Least Squares Estimation

SIO 223A
Lecture 15
2/25/2020

- Simple examples
- Fitting a straight line
- Assessing fit
- Correlation and Regression
- Matrix approach to Least Squares
- Statistical properties
- Inferences about theta

Matrix Representation for Linear Least Squares

$$\vec{Y} = X\vec{\theta} + \vec{\epsilon} \tag{28}$$

$\vec{Y} \in \mathbb{R}^n$, $\vec{\theta} \in \mathbb{R}^p$. X is an $n \times p$ matrix (not a random variable) and is known as the **design matrix**; its rows are the $x_1, x_2, \ldots, x_p$'s for each measurement of y:

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \vec{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1,p} \\ x_{21} & x_{22} & \cdots & x_{2,p} \\ \vdots & \ddots & \vdots & \\ x_{n1} & x_{n2} & \cdots & x_{n,p} \end{pmatrix} \tag{29}$$

We can also define a **residual vector** $\vec{r} \in \mathcal{R}^n$

$$\vec{r} = \vec{Y} - X\vec{\theta} \tag{30}$$

We see that $S = \vec{r}.\vec{r}$ or the Euclidean length of $\vec{r}$. In other words the least-squares solution is the one with the smallest misfit to the measurements, as given by

$$\vec{r}^T\vec{r} = r_i r_i = \sum_{i=1}^{n} r_i^2 = \vec{r}.\vec{r} = \|\vec{r}\|_2^2 \tag{31}$$

We want $\vec{\theta} = \hat{\theta}$ such that

$$\nabla_{\vec{\theta}}[\vec{r}(\vec{\theta}).\vec{r}(\vec{\theta})] = \vec{0} \tag{32}$$

Equivalently,

$$\begin{aligned} \nabla_{\vec{\theta}}\left[(\vec{Y} - X\vec{\theta})^T(\vec{Y} - X\vec{\theta})\right] &= \vec{0} \\ \nabla_{\vec{\theta}}\left[\vec{Y}^T\vec{Y} - \vec{Y}^T X\vec{\theta} - \vec{\theta}^T X^T\vec{Y} + \vec{\theta}^T X^T X\vec{\theta}\right] &= \vec{0} \end{aligned} \tag{33}$$

Since $\vec{Y}^T X\vec{\theta} = \vec{\theta}^T X^T\vec{Y} = (X^T\vec{Y})^T\vec{\theta}$ this becomes

$$\nabla_{\vec{\theta}}\left[\vec{Y}^T\vec{Y} - 2(X^T\vec{Y})^T\vec{\theta} + \vec{\theta}^T X^T X\vec{\theta}\right] = 0 \tag{34}$$

whence

$$-2X^T\vec{Y} + 2X^T X\vec{\theta} = 0 \tag{35}$$

which is to say

$$\hat{\theta} = (X^T X)^{-1}X^T\vec{Y} \tag{36}$$

(36) are the **normal equations**.

Simple Example: y=bx

The matrix notation is readily understood if we use as an example the straight line fitting from an earlier section. In this context (29) produces

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \vec{\theta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{37}$$

We can form the normal equations as in (36) by

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{38}$$

yielding

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

Inverting this we find

$$(X^T X)^{-1} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}$$

along with

$$X^T Y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

Now we get $\hat{\beta}$ using (36)

$$\begin{aligned} \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} &= (X^T X)^{-1} X^T \vec{Y} \\ &= \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} \\ &= \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} (\sum_i y_i)(\sum_i x_i^2) & -(\sum_i x_i)(\sum_i x_i y_i) \\ n(\sum_i x_i y_i) & -(\sum_i x_i)(\sum_i y_i) \end{bmatrix} \end{aligned}$$

Statistical Properties of LS Estimators

If the errors in the original measurements are uncorrelated, *i.e.*, $Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall\, i \neq j$ and they all have the same variance, σ^2, then we write the data covariance matrix as $C_{\epsilon\epsilon} = \sigma^2 I$. I is an n by n identity matrix. When this property holds for the data errors, each $\hat{\theta}_k$ is an unbiased estimate of θ_k

$$\mathcal{E}(\hat{\theta}_k) = \theta_k \quad \text{for all} \;\; k = 1, \ldots, p \tag{43}$$

Also the variance-covariance matrix for $\hat{\theta}$ is $C_{\hat{\theta}\hat{\theta}} = \sigma^2 (X^T X)^{-1}$, so that

$$\begin{aligned} Cov(\hat{\theta}_k, \hat{\theta}_l) &= k, \;\; l\text{th element of } \sigma^2 (X^T X)^{-1} \\ Var(\hat{\theta}_k) &= k, \; k\text{th diagonal element of } \sigma^2 (X^T X)^{-1} \end{aligned} \tag{44}$$

$C_{\hat{\theta}\hat{\theta}}$ is a p by p matrix, and (44) is just a generalization of (13). Observe that even though the uncertainties in the original measurements are uncorrelated the parameter estimates derived from them are in general correlated. Under this model an unbiased estimate for σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{||\vec{Y} - \hat{Y}||^2}{n - p} = \frac{RSS}{n - p} \tag{45}$$

LS Estimators are BLUE, best, linear, unbiased estimates

- BLUE versus MMSE
- Inferences about the estimates
- Squared multiple correlation coefficient and variance reduction

SIOG 223A: Lecture 16
2/27/2020

- Equivalence of ML and LS Estimates for Multivariate Normal

- Weighted LS

- Numerical Stability

- NLLS

- A little bit about optimization

The Arrhenius relationship for thermally activated semiconduction in minerals is $\sigma(t) = \sigma_0 e^{-A/kt}$ *where* $\sigma(t)$ *is the electrical conductivity at temperature* t*,* k *is Boltzmann's constant and* A *is the activation energy. This has been used to model the electrical conductivity data for the mineral olivine as a function of temperature. Olivine is a major constituent of Earth's mantle, and it is of some interest to understand the relationship between conductivity and the temperature and other physical properties of Earth's interior. For the conductivity example we can solve for the activation energy* A *and* σ_0 *simply by working in the log domain, and the transformed model is shown in Figure 7-1 for conductivity data derived from a sample of Jackson County dunite.*

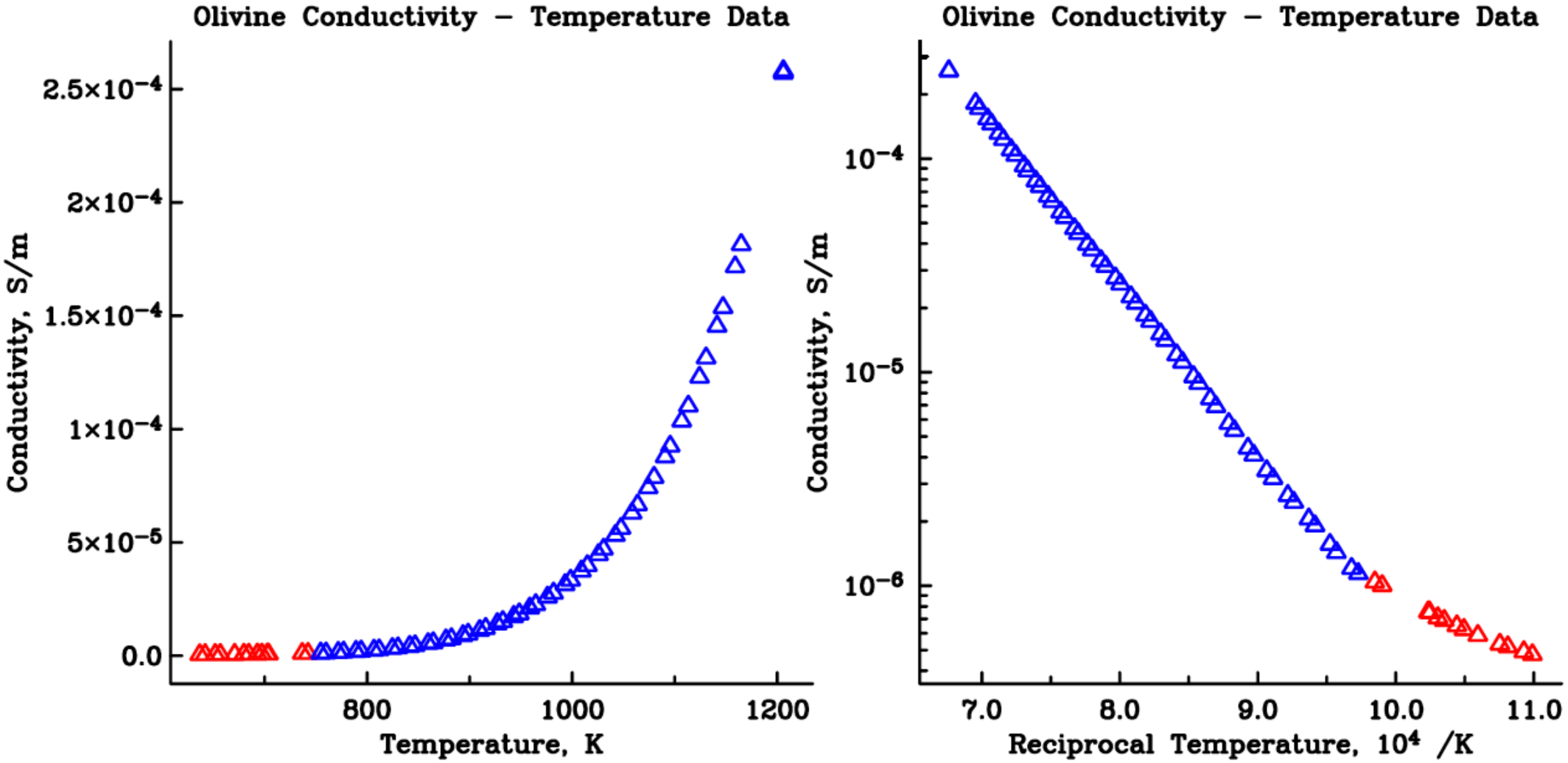


Figure 7 -1: Temperature - conductivity data for Jackson County dunite. The blue symbols appear to follow the Arrhenius relationship reasonably well, but the red parts will require additional non-linear terms.

