

Statistical Data Analysis for Physicists

Homework 4

Artur Palha n°46724
Luís Filipe Figueira n°46740

4th November 2002

Exercise 1. Asymptotic properties of the Maximum Likelihood Estimator.

- (a) Consistency. Show that the MLE $\hat{\theta}_n$ is a consistent estimator of θ , given some conditions you will explicitly write down. You can use the normalization of the p.d.f., the Central Limit Theorem. Prove it only for one parameter.

Proof. What we wish to prove is that $\hat{\theta}_n \xrightarrow{P} \theta_0$ where $\hat{\theta}_n$ is the Maximum Likelihood Estimator (MLE), θ_0 is the actual value of the parameter and

$$\hat{\theta}_n \xrightarrow{P} \theta_0 \Leftrightarrow \forall \epsilon > 0: \lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta_0| < \epsilon) = 1$$

The estimator $\hat{\theta}_n$ is defined as

$$\forall \theta: \mathcal{L}(\theta|x) \leq \mathcal{L}(\hat{\theta}_n|x)$$

where $\mathcal{L}(\theta|x)$ is the likelihood function, i.e., the probability of having the parameter θ given the observed data x . Instead of the likelihood function, it is used the log-likelihood $\ln \mathcal{L}(\theta|x)$ instead.

Before, we shall need a few concepts and results needed for the complete proof. A function f is said to be *convex* if, given any two points a and b , the graph of f evaluated on a point $x \in [a, b]$ passes under the line joining a and b . If that happens, then f has no maximum and $f''(x) \geq 0$ for all x in its domain. It is easily seen that, if f is smaller than any point in the line that joins $f(a)$ and $f(b)$, then the area under the graph in the region $[a, b]$ shall necessarily be smaller than the trapezoid formed by joining a , $f(a)$, $f(b)$ and b . Using infinitesimal intervals, we have

$$\frac{f(x) + f(x + dx)}{2} \geq f\left(\frac{x + x + dx}{2}\right) \Leftrightarrow \frac{f(x) + df + f(x)}{2} \geq f\left(x + \frac{1}{2}dx\right)$$

where we've used the formula from non-standard analysis $df = f(x + dx) - f(x)$ (or, equivalently, a first-order Taylor expansion on dx). We can now do the same for a very large number of intervals and trapezoids and, summing over the dx (or, integrating), we have

$$\int dx \left\{ \frac{f(x) + df + f(x)}{2} \right\} \geq f\left(\int dx \left\{ x + \frac{1}{2}dx \right\}\right)$$

Considering only the first order differentials, we get Jensen's inequality¹ (here without a more formal proof), which states that, for a smooth continuous convex function f ,

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

Since we're interested in studying the log-likelihood, we note that the function $f(x) = -\ln x$ is convex, for $f''(x) = 1/x^2 \geq 0 \forall x \in]0, \infty[$.

Now, let us consider the function $f(x) = -\ln \frac{\mathcal{L}(\theta^*|x)}{\mathcal{L}(\theta_0|x)}$. As we've seen, the logarithm is convex, and we can thus use Jensen's inequality to state

$$\mathbb{E}_0 \left[\ln \frac{\mathcal{L}(\theta^*|x)}{\mathcal{L}(\theta_0|x)} \right] < \ln \mathbb{E}_0 \left[\frac{\mathcal{L}(\theta^*|x)}{\mathcal{L}(\theta_0|x)} \right]$$

for all $\theta^* \neq \theta_0$ and where \mathbb{E}_0 is the average, taken with $\theta = \theta_0$. The equality is verified if $\theta^* = \theta_0$. The mean value on the right side is just

$$\mathbb{E}_0 \left[\frac{\mathcal{L}(\theta^*|x)}{\mathcal{L}(\theta_0|x)} \right] = \int \frac{\mathcal{L}(\theta^*|x)}{\mathcal{L}(\theta_0|x)} \mathcal{L}(\theta_0|x) dx = \int \mathcal{L}(\theta^*|x) dx = 1$$

and, obviously, $\ln 1 = 0$, so

$$\mathbb{E}_0 \left[\ln \frac{\mathcal{L}(\theta^*|x)}{\mathcal{L}(\theta_0|x)} \right] < 0 \quad \Leftrightarrow \quad \mathbb{E}_0 \left[\frac{1}{n} \ln \frac{\mathcal{L}(\theta^*|x)}{\mathcal{L}(\theta_0|x)} \right] < 0$$

