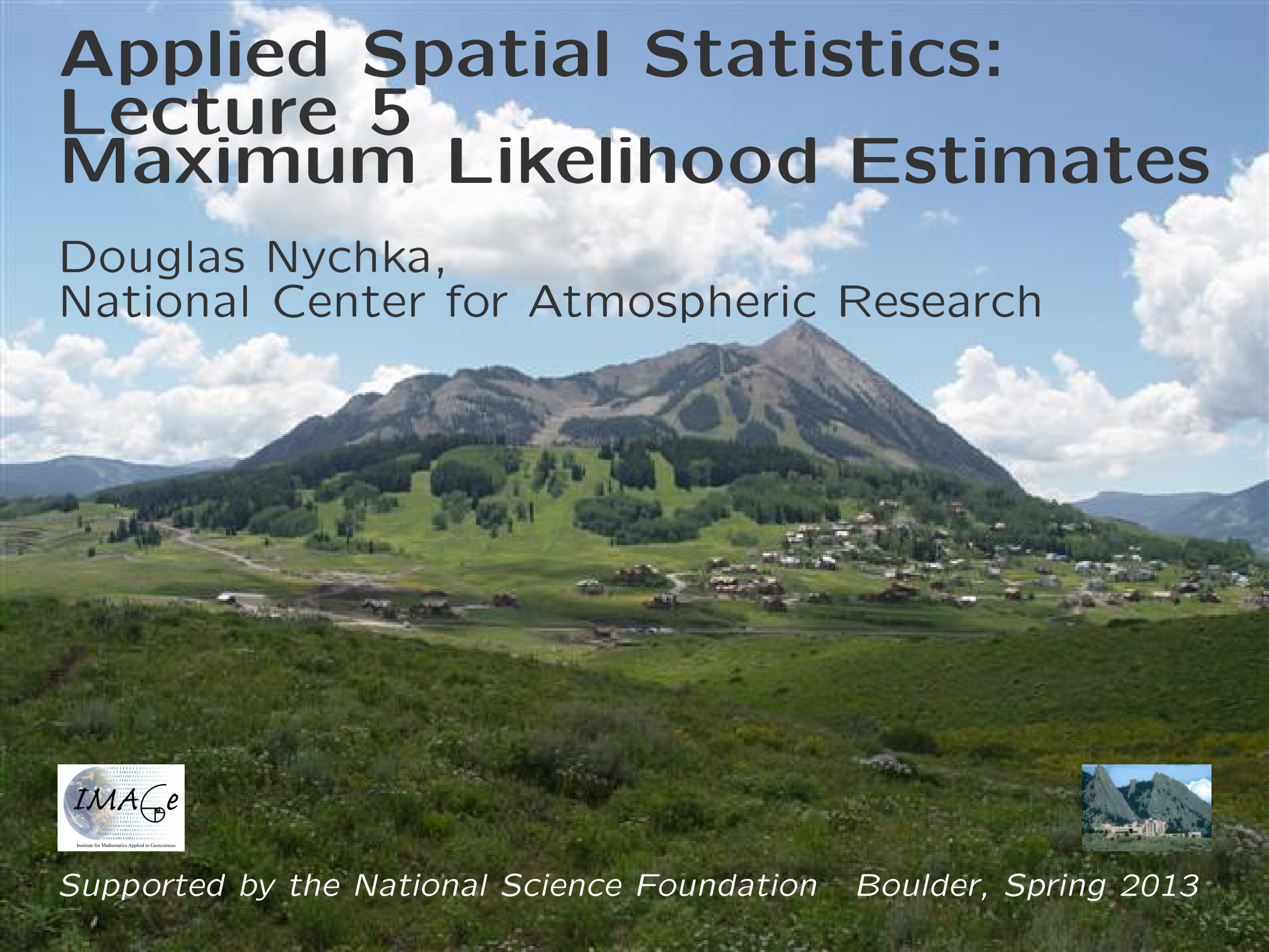
Applied Spatial Statistics: Lecture 5 Maximum Likelihood Estimates

Douglas Nychka,
National Center for Atmospheric Research



Supported by the National Science Foundation Boulder, Spring 2013

Outline

- The likelihood
- Maximization for univariate cases
- Regression
- Inference and profiling.
- Bayes connection

Estimating statistical parameters

Setup:

- A univariate probability density function, $f_\theta(y)$

– depends on parameter θ

- A sample of independent observations $\{Y_1, Y_2, ..., Y_n\}$ from $f_\theta(y)$

Estimate θ based on sample.

The plan:

- find joint density of observations
- substitute in observations to form likelihood
- maximize likelihood
- curvature of likelihood surface $\equiv$ uncertainty of estimate

Example from exponential

$$f_\theta(y) = (1/\theta)e^{-y/\theta}$$

Independent observations,

g joint density function.

$$g_\theta(y_1, y_2, ..., y_n) = f_\theta(y_1)f_\theta(y_2)\dots f_\theta(y_n)$$

$$= (1/\theta)e^{-y_1/\theta}(1/\theta)e^{-y_2/\theta}(1/\theta)\dots e^{-y_n/\theta} = (1/\theta)^n e^{-\sum_i y_i/\theta}$$

Will abuse notation: $g_\theta(\boldsymbol{y})$ this is the probability of drawing $\boldsymbol{y}$ as a simple.