Total Least Squares and Robust Methods

SIO 223A
Lecture 17

- What if the design matrix in least squares is random, as well as the observations?

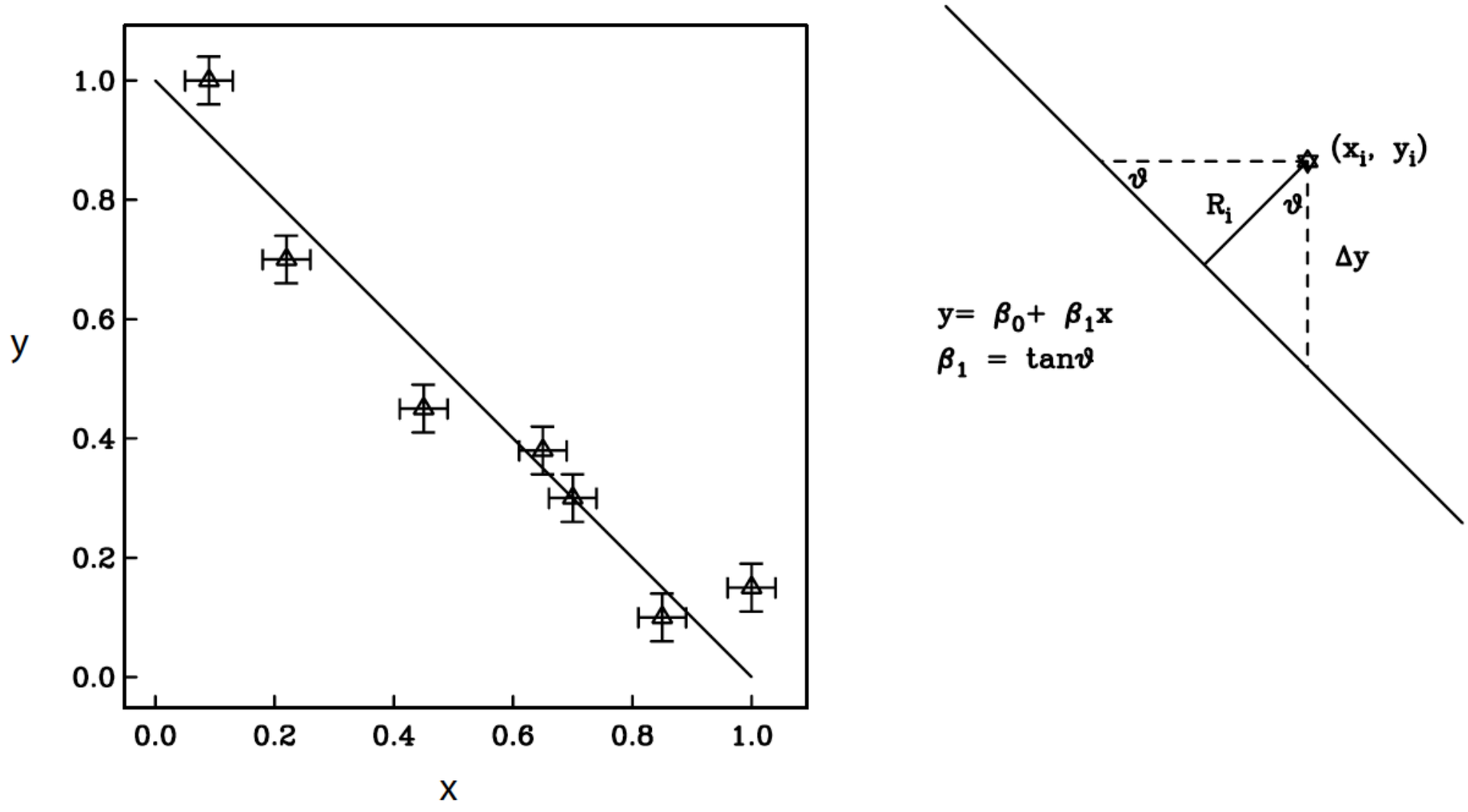


Figure 8.1: The total least squares problem for a straight line. Note that in the illustration the uncertainties in x and y are equal.

- Robust methods and iteratively reweighted least squares

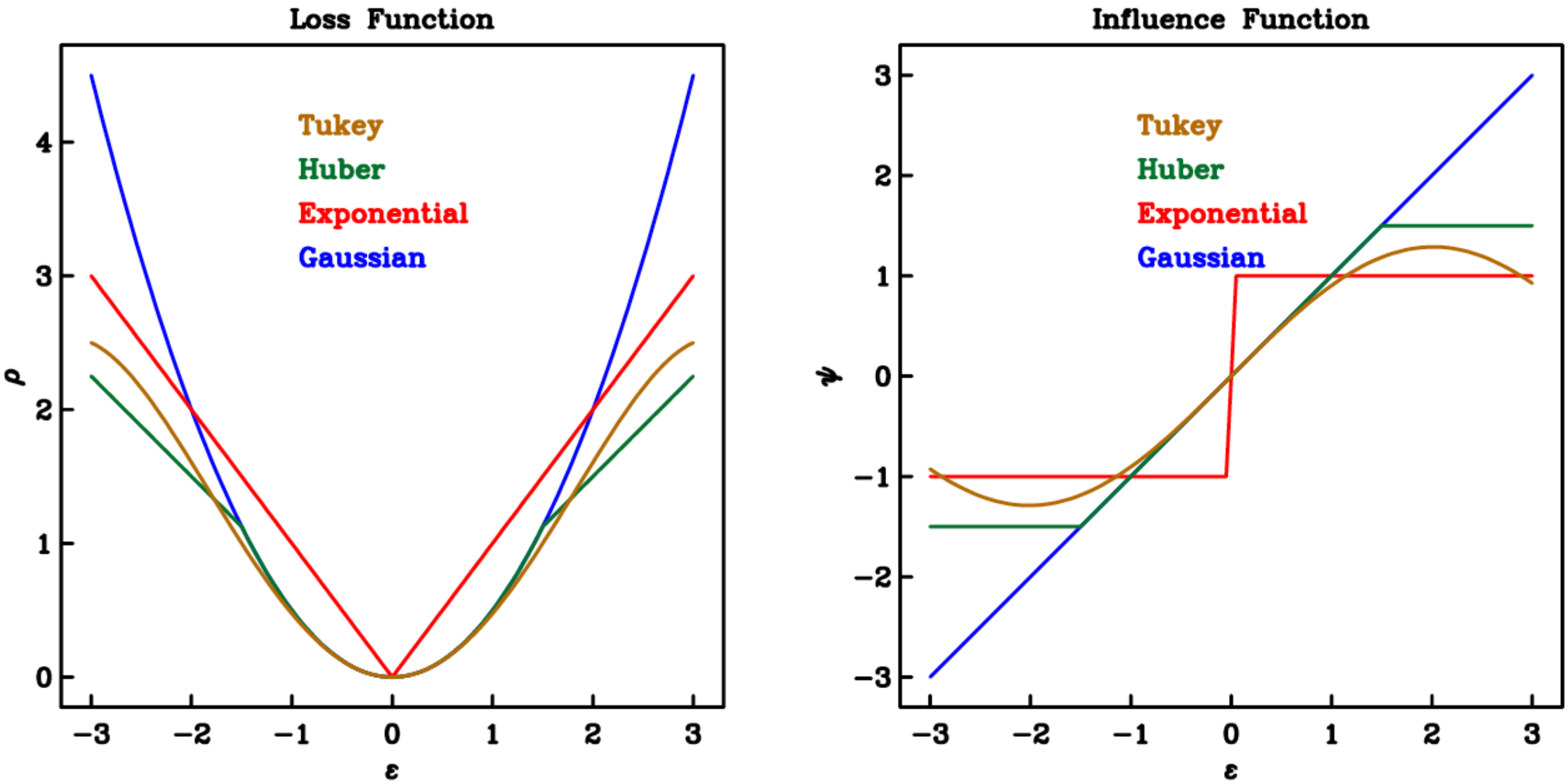


Figure 8.2: Loss and influence function for ML estimation with Gaussian and exponentially distributed noise. Robust M-type loss and influence functions for Huber's t-function with $t = 1.5$, and Tukey's biweight with $t = 4.5$.

