 2008, interface to `cddlib` computational geometry package, Fukuda, 2008).

Trendiness?

So isn't this all old seventies stuff? Who cares?

The bandwagon of the oughts (this decade), small n large p, genomics, data mining, model selection, model averaging, etc. is what's trendy now.

In linear models, if you have small n large p, then you have collinearity, and old seventies stuff like ridge regression becomes relevant again, a competitor of LASSO and the like.

In generalized linear models, if you have small n large p, then you have nonexistence of the MLE — what this talk is about!

Binomial Example

x is Binomial(n, p). MLE is $\hat{p} = x/n$.

Natural parameter is

$$\theta = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

When $\hat{p} = 0$ or $\hat{p} = 1$, MLE of the natural parameter $\hat{\theta} = \text{logit}(\hat{p})$ does not exist.

When $\hat{p} = 0$ or $\hat{p} = 1$, distribution for MLE is degenerate.

When $\hat{p} = 0$ or $\hat{p} = 1$, "usual" confidence interval

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

does not work.

Natural Affine Submodels

If a is known vector and M is known matrix (*offset vector* and *model matrix*), change-of-parameter

$$\theta = a + M\beta$$

gives log likelihood

$$l_{\mathsf{sub}}(\beta) = \langle y, a + M\beta \rangle - c(a + M\beta)$$
$$= \langle y, a \rangle + \langle y, M\beta \rangle - c(a + M\beta)$$

Term $\langle y, a \rangle$ does not contain parameters and can be dropped. Also

$$y^T M\beta = \langle y, M\beta \rangle = \langle M^T y, \beta \rangle = \beta^T M^T y$$

where bilinear forms have different dimensions.

Natural Affine Submodels (cont.)

Conclusion: natural affine submodel has log likelihood

$$l_{\mathsf{sub}}(\beta) = \langle M^T y, \beta \rangle - c_{\mathsf{sub}}(\beta)$$

hence is itself exponential family with natural statistic $M^T y$ and natural parameter β.

Exponential family theory covers not only saturated generalized linear model, loglinear model, or aster model but also all (natural affine) submodels too!

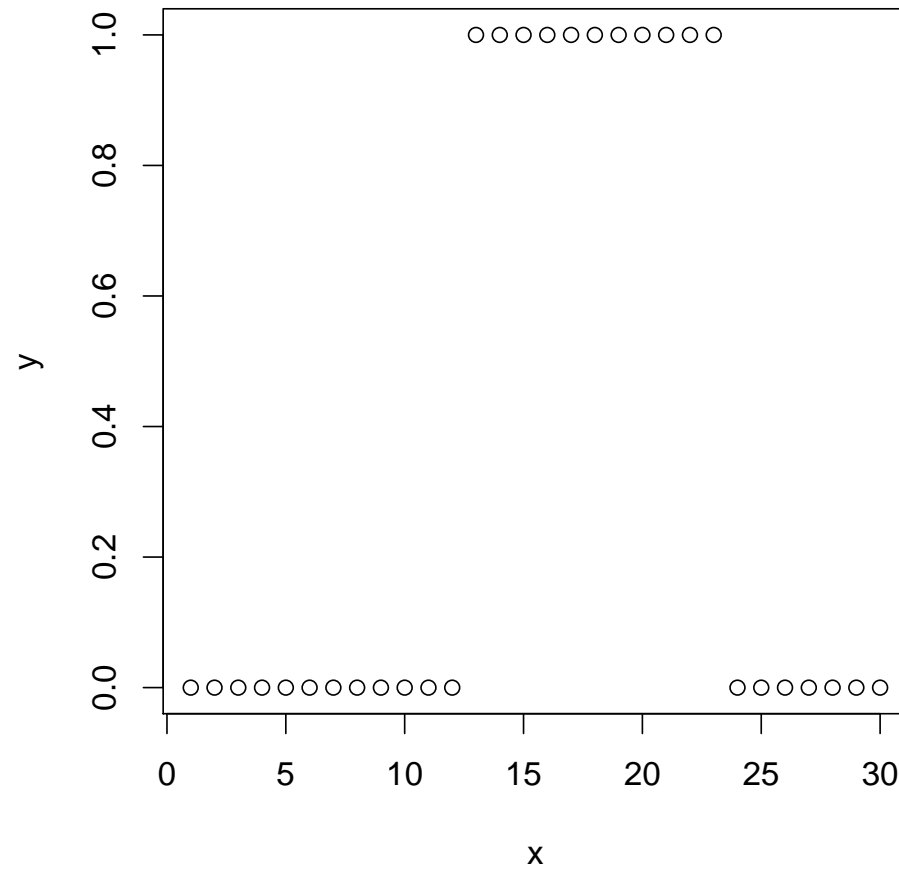From now on we drop the "sub" from l_{sub} and c_{sub}.