This can be written as

$$\mathbb{E}_0 \left[\frac{1}{n} \sum_i \ln f(x_i|\theta^*) \right] < \mathbb{E}_0 \left[\frac{1}{n} \sum_i \ln f(x_i|\theta_0) \right]$$

This last expression can be identified with the sample mean of a set of identical random variables $\ln f(\theta|x)$ where f is the probability density for the observable x . According to the Strong Law of Large Numbers, the sample mean will converge with almost certain probability to the expected value,

$$\forall \epsilon > 0: \lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_i \ln f(x_i|\theta) - \mathbb{E} \left[\frac{1}{n} \ln [L](x_i|\theta) \right] \right| < \epsilon \right) = 1$$

which implies that, when $n \rightarrow \infty$, we have the following inequality with almost certain probability:

$$\forall \theta^* \neq \theta_0: \ln \mathcal{L}(\theta^*|x) < \ln \mathcal{L}(\theta_0|x)$$

In particular, for $\theta^* = \hat{\theta}_n$. However, taking $\theta = \theta_0$ in the definition of the ML estimator, we have

$$\mathcal{L}(\theta_0|x) \leq \mathcal{L}(\hat{\theta}_n|x)$$

and so, combining both inequalities,

$$\ln \mathcal{L}(\hat{\theta}_n|x) < \ln \mathcal{L}(\theta_0|x) \leq \mathcal{L}(\hat{\theta}_n|x)$$

the only overall valid condition is therefore

$$\mathcal{L}(\hat{\theta}_n|x) = \mathcal{L}(\theta_0|x)$$

If, furthermore, the likelihood function is a one-to-one function of the parameter, we then have, asymptotically with almost certain probability, the desired result: $\hat{\theta}_n = \theta_0$. \square

¹To get Jensen's inequality, we should integrate using a measure function $P(x)dx$, where $P(x)$ could be identified with a probability density.

(b) *Asymptotic normality.*

Show that the MLE asymptotically follows a normal distribution of the estimates, under conditions you will explicitly formulate. You can use a Taylor series expansion of the estimates as well as the Central Limit Theorem. In other words we have

$$\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, 1)$$

Taylor's theorem tells us that:

$$\left(\frac{\partial \mathcal{L}}{\partial \theta_i} \right) \Big|_{\theta_0} = \sum_i^N (\hat{\theta}_j - \theta_{j,0}) \left(\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} \right) \Big|_{\theta^*} \quad (1)$$

with θ^* between $\hat{\theta}$ and θ_0 .

Since $\hat{\theta}$, as it was seen in the previous question, is consistent we have:

$$\hat{\theta} \xrightarrow{P} \theta_0$$

Which means that:

$$\theta^* \xrightarrow{P} \theta_0$$

Therefore the derivatives on the right side converge, in probability, to their expected values. We can rewrite Equation (1) as:

$$\mathbf{y} = V^{-1} \mathbf{z} \quad (2)$$

where:

$$\mathbf{y} = \left[\left(\frac{\partial \mathcal{L}}{\partial \theta_i} \right) \Big|_{\theta_0} \right]; \quad \mathbf{z} = [\hat{\theta}_j - \theta_{j,0}]; \quad V^{-1} = \left[\left(\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} \right) \Big|_{\theta^*} \right]$$

By the Central Limit Theorem, in the multivariate case:

$$\mathbf{y} \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{\mu}, M)$$

where $\boldsymbol{\mu}$ is the expected value of \mathbf{y} and $\{M\}$ is the covariance matrix of \mathbf{y} . If, for each i , verifies:

$$\mathbb{E} \left[\frac{\partial \mathcal{L}}{\partial \theta_i} \right] = 0$$

Implies that:

$$\mathbf{y} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, M)$$

Now, to compute M . We just have to remember that $M = \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^t]$, which means:

$$\begin{aligned} M &= \mathbb{E}[\mathbf{y}\mathbf{y}^t] \\ &= \mathbb{E} \left[\frac{\partial \mathcal{L}}{\partial \theta_j} \right] \\ &= -\mathbb{E} \left[\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} \right] \\ &= V^{-1} \end{aligned}$$