Non-Parametric Density Function Estimation

SIO 223A
Lecture 18

Background Reading: Chapter 9 of notes,
Dekking et al, Chapter 15
Constable (2000), doi: 10.1016s0031-9201(99)00139-9

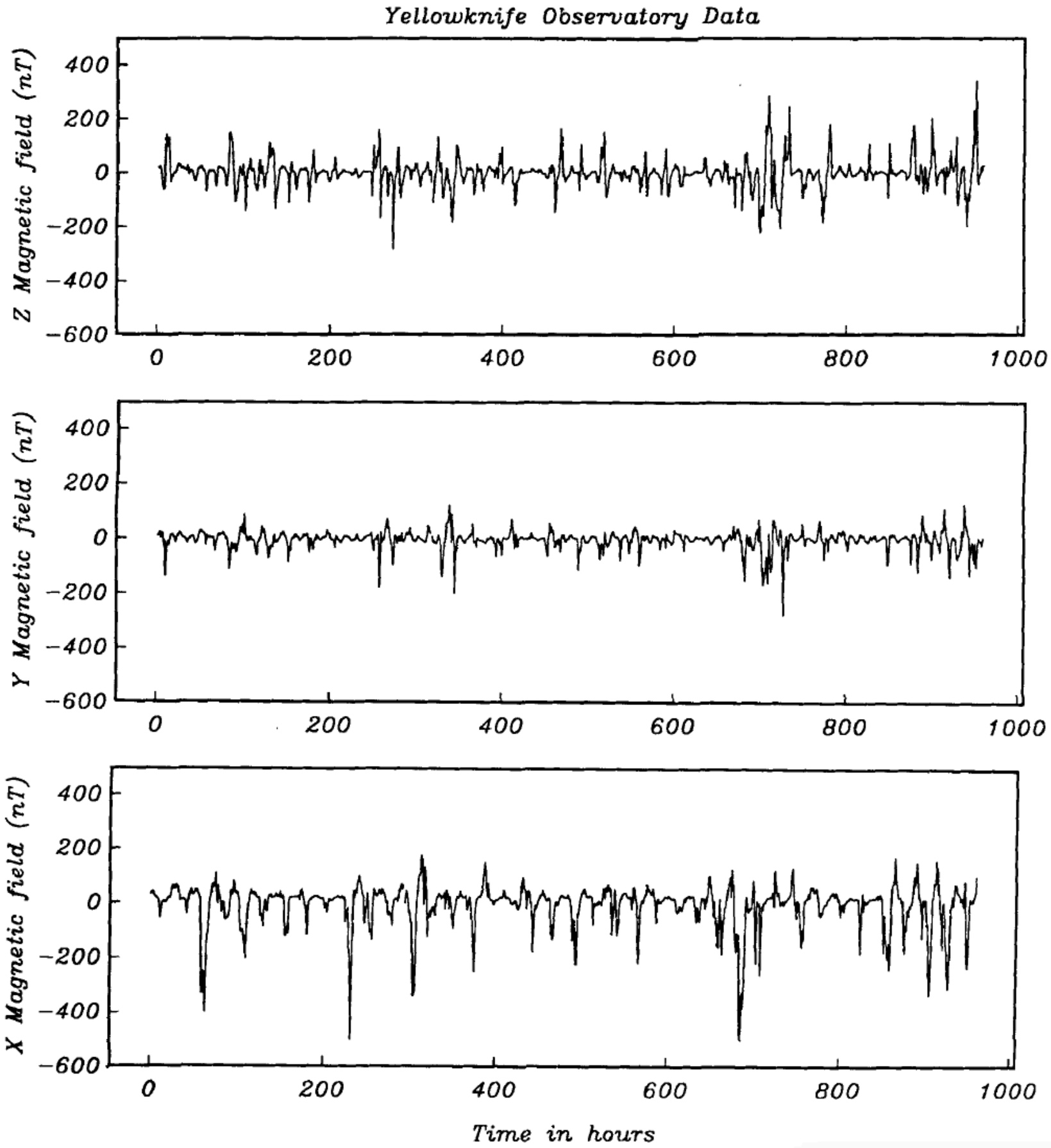


Figure 2. Sample of hourly average values of geomagnetic observatory data from Yellowknife, Canada.

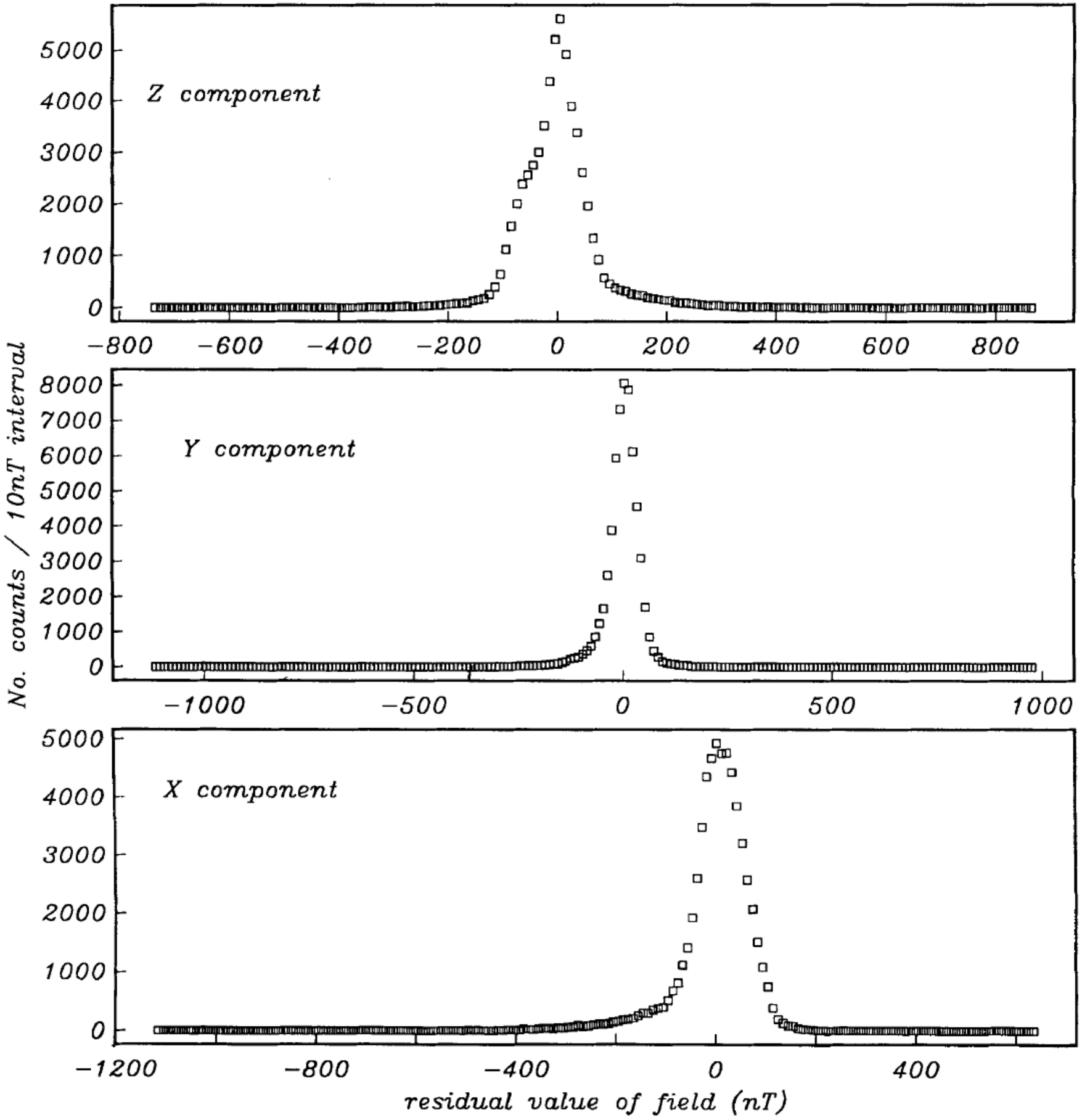


Figure 3. Histograms of residuals from LS fit of lines to Yellowknife data set.