Likelihood: $L(\boldsymbol{Y}, \theta) = g_\theta(\boldsymbol{Y})$ Just substitute observations into joint density. The "likelihood" of observing this data given the parameter value θ.

Maximize Likelihood over θ

Maximze

$$L(\boldsymbol{Y}, \theta) = 1/(\theta^n) e^{-(\sum_i \boldsymbol{Y}_i)/\theta}$$

Maximizing likelihood is the same as maximizing log likelihood. In general the log likelihood is the natural scale to work on because it involves sums.

$$\log L(\boldsymbol{Y}, \theta) = -n \log(\theta) - (\sum_i \boldsymbol{Y}_i)/\theta$$

Maximum Likelihood Estimator (MLE) for exponential

$$\widehat{\theta} = (\sum_i Y_i)/n \equiv \overline{\boldsymbol{Y}}$$

How about normal?

$$f_\theta(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

log likelihood for a random sample

$$logL(\boldsymbol{Y}, \mu, \sigma) = -log(\sqrt{2\pi}\sigma) - \sum_i \frac{(\boldsymbol{Y}_i - \mu)^2}{2\sigma^2}$$

$$= -log(\sqrt{2\pi}) - log(\sigma) - \sum_i \frac{(\boldsymbol{Y}_i - \mu)^2}{2\sigma^2}$$

Take partials w/r to μ and σ set to set zero and solve

$$\frac{\partial log L(\boldsymbol{Y}, \mu, \sigma)}{\partial \mu} = \frac{\sum_i (\boldsymbol{Y}_i - \mu)}{\sigma^2} = 0$$

$$\frac{\partial log L(\boldsymbol{Y}, \mu, \sigma)}{\partial \mu} = -(1/\sigma) + \frac{\sum_i (\boldsymbol{Y}_i - \mu)^2}{\sigma^3} = 0$$

- From first equation $\widehat{\mu} = \overline{\boldsymbol{Y}}$ for all σ
- From second equation solution is simplified by solving for σ^2

$$\widehat{\sigma^2} = (1/n) \sum_i (\boldsymbol{Y}_i - \widehat{\mu})^2$$

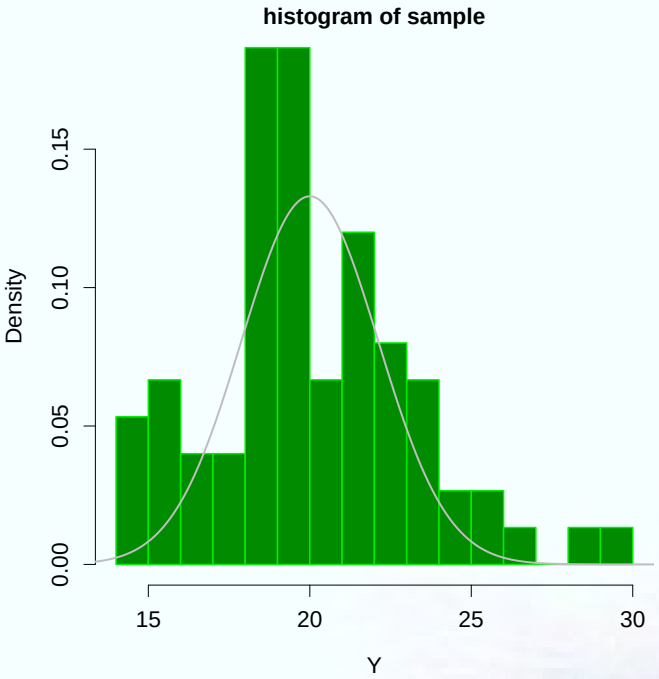
MLE for σ is $\sqrt{(1/n) \sum_i (\boldsymbol{Y}_i - \widehat{\mu})^2}$