Hence:

$$\mathbf{y} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, V^{-1})$$

The exponent of this multinormal distribution is given by the quadratic form:

$$-\frac{1}{2} \mathbf{y}^t V \mathbf{y}$$

Using the transformation of coordinates, Equation (2) the quadratic form turns to:

$$-\frac{1}{2}\mathbf{z}^t V^{-1} \mathbf{z} \quad (3)$$

Because we will have:

$$\begin{aligned} -\frac{1}{2}\mathbf{y}^t V \mathbf{y} &= -\frac{1}{2}\mathbf{z}^t (V^{-1})^t V V^{-1} \mathbf{z} \\ &= -\frac{1}{2}\mathbf{z}^t (V^{-1})^t \mathbf{z} \end{aligned}$$

But V^{-1} is symmetric since it's constituted of second order partial derivatives of the type $\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j}\right)$. Therefore:

$$-\frac{1}{2}\mathbf{y}^t V \mathbf{y} = -\frac{1}{2}\mathbf{z}^t V^{-1} \mathbf{z}$$

Equation (3) means that the dispersion matrix of \mathbf{z} is $(V^{-1})^{-1} = V$. Hence:

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, V)$$

Which is just:

$$\sqrt{V}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}_p(0, 1)$$

But $V = I_n$. Which finishes our demonstration. ■

(c) *Bias and variance.*

From the Taylor's expansion we deal in the previous question, we introduce a bias. Compute $b_n(\hat{\boldsymbol{\theta}})$ in the one parameter case, and the variance (much quicker) to order $\frac{1}{n}$ using a Taylor series approximation of the estimates.

First, we think it's necessary to define (introduce) some concepts such as *cumulants* and *k-statistics*.

Definition 1. For cumulants we take as the κ_n such that:

$$\exp\left(\sum_{n=1}^{\infty} \frac{\kappa_n}{n!} t^n\right) = \sum_{n=0}^{\infty} \mu'_n \frac{t^n}{n!}$$

Or, making substituting t for it , we have:

$$\begin{aligned} \exp\left(\sum_{n=1}^{\infty} \frac{\kappa_n}{n!} (it)^n\right) &= \sum_{n=0}^{\infty} \mu'_n \frac{(it)^n}{n!} \\ &= \int e^{itx} dF \\ &= \phi(t) \end{aligned}$$

Therefore κ_n is just the coefficient of $\frac{(it)^n}{n!}$ of $\ln \phi(t)$ if there exists a power series expansion.

Definition 2. A k_n -statistics is the unique symmetric unbiased estimator of the cumulant κ_n of a given distribution, i.e., is defined so that:

$$\mathbb{E}[k_n] = \kappa_n, \quad \text{with } \kappa_n \text{ a cumulant.}$$

The k_n -statistics can be given in terms of the sums of the n th powers of the data points as:

$$S_n \equiv \sum_i^N x_i^n$$

Then:

$$\begin{aligned} k_1 &= \frac{S_1}{n} \\ k_2 &= \frac{nS_2 - S_1^2}{n(n-1)} \\ k_3 &= \frac{2S_1^3 - 3nS_1S_2 + n^2S_3}{n(n-1)(n-2)} \end{aligned}$$

Computing this general expressions gives:

$$\begin{aligned} k_1 &= \bar{X}_n \\ k_2 &= \frac{n}{n-1} m_2 \\ k_3 &= \frac{n^2}{(n-1)(n-2)} m_3 \end{aligned}$$

Where m_2 is the sample variance and m_i is the i th sample central moment.

Given this we can answer the question.