Logistic Regression Example

Response vector y, predictor vector x, and y_i is Bernoulli(p_i) with

$$\theta_i = \text{logit}(p_i) = \beta_1 + x_i\beta_2 + x_i^2\beta_3$$

(quadratic logistic regression). How hard can it be?

Logistic Regression Example (cont.)

If data as shown, MLE does not exist!

Generic Directions of Recession

A *generic direction of recession* (GDOR) is a vector δ such that

$$s \mapsto l(\beta + s\delta)$$

is strictly increasing function for each fixed β and there exists $\widehat{\beta}$ such that

$$\lim_{s \to \infty} l(\widehat{\beta} + s\delta) = \sup_{\beta \in \mathbb{R}^q} l(\beta)$$

MLE "is" $\widehat{\beta}$ sent to infinity in the direction δ.

Theorem: MLE does not exist in the conventional sense if and only if a GDOR exists.

Probability density function of distribution with parameter β with respect to distribution with parameter ψ is

$$f_\beta(\omega) = e^{\langle M^T Y(\omega), \beta - \psi \rangle - c(\beta) + c(\psi)}$$

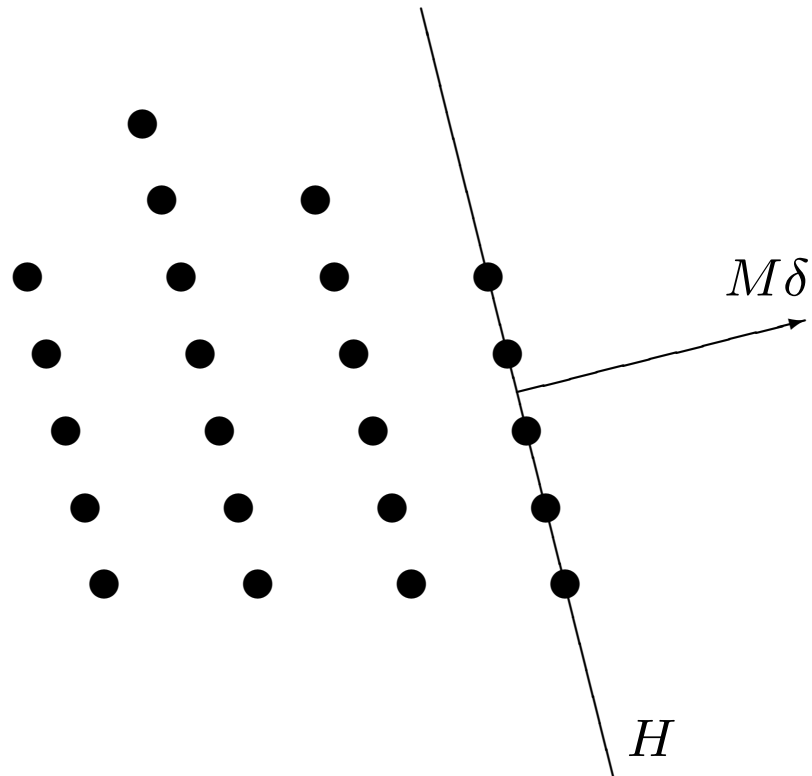If δ is GDOR and y is observed data vector, define

$$H = \{\, w \in \mathbb{R}^p : \langle w - y, M\delta \rangle = 0 \,\}$$

and suppose $\Pr_\beta(Y \in H) > 0$, then

$$\lim_{s \to \infty} f_{\beta + s\delta}(\omega) = \begin{cases} 0, & \langle Y(\omega) - y, M\delta \rangle < 0 \\ f_\beta(\omega)/\Pr_\beta(Y \in H), & \langle Y(\omega) - y, M\delta \rangle = 0 \\ +\infty, & \langle Y(\omega) - y, M\delta \rangle > 0 \end{cases}$$

Right-hand side is $f_\beta(\omega \mid Y \in H)$, and $\langle Y - y, M\delta \rangle \leq 0$ almost surely.