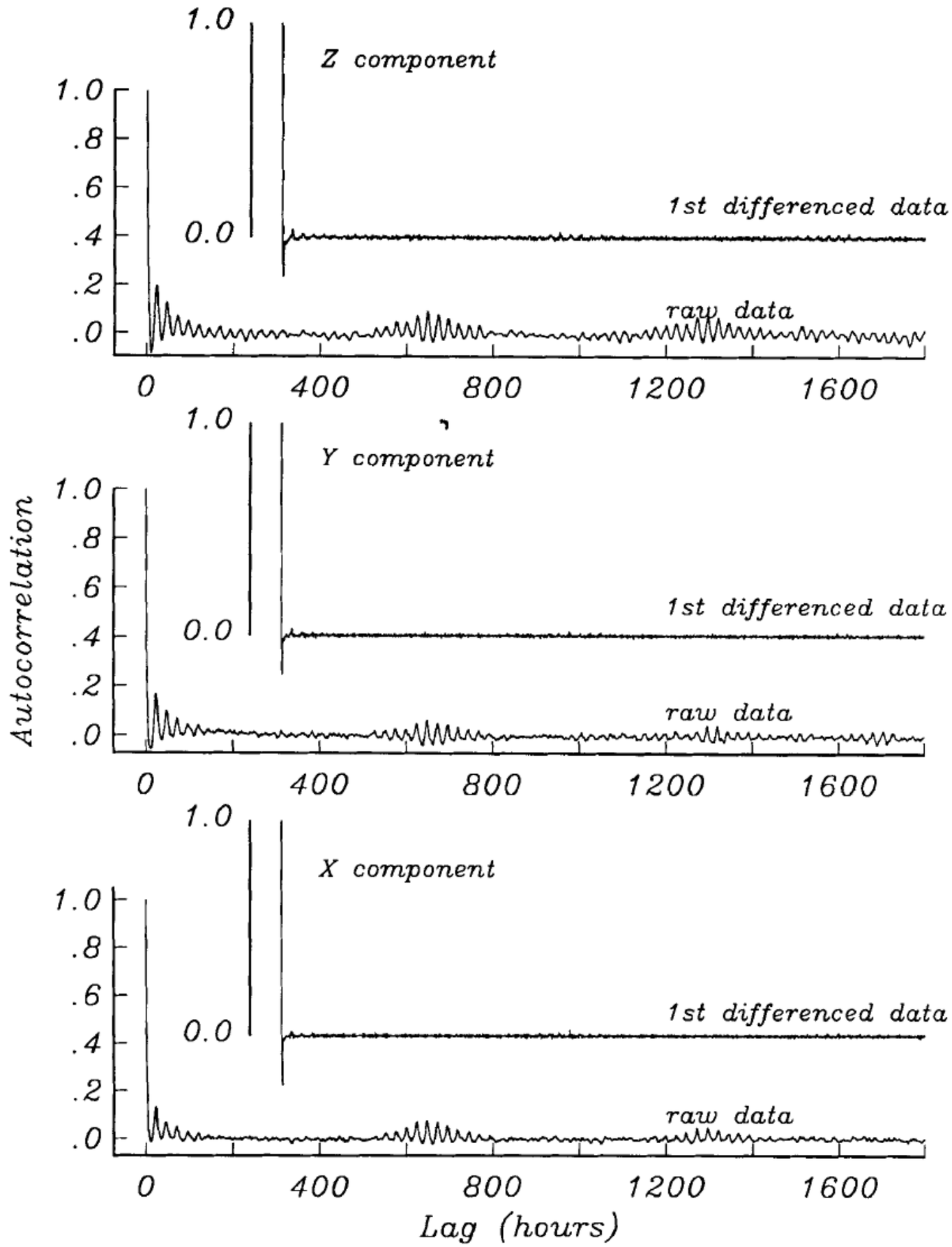


Figure 4. Autocorrelation functions for the residuals from a LS fit to the Yellowknife data. The lower curve in each part of the figure is for the unfiltered data, the upper part the autocorrelation for the residuals from the first differenced data.

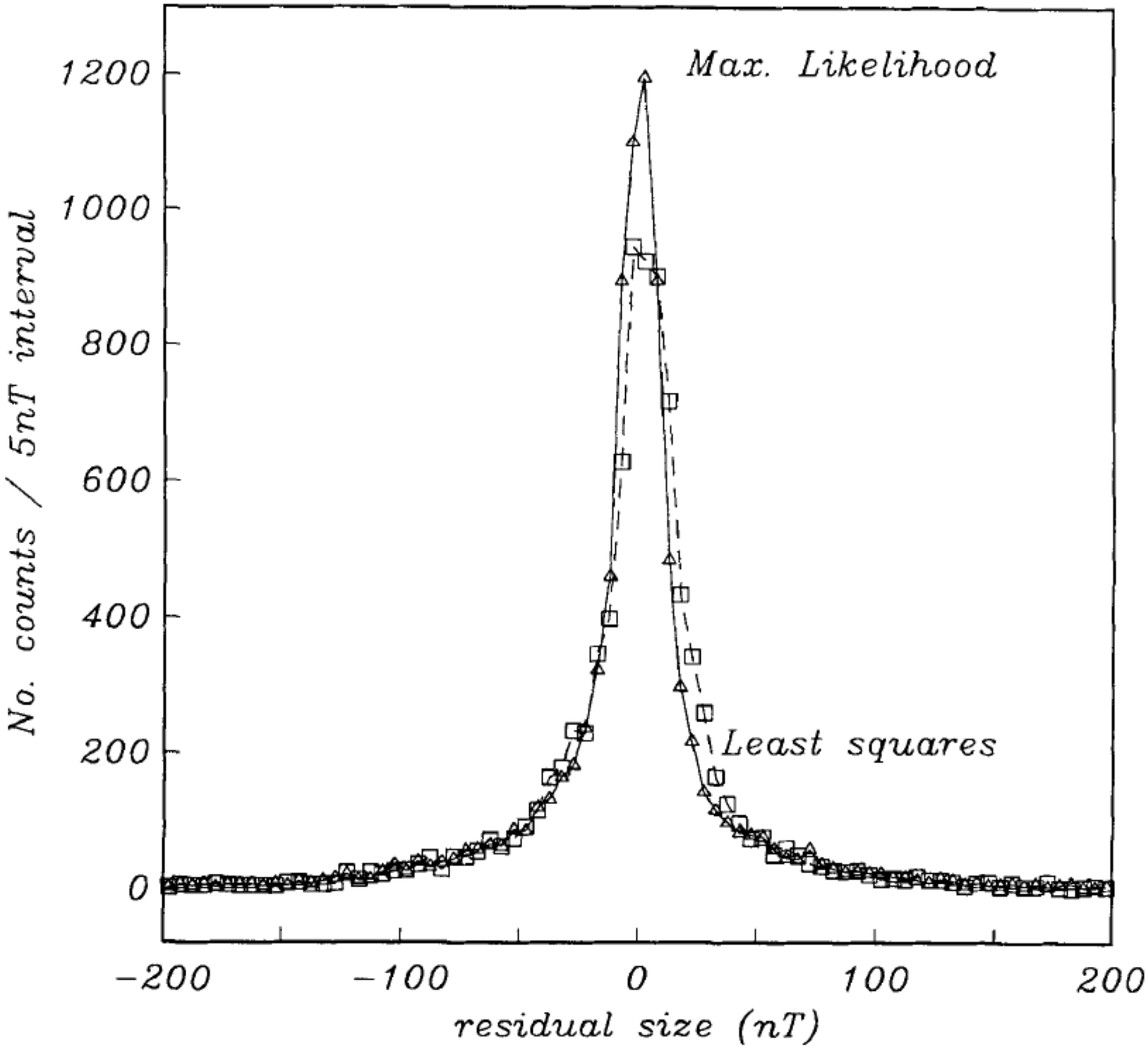


Figure 6. Comparison of the histograms of residuals for least-squares (dashed line and squares) and maximum likelihood (solid line and triangles) estimation.