Suppose that $\hat{\theta}_n$ is a biased estimator, with n the number of observations. Suppose also, that $\hat{\theta}_n$ is a function of the sample k -statistics k_i . Which are unbiased estimators of the population cumulants κ_i . Assuming all of those exist. If we expand $\hat{\theta}_n$ in a Taylor series about θ , we have:

$$\hat{\theta}_n - \theta = \sum_i (k_i - \kappa_i) \left(\frac{\partial \hat{\theta}_n}{\partial k_i} \right) + \frac{1}{2} \sum_{ij} (k_i - \kappa_i)(k_j - \kappa_j) \left(\frac{\partial^2 \hat{\theta}_n}{\partial k_i \partial k_j} \right)$$

The derivatives taken in $k_i = \kappa_i$. It follows that, taking the expected values:

$$\mathbb{E} \left[\hat{\theta}_n \right] - \theta = \sum_r \frac{a_r}{n^r} \quad (4)$$

Now it's interesting to see why this is true. Suppose that n_r is the number of times a certain value appears in a sample (with n elements), such that: $\sum_r n_r = n$. Being $\pi_r = f(x_r)$, where $f(x)$ is the pdf of the sample. Therefore, we can write:

$$\mathcal{L} \propto \prod_r \pi_r^{n_r}$$

Which means that:

$$\ln \mathcal{L} = \sum_r n_r \ln \pi_r + \text{constant}$$

Taking the derivative, we get:

$$\left. \frac{\partial \mathcal{L}}{\partial \theta} \right|_{\theta_0} = \sum_r n_r \left. \frac{\pi_r'}{\pi_r} \right|_{\theta_0} = 0 \quad (5)$$

Expanding in Taylor series:

$$\begin{aligned} \pi_r(\hat{\theta}) &= \pi_r(\theta_0) + (\hat{\theta} - \theta_0) \pi_r' + \frac{1}{2} (\hat{\theta} - \theta_0)^2 \pi_r'' + \dots \\ \pi_r'(\hat{\theta}) &= \pi_r'(\theta_0) + (\hat{\theta} - \theta_0) \pi_r'' + \frac{1}{2} (\hat{\theta} - \theta_0)^2 \pi_r''' + \dots \end{aligned} \quad (6)$$

(7)

If we set:

$$\begin{aligned}
A_i &= \sum_r \frac{[\pi'_r(\theta_0)]^{i+1}}{[\pi_r(\theta_0)]^i} \\
B_i &= \sum_r \frac{[\pi'_r(\theta_0)]^i \pi''_r(\theta_0)}{[\pi_r(\theta_0)]^i} \\
C_i &= \sum_r \frac{[\pi'_r(\theta_0)]^{i-1} [\pi''_r(\theta_0)]^2}{[\pi_r(\theta_0)]^i} \\
D_i &= \sum_r \frac{[\pi'_r(\theta_0)]^i \pi'''_r(\theta_0)}{[\pi_r(\theta_0)]^i} \\
\alpha_i &= \sum_r \frac{[\pi'_r(\theta_0)]^i \left[\frac{n_r}{n} - \pi_r(\theta_0) \right]}{[\pi_r(\theta_0)]^i} \\
\beta_i &= \sum_r \frac{[\pi'_r(\theta_0)]^{i-1} \pi''_r(\theta_0) \left[\frac{n_r}{n} - \pi_r(\theta_0) \right]}{[\pi_r(\theta_0)]^i} \\
\delta_i &= \sum_r \frac{[\pi'_r(\theta_0)]^{i-1} \pi'''_r(\theta_0) \left[\frac{n_r}{n} - \pi_r(\theta_0) \right]}{[\pi_r(\theta_0)]^i}
\end{aligned}$$

we get, combining Equations (6) with Equation (5):

$$\begin{aligned}
\alpha_1 - (A_1 + \alpha_2 - \beta_1)(\hat{\theta} - \theta_0) + \frac{1}{2}(2A_2 - 3B_1 + 2\alpha_3 - 3\beta_2 + \delta_1)(\hat{\theta} - \theta_0)^2 + \\
+ \frac{1}{6}(6A_3 - 12B_2 + 3C_1 + 4D_1)(\hat{\theta} - \theta_0)^3 + \dots = 0
\end{aligned}$$