13

Dots are support of original family. If δ is GDOR, then limiting conditional distribution is concentrated on H.

14

Limiting Conditional Model

Limiting conditional model (LCM)

$$\{\, f_\beta(\,\cdot\, \mid Y \in H) : \beta \in \mathbb{R}^q \,\}$$

has log likelihood

$$l_{\text{cond}}(\beta) = l(\beta) - \log \text{Pr}_\beta(Y \in H)$$

hence is exponential family with same natural statistic $M^T y$ and natural parameter β as the original model. Since

$$l_{\text{cond}}(\beta) > l(\beta), \qquad \text{for all } \beta$$

MLE for LCM, if it exists, is MLE in Barndorff-Nielsen completion of exponential family.

Theorem: if δ is GDOR, then MLE always exists in LCM under regularity conditions of Brown (1986) that hold for all practical applications.

Finding a GDOR

```
>  x <- 1:30
>  y <- c(rep(0, 12), rep(1, 11), rep(0, 7))
>  out <- glm(y ~ x + I(x^2), family = binomial, x = TRUE)
Warning messages:
1: In glm.fit(x = X, y = Y, weights = weights, start = start, etasta
  algorithm did not converge
2: In glm.fit(x = X, y = Y, weights = weights, start = start, etasta
  fitted probabilities numerically 0 or 1 occurred
```

This is the `glm` function's somewhat indirect way of suggesting the MLE may not exist.

Finding a GDOR (cont.)

```
>   library(rcdd)
>   tanv <- out$x
>   tanv[y == 1, ] <- (-tanv[y == 1, ])
>   vrep <- cbind(0, 0, tanv)
>   lout <- linearity(vrep, rep = "V")
>   lout
integer(0)
>   p <- ncol(tanv)
>   hrep <- cbind(-vrep, -1)
>   hrep <- rbind(hrep, c(0, 1, rep(0, p), -1))
>   objv <- c(rep(0, p), 1)
>   pout <- lpcdd(hrep, objv, minimize = FALSE)
>   gdor <- pout$primal.solution[1:p]
>   gdor
[1] -53.3636364    6.5454545   -0.1818182
```

Finding a GDOR (cont.)

In the immortal words of a comment in the UNIX source code, "you are not expected to understand this," but the output tells us two things:

```
>  gdor
[1] -53.3636364    6.5454545   -0.1818182
```

gives the GDOR and

```
>  lout
integer(0)
```

says that the LCM fixes all components of the data vector at their observed values.

Directions of Constancy

A *direction of constancy* is a vector δ such that

$$s \mapsto l(\beta + s\delta)$$

is a constant function for each fixed β, in which case β and $\beta + s\delta$ correspond to the same probability distribution for all s.

Only way the MLE of an exponential family can be nonunique.

Theorem: δ is a direction of constancy if and only if $\langle Y, M\delta \rangle$ is almost surely constant.

Logistic Regression Example (cont.)

The LCM for our example is concentrated at one point. Hence $l_{\text{cond}}(\beta)$ is constant function, every direction is a direction of constancy for the LCM, and every vector $\widehat{\beta}$ is an MLE for the LCM and satisfies

$$\lim_{s \to \infty} l(\widehat{\beta} + s\delta) = \sup_{\beta \in \mathbb{R}^3} l(\beta)$$

where δ is the GDOR found by the computer.

Logistic Regression Example (cont.)

So what? The MLE distribution in the LCM is degenerate, concentrated at one point. It says we could never observe data different from what we did observe. Nobody believes that.

The sample is not the population. Estimates are not parameters.

We need confidence intervals, necessarily one-sided, saying how close s is to infinity in $\widehat{\beta} + s\delta$ and how close the corresponding mean value parameters $\mu_i = E_\beta(Y_i)$ are to their observed values.