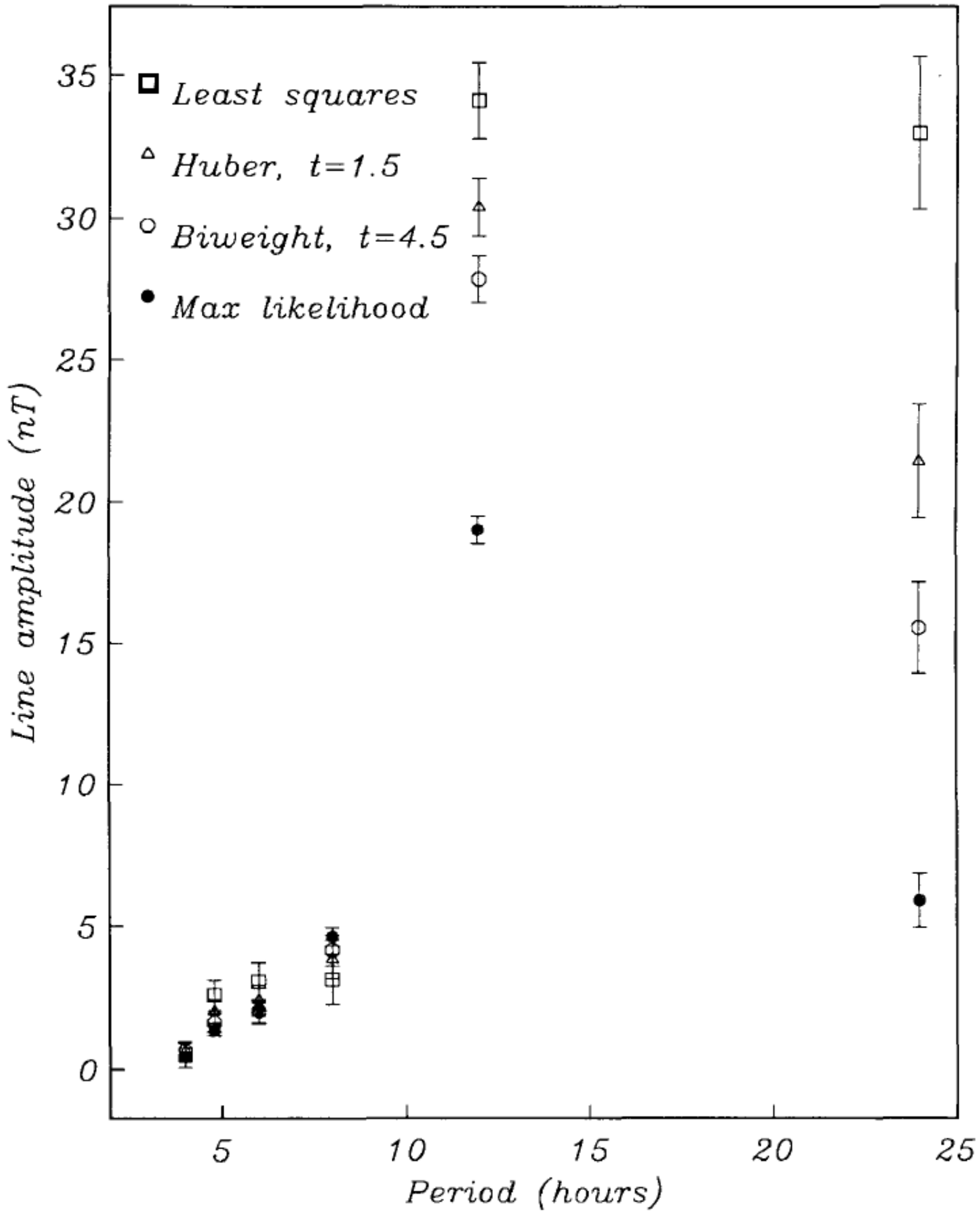
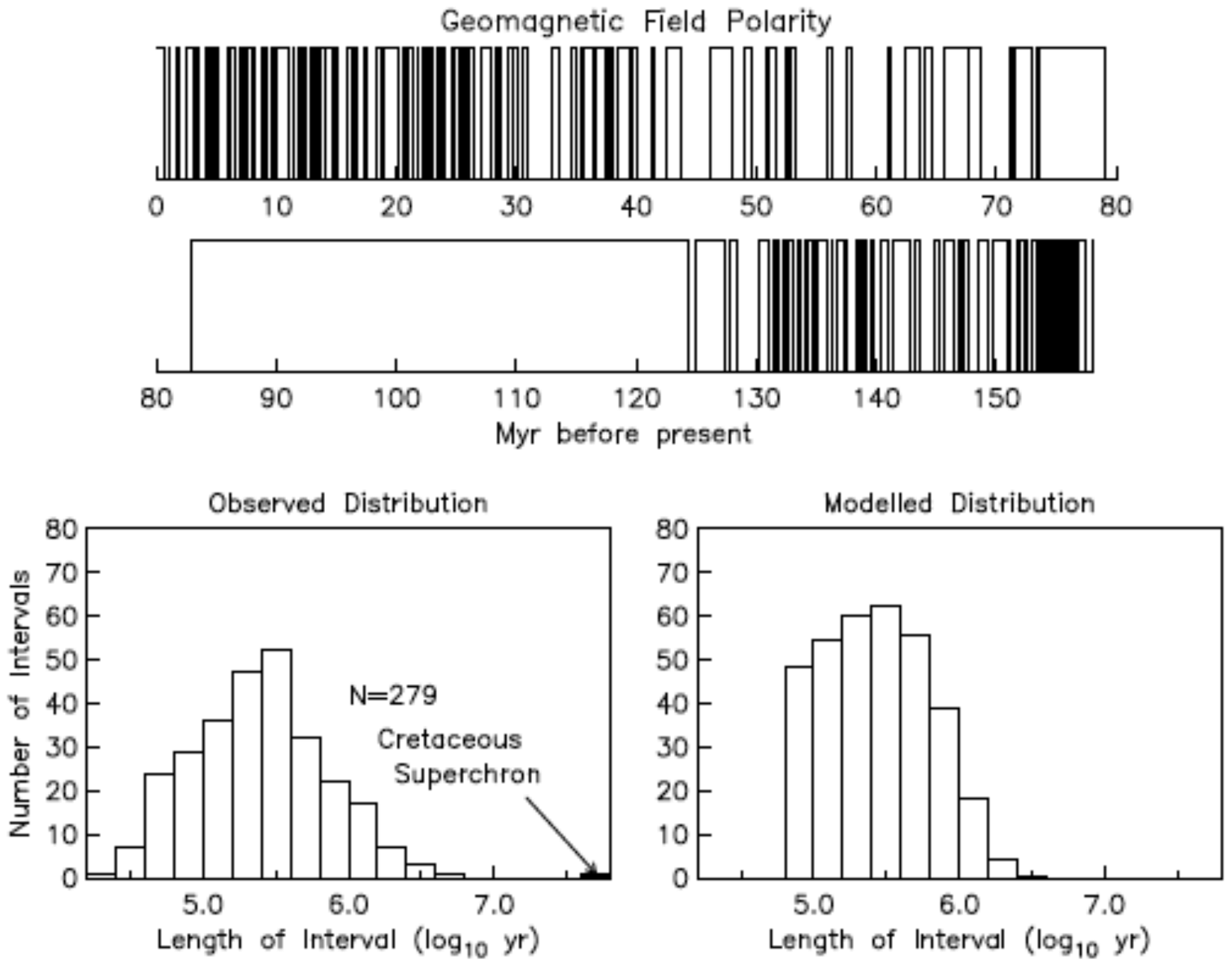


Figure 7. Line amplitudes obtained by the various methods discussed in the text. Error bars are one standard deviation as computed by (11).

Today's topics

- Non-parametric density estimates
- Histograms, Naive estimators, Kernel density estimates, pros and cons
- Choosing bandwidth for an estimate
- Comparing with the sample distribution function
- Adaptive estimation: nearest neighbors and variable kernels
- Maximum penalized likelihood estimation