For large n , the last equation can be inverted using Lagrange's theorem, resulting:

$$\begin{aligned}
(\hat{\theta} - \theta_0) &= A_1^{-1} \alpha_1 + A_1^{-3} \alpha_1 \left[(A_2 - \frac{3}{2} B_1) \alpha_1 - A_1 (\alpha_2 - \beta_1) \right] + \\
&+ A_1^{-5} \alpha_1 \left[(2(A_2 - \frac{3}{2} B_1)^2 - A_1 (A_3 - 2B_2 + \frac{1}{2} C_1 + \frac{2}{3} D_1)) \alpha_1^2 - \right. \\
&- 3A_1 (A_2 - \frac{3}{2} B_1) \alpha_1 (\alpha_2 - \beta_1) + \frac{1}{2} A_1^2 \alpha_1 (2\alpha_3 - 3\beta_2 + \delta_1) + \\
&\left. + A_1^2 (\alpha_2 - \beta_1)^2 \right] + \mathcal{O}(n^{-3})
\end{aligned} \tag{8}$$

Which is just what we have used in Equation (4). Therefore the bias of order $\frac{1}{n}$ is just the terms of order $\frac{1}{n}$ of Equation (8). ■

Exercise 2. Non linear least squares fit.

The decay rate of orthopositronium has been measured by atomic physicists. We have a set of data in which there is the number of decays in 10ns bins every 50ns starting at 300ns. The models consist of a background b and the traditional decay:

$$R(t) = Ae^{-\lambda t} + b$$

Use the least-squares method to find estimates of A , λ and b , with associated estimated errors (with the help of computer programs). Suppose you would like to predict the rate which would have been measured at 525ns. Find the predicted rate and its error.

Region (ns)	N	\sqrt{N}
300	803000	896
350	581083	762
400	429666	655
450	320016	566
500	242783	493
550	188487	434
600	150737	388
650	124103	352
700	105397	325
750	92748	305

We used IDL to solve this problem. The program used is transcribed here:

```

;-----
;
; Program to adjust, by chi^2 minimization, the function
; R(t)=A*exp(-lambda*t)+b
;
; Should be initialized with: [1000000,0.000001,0].
; the syntax to use this program is something like:
; ex2,[start_A,star_lambda,star_b]
;-----
;Defines the function to be adjusted
;-----
function func,x,P

temp = P(0)*exp(-P(1)*x)+P(2)
return,temp

end
;-----
;
;Main program
;
;-----
pro ex2,start
;
; Data to be adjusted
;
; Region N
; 300, 803000,
; 350, 581083,
; 400, 429666,
; 450, 320016,
; 500, 242783,
; 550, 188487,
; 600, 150737,
; 650, 124103,
; 700, 105397,
; 750, 92748]

; Definition of the vector that stores the independent variables
x_i = [ 300d, 350d, 400d, 450d, 500d, 550d, 600d,

```

```

        650d, 700d, 750d]
; Offset of the previous vector, so that the independent variables
; falls in the middle
; of the measurement
x = x_i+25d
; Definition of the vector that stores the dependent variable
n = [803000d,581083d,429666d,320016d,242783d,188487d,150737d,
    124103d,105397d,92748d]

; Definition of the sigma error
err = sqrt(n)

; Main function to perform the fit
result = MPFITFUN('func', x, n, err, start,covar=mat_temp)

; This matrix stores the covariance matrix of the fit, the square
; root of the diagonal elements is the error on the corresponding
; parameter
mat=sqrt(mat_temp)

; Computation of the rate at t=525ns
medida=func(525,result)/50
; It's error by quadratic propagation
erro_medida=sqrt(((mat[2,2])^2)+((exp(-result[1]*525)*mat[0,0])^2)
    +((result[1]*result[0]*exp(-result[1]*525)*mat[1,1])^2))

;Display of the results
print,'A=      ',result[0],',', sigma=',mat[0,0]
print,'lambda=',result[1],',', sigma=',mat[1,1]
print,'b      ',result[2],',', sigma=',mat[2,2]

print,'Valor do 525=',medida, ', sigma=',erro_medida/50

set_plot,'ps'
device,/encapsulated,filename='ex2.eps'
ploterr,x,n,err
x=indgen(500)+300
oplot,x,func(x,result),thick=3
text = "A =" +string(result[0])+" +/-" +string(mat[0,0])
xyouts,450,8e5,text
text = "lambda =" +string(result[1])+" +/-" +string(mat[1,1])
xyouts,450,7.5e5,text
text = "b =" +string(result[2])+" +/-" +string(mat[2,2])
xyouts,450,7e5,text
device,/close
set_plot,'x'

end