One-Sided Confidence Intervals

The upper-tailed test with null hypothesis $\widehat{\beta} + s\delta$ and test statistic $\langle Y, M\delta \rangle$ has conservative P-value

$$\mathsf{Pr}_{\widehat{\beta}+s\delta}(\langle Y, M\delta \rangle \geq \langle y, M\delta \rangle) = \mathsf{Pr}_{\widehat{\beta}+s\delta}(Y \in H)$$

where y is observed value of Y.

A one-sided $1 - \alpha$ confidence interval consists of all s such that

$$\mathsf{Pr}_{\widehat{\beta}+s\delta}(Y \in H) \geq \alpha$$

One-Sided Confidence Intervals (cont.)

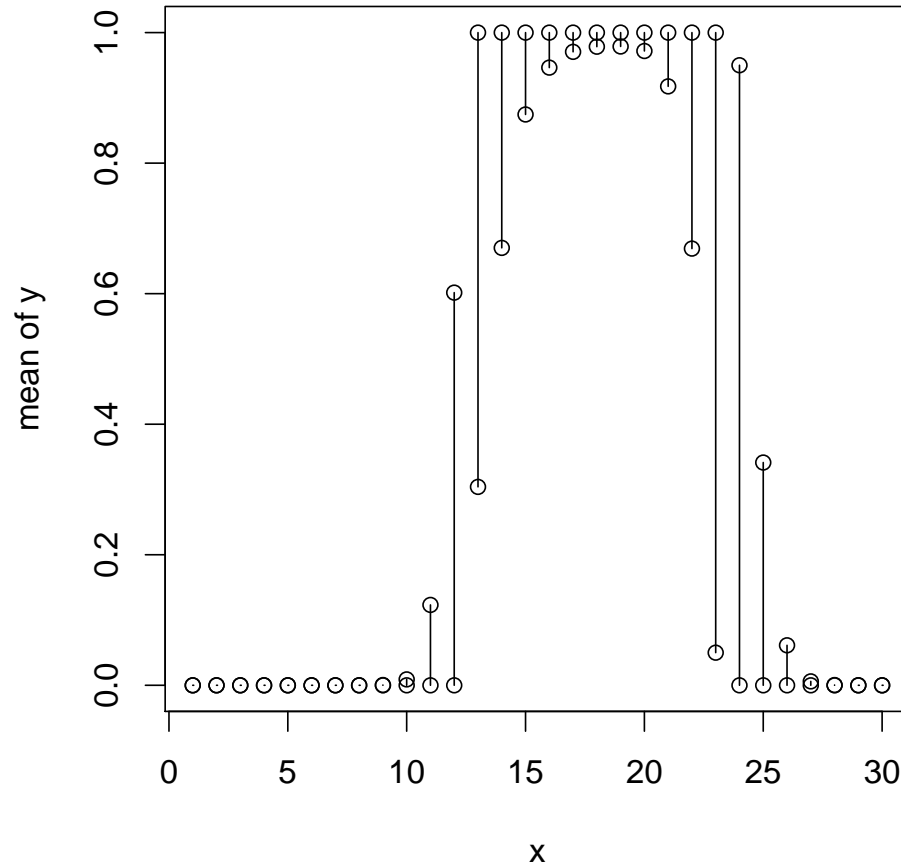Let B denote the set of all MLE for the LCM. In our logistic regression example $B = \mathbb{R}^3$.

For each $\beta \in B$, find the one-sided confidence interval (s_β, ∞) as on the previous slide.

Since all of these confidence intervals use the same test statistic, they have simultaneous coverage

$$\{\,\beta + s\delta : \beta \in B \text{ and } s > s_\beta\,\}$$

is a valid confidence region.

One-sided exact simultaneous 95% confidence intervals for mean value parameters $\mu_i = E_\beta(Y_i)$ made using multiple different $\widehat{\beta}$.

Lessons Learned from Logistic Regression Example

GDOR notion. General.

LCM construction. General.

LCM concentrated at one point so any β is an MLE for the LCM. Not general.

Usually LCM fixes only some, not all components of response vector at observed values. Hence need to find MLE for LCM.

Loglinear Model Example

$2 \times 2 \times \cdots \times 2$ contingency table with seven dimensions hence $2^7 = 128$ cells.

```
> dat <- read.table(
+     url("http://www.stat.umn.edu/geyer/gdor/catrec.txt"),
+     header = TRUE)
```

gets the data.