A pdf for polarity interval length



Histograms

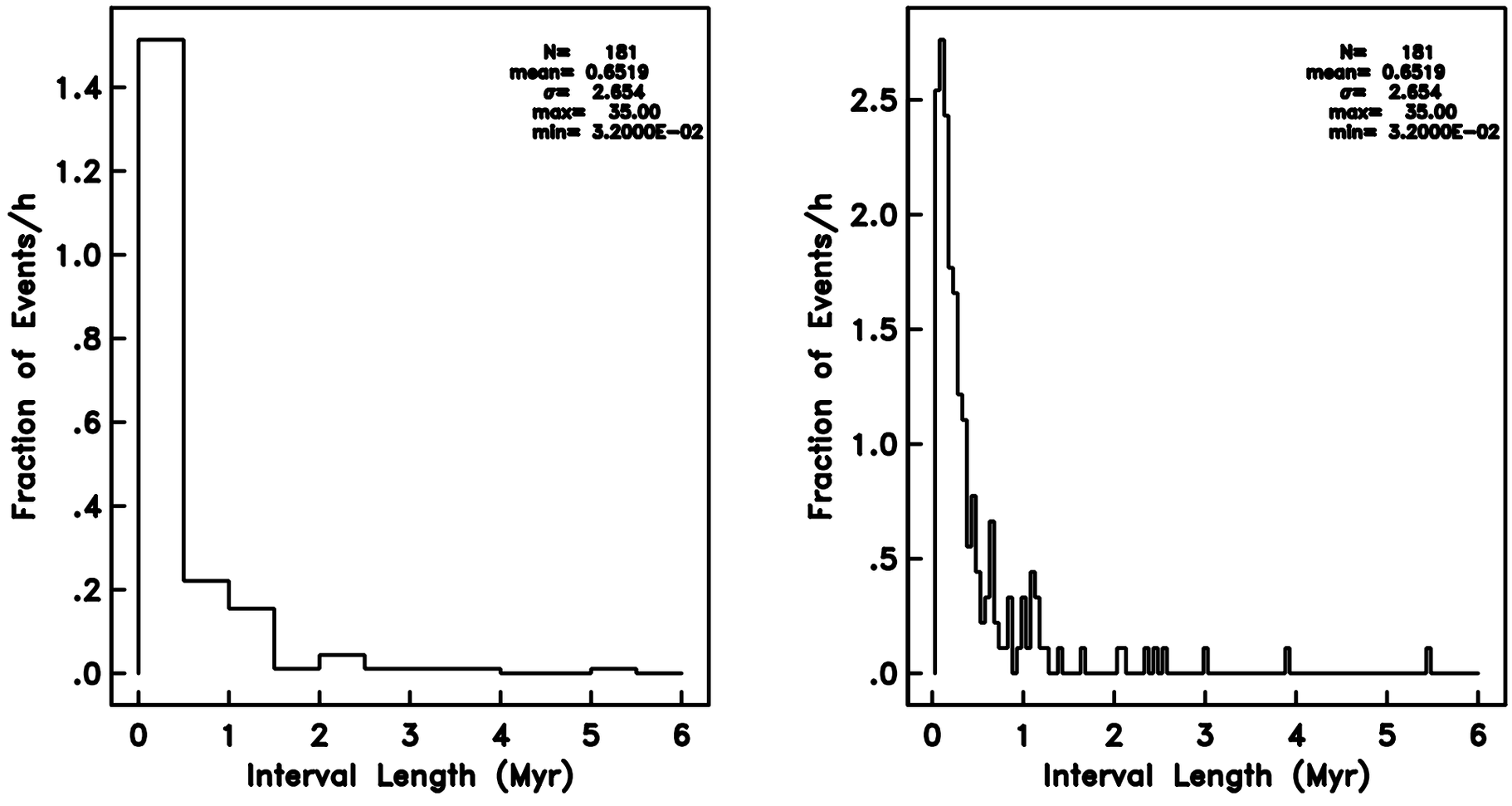


Figure 9-1

Sample Distribution functions and various estimates of pdf

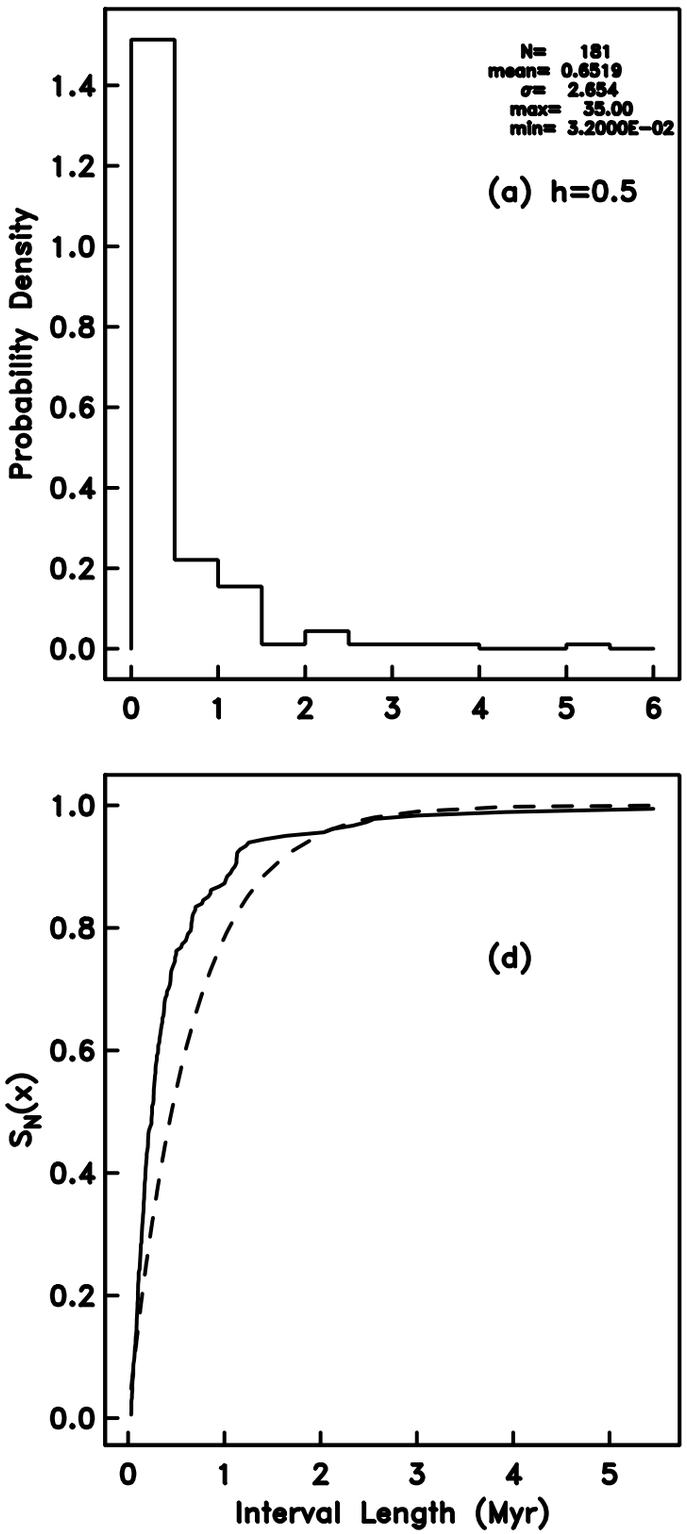
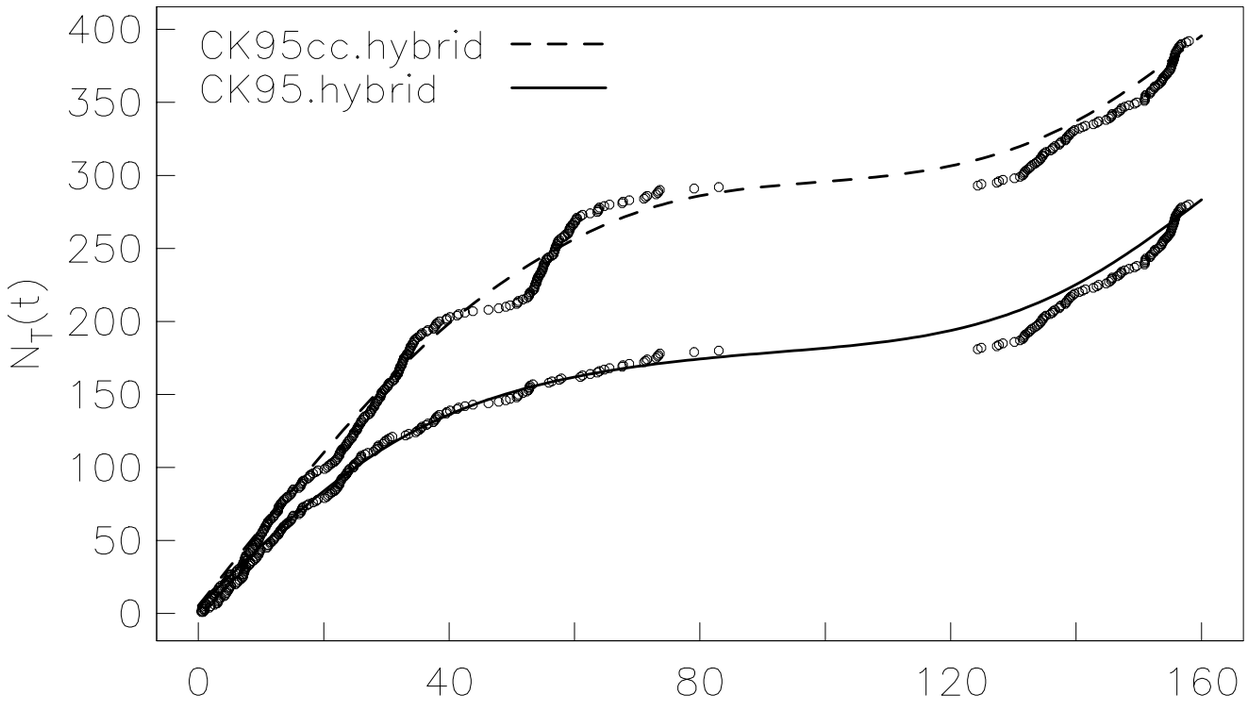


Figure 9-2

Estimating Reversal Rate as a function of time
Constable, 2000, PEPI

Finding the time-varying rate for a Poisson process - adaptive density estimate



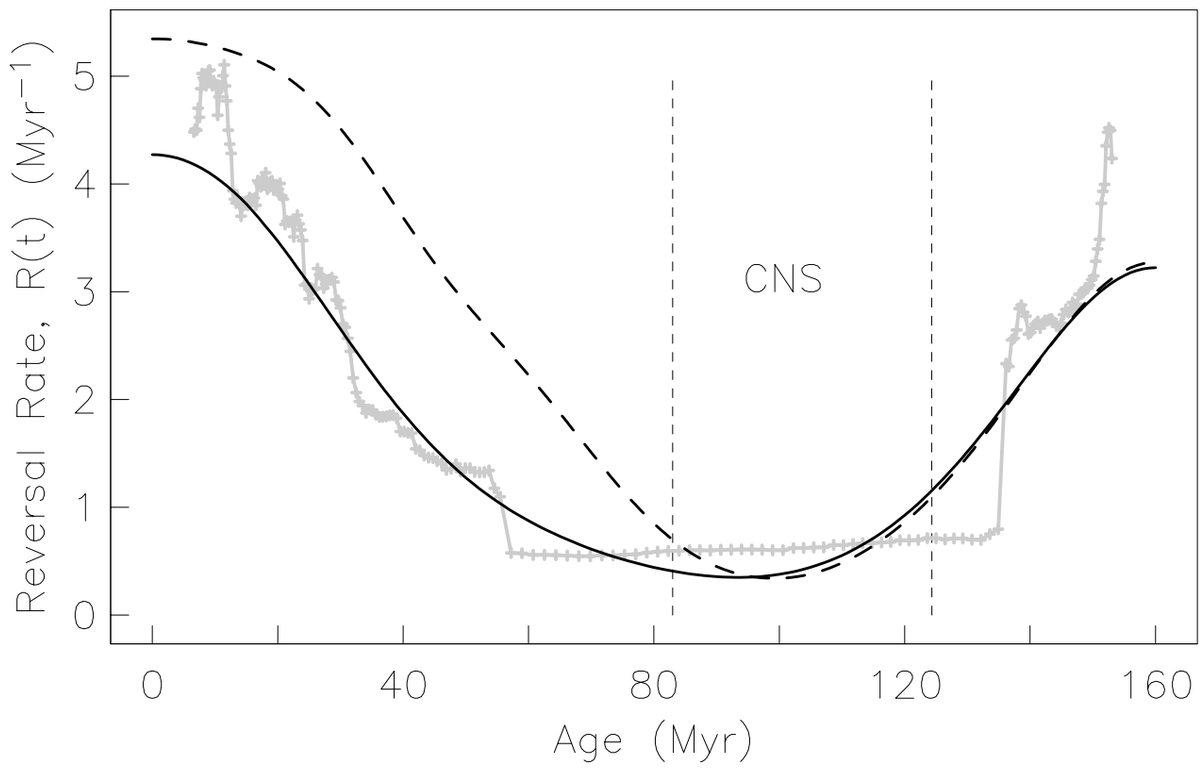


Figure 1

Finding the time-varying rate for a Poisson process - separate adaptive density estimate before and after the CNS