```

With this, as you could see running the program, we obtained the results presented on Table (1): For the computation of the predicted rate at 525ns we used, for the errors the formula of quadratic

Param.	Value	σ
A	$6.33 \times 10^6 (50\text{ns})^{-1}$	$0.03 \times 10^6 (50\text{ns})^{-1}$
λ	$7.03 \times 10^{-3} \text{ns}^{-1}$	$0.02 \times 10^{-3} \text{ns}^{-1}$
b	$6.11 \times 10^4 (50\text{ns})^{-1}$	$0.04 \times 10^4 (50\text{ns})^{-1}$

Table 1: Results of χ^2

propagation:

$$\sigma_{t=525\text{ns}} = \sqrt{\left(\left(\frac{\partial R}{\partial A}\sigma_A\right)\right)_{t=525\text{ns}}^2 + \left(\left(\frac{\partial R}{\partial \lambda}\sigma_\lambda\right)\right)_{t=525\text{ns}}^2 + \left(\left(\frac{\partial R}{\partial b}\sigma_b\right)\right)_{t=525\text{ns}}^2}$$

Which gave the result $R(t = 525\text{ns}) = 4.86 \times 10^3 \pm 0.02 \times 10^3 \text{ns}^{-1}$.

The result of the fit is presented in Figure (1).

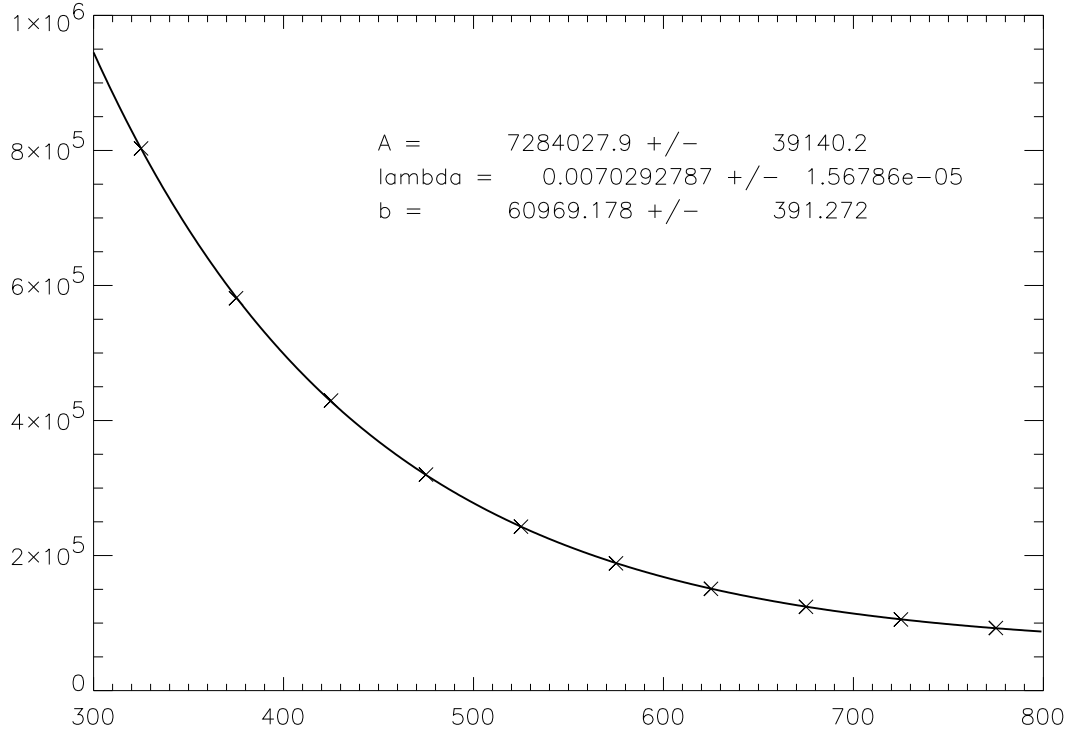


Figure 1: Non-linear fit of exponential decay.

■