Treat as GLM. Poisson sampling and multinomial sampling give same MLE (known fact from categorical data analysis).

```
> out2 <- glm(y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^2,
+       family = poisson, data = dat, x = TRUE)
> out3 <- glm(y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3,
+       family = poisson, data = dat, x = TRUE)
> anova(out2, out3, test = "Chisq")
Analysis of Deviance Table

Model 1: y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^2
Model 2: y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        99    191.629
2        64     31.291 35  160.338 5.819e-18
```

27

Loglinear Model Example (cont.)

R function `glm` gives no error or warning, but MLE does not exist for model with formula

```
y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3
```

which is all 3-way interactions but no higher order interactions.

R would produce nonsense if asked for confidence intervals for this model or if asked for hypothesis test with this model as null hypothesis.

Loglinear Model Example (cont.)

```
>  tanv <- out3$x
>  vrep <- cbind(0, 0, tanv)
>  vrep[dat$y > 0, 1] <- 1
>  lout <- linearity(vrep, rep = "V")
>  linear <- dat$y > 0
>  linear[lout] <- TRUE
>  sum(linear)
[1] 112
>  length(linear) - sum(linear)
[1] 16
```

The LCM fixes 16 components of the response vector at their observed value zero and leaves 112 components random.

Loglinear Model Example (cont.)

Fit LCM

```
>  dat.cond <- dat[linear, ]
>  out3.cond <- glm(y ~ (v1 + v2 + v3 + v4 + v5 + v6 + v7)^3,
+        family = poisson, data = dat.cond)
>  summary(out3.cond)
```

Voluminous output not shown (64 regression coefficients!). This fit out3.cond can be used to produce valid hypothesis tests and confidence intervals about the 112 components of the response not fixed in the LCM.

Loglinear Model Example (cont.)

For the 16 components of the response fixed at zero in LCM, proceed as before. Find GDOR.

```
>  p <- ncol(tanv)
>  hrep <- cbind(0, 0, -tanv, 0)
>  hrep[!linear, ncol(hrep)] <- (-1)
>  hrep[linear, 1] <- 1
>  hrep <- rbind(hrep, c(0, 1, rep(0, p), -1))
>  objv <- c(rep(0, p), 1)
>  pout <- lpcdd(hrep, objv, minimize = FALSE)
>  gdor <- pout$primal.solution[1:p]
```

and find one-sided confidence interval for s in $\widehat{\beta} + s\delta$.

Loglinear Model Example (cont.)

One-sided exact simultaneous 95% confidence intervals for mean value parameters $\mu_i = E_\beta(Y_i)$ based on multinomial sampling (not Poisson). Sample size `sum(dat$y)` is 544.

v_1	v_2	v_3	v_4	v_5	v_6	v_7	lower	upper
0	0	0	0	0	0	0	0	0.2855
0	0	0	1	0	0	0	0	0.1404
1	1	0	0	1	0	0	0	0.2194
1	1	0	1	1	0	0	0	0.4198
0	0	0	0	0	1	0	0	0.0892
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	0	1	1	0	1	0	0.2639
0	0	0	0	0	1	1	0	0.0665
0	0	0	1	0	1	1	0	0.1543
1	1	0	0	1	1	1	0	0.1406
1	1	0	1	1	1	1	0	0.3230

Hypothesis Tests

"Usual" hypothesis tests valid if MLE exists in the conventional sense for null hypothesis.

If not, then base test on LCM for null hypothesis (S. Fienberg, personal communication).

Other Uses for RCDD

Exact, infinite-precision, rational arithmetic. Can be used for computational geometry operations, including the `linearity` and `lpcdd` functions, and for ordinary arithmetic and comparison.

When exact arithmetic is used, computer proofs are as rigorous as pencil-and-paper proofs.

See package vignette for everything `rcdd` can do.

Web Pages

http://www.stat.umn.edu/geyer/gdor/

Submitted paper.

Technical report done with `Sweave` so every number in the paper and this talk is reproducible by anyone who has R.

Slides for this talk.

http://www.stat.umn.edu/geyer/aster/

Aster models.