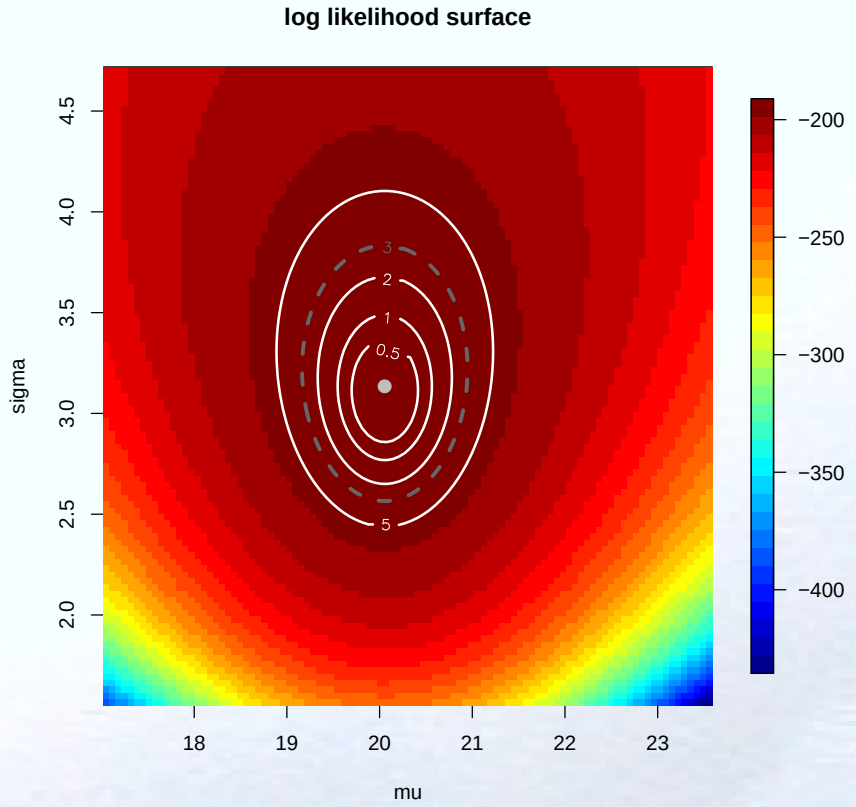
MLEs are transform invariant
– the maximizer is always the maximizer!
i.e. MLE for $\omega(\theta)$ is just $\omega(\widehat{\theta})$

Approximate inference

Key is to examine $2L(Y, \boldsymbol{\theta})$ as a function of parameter values and compare to maximum at MLE. Expect a quadratic surface about MLE.

Normal example: $(n = 75, \mu = 20,\ \sigma = 3)$, MLEs=(19.9, 3.05)





Contours are the difference between log likelihood and the maximum.

Approximate confidence sets are found by the contour that is offset by a chi square value from the maximum.

$(1-\alpha)\%$ *Confidence set:*

$$\left\{\boldsymbol{\theta} : 2logL(Y,\hat{\boldsymbol{\theta}}) - 2logL(Y,\boldsymbol{\theta}) < \chi^2_p(\alpha)\right\}$$

$p =$ number of parameters

i.e. All parameters with lnLike large enough.

For normal example $\chi^2_p(.05) = 5.99$ (In R: `qchisq( .95, 2)`).
Grey contour (5.99/2) in previous figure.

Regression models

This is mainly to review using vector/matrix notation in likelihoods.

A linear model:

$$\boldsymbol{y} = X\boldsymbol{d} + \boldsymbol{e}$$

$\{e_1, e_2, \ldots, e_n\}$ are independent, normal mean 0 and variance σ^2. X a known matrix of covariates and d are the parameters.

$\boldsymbol{y}_i$ are normal mean $[X\boldsymbol{d}]_i$ and variance σ^2

$$g(\boldsymbol{y}) = f_1(\boldsymbol{y}_1) f_2(\boldsymbol{y}_2) \ldots f_n(\boldsymbol{y}_n)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_1 - [Xd]_1)^2}{2\sigma^2}} \ldots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_n - [Xd]_n)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{(\boldsymbol{y} - \boldsymbol{Xd})^T(\boldsymbol{y} - \boldsymbol{Xd})}{2\sigma^2}}$$

log Likelihood (least squares!):

$$-(n/2)\log(2\pi) - n\log(\sigma) - \frac{(\boldsymbol{y} - \boldsymbol{Xd})^T(\boldsymbol{y} - \boldsymbol{Xd})}{2\sigma^2}$$

Partial derivatives set to zero:

$$X^T(\boldsymbol{y} - X\boldsymbol{d}) = 0$$

$$-n/\sigma - \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{d})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{d})}{2\sigma^2} = 0$$

MLEs

$$\hat{\boldsymbol{d}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

$$\hat{\sigma}^2 = (1/n)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{d})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{d})$$

Note: this will work for a nonlinear mean function depending on parameter vector $\boldsymbol{d}$ and will lead to minimizing the sums of squares but the MLEs may not have a closed form.

(This provides a framework for fitting variograms by nonlinear regression.)

Covariance estimation

Using a fields functions to fit a Matern covariance (smoothness=1.0) using maximum likelihood.

```
fit3<- MLE.Matern( xHW1,yHW1, Z=elevHW1, smoothness=1.0, m=2,
                   theta.grid=seq(.2,4,,150))
plot( fit3$REML.grid[,1], fit3$REML.grid[,2])
yline( max( fit3$REML.grid[,2]) - qchisq(.95,1)/2)
```

Profile likelihood: For each value of θ maximize over the sill and nugget (ρ and σ). This is a way to explore uncertainty in estimating a specific parameter. In this case the chi square degrees of freedom is just one (1), the total number of parameters (3) minus the number of parameters being maximized (2).