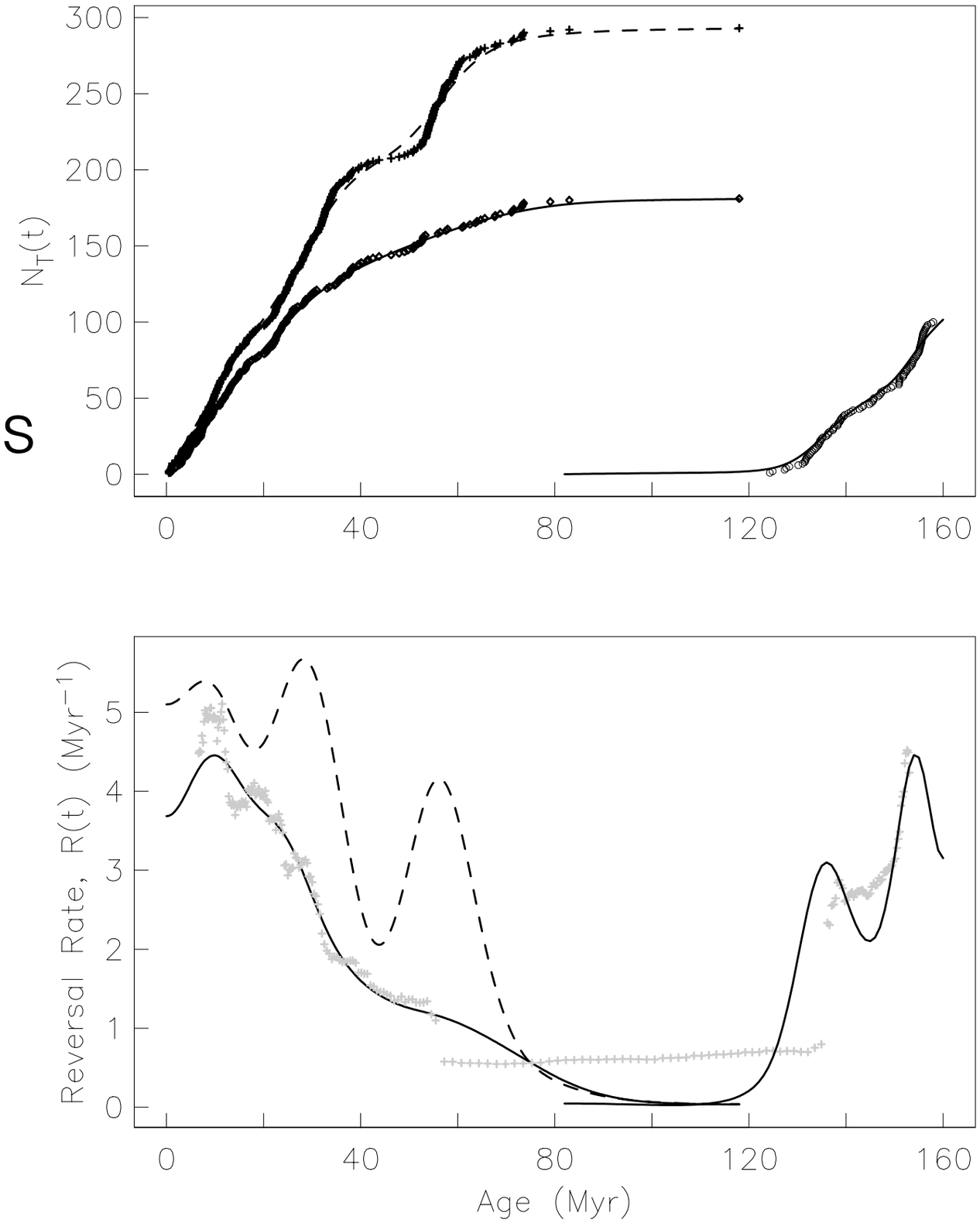
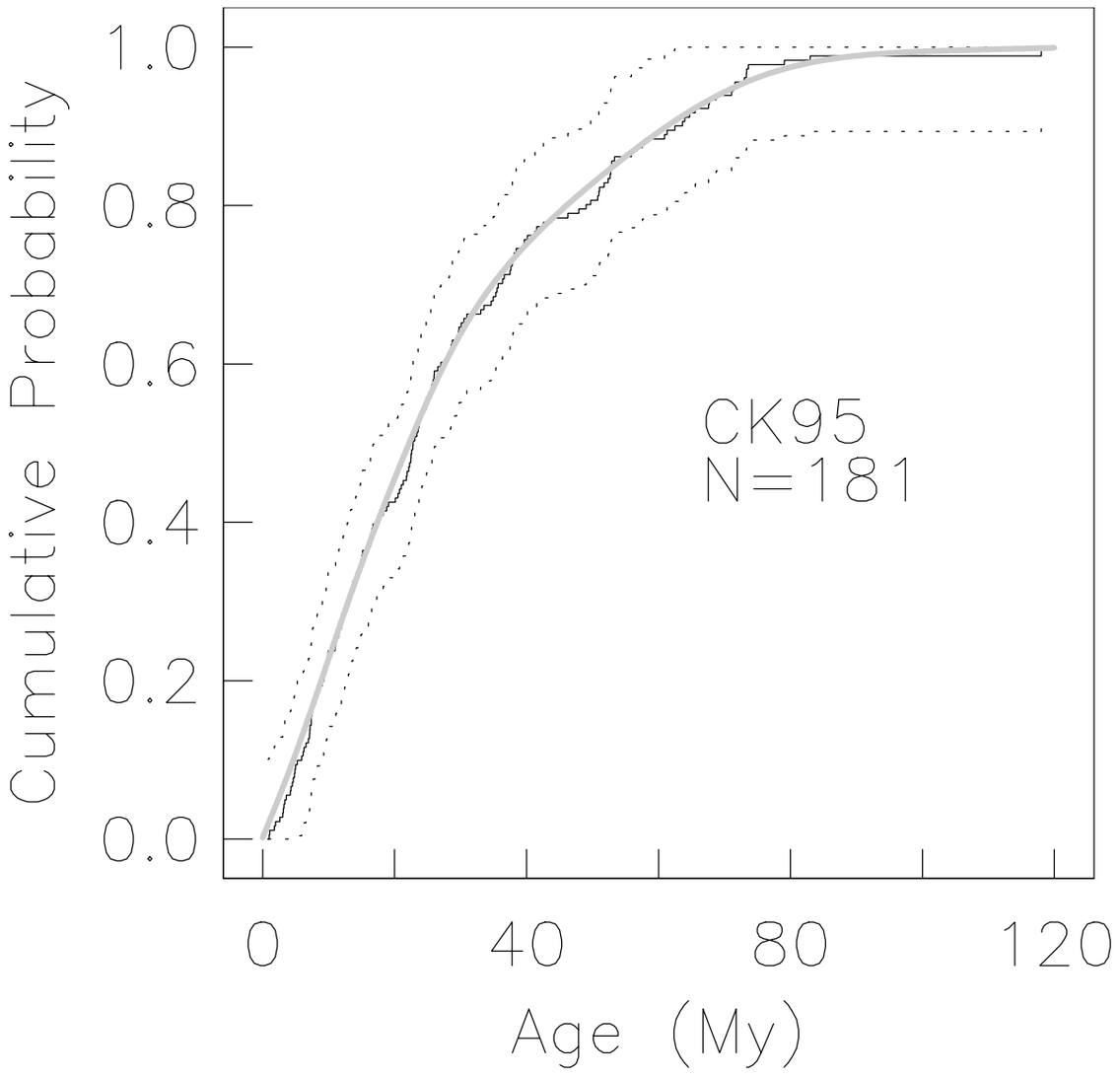


Figure 2

Putting bounds on the rate estimate using the Kolmogorov Smirnov Statistic



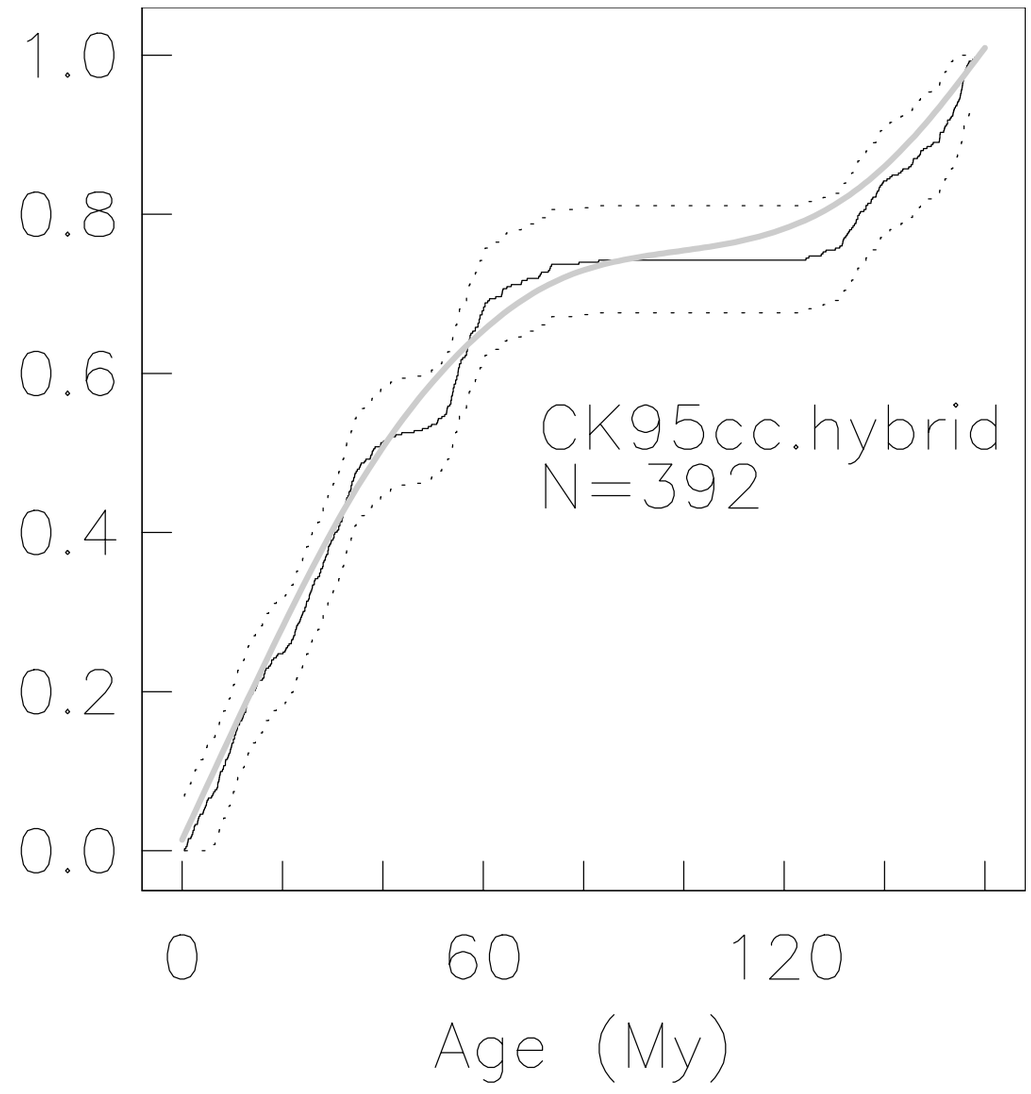


Figure 3

Putting Bounds on Reversal Rates, assuming monotonic variation of the pdf over 2 intervals

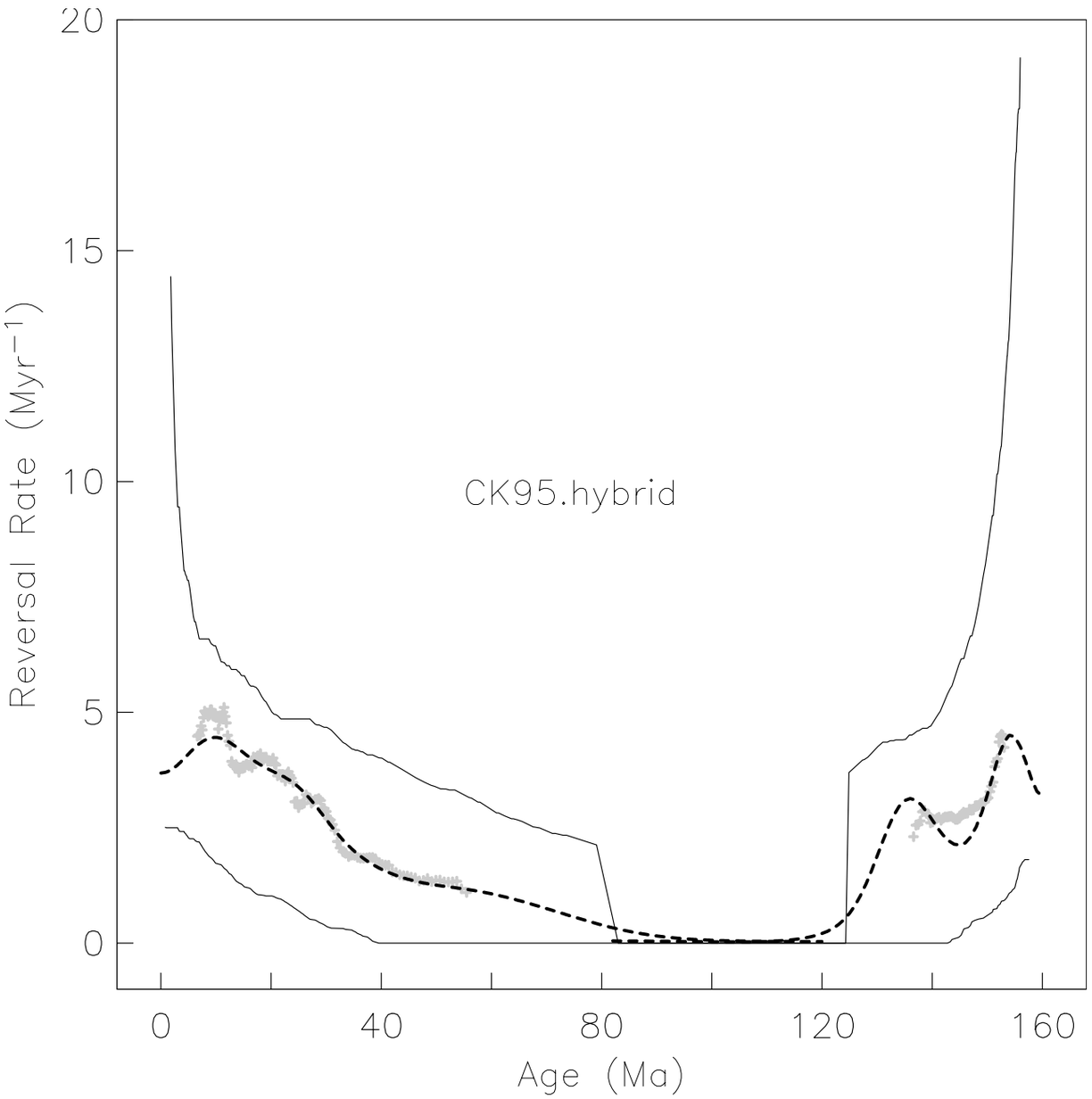


Figure 4

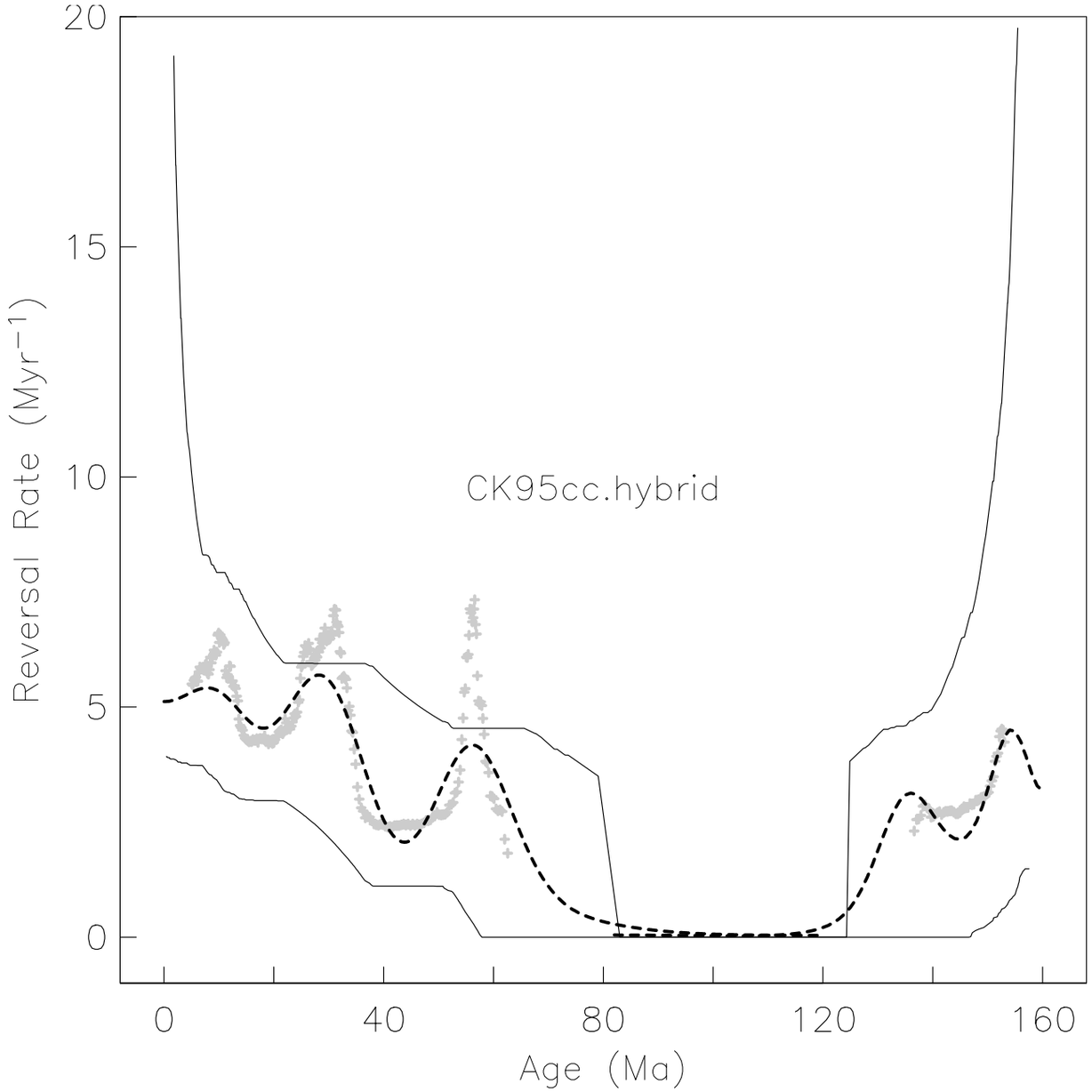


Figure 5