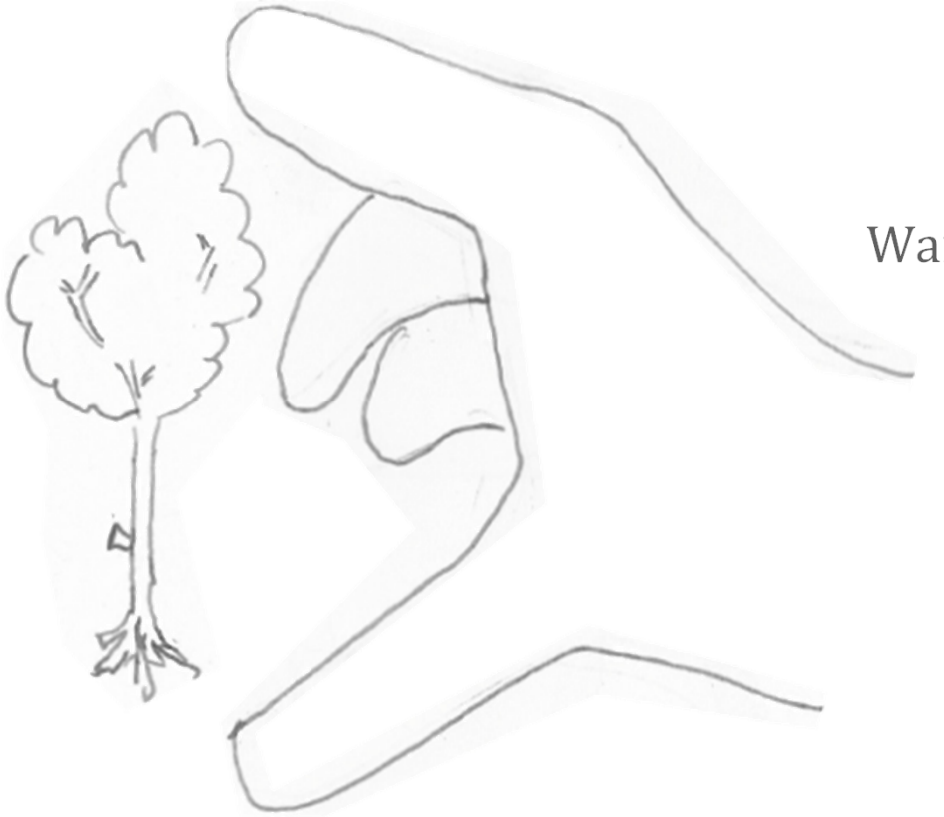
Summarizing Performance Data
Part 2



Important

Easy to Difficult

Warning: some mathematical content

Estimation Theory again

- We will see some results that are a little more complicated to formulate but are interesting due to their generality
- We will also see the impact of non iid-ness

Parametric Estimation Theory

Consider a data set x_i, $i = 1..., n$, that we view as the realization of a stochastic system (in other words, the output of a simulator). The framework of parametric estimation theory consists in assuming that θ is fixed, but unknown. We usually assume that the model has a density of probability, and that the density of probability that the output is $x_1, ..., x_n$ depends on the parameter θ; we denote it with $f(x_1, ..., x_n|\theta)$. It is also called the *likelihood* of the observed data. An *estimator* of θ is any function $T()$ of the observed data. A good estimator is one such that, in average, $T(X_1, ..., X_n)$ is "close" to the true value θ.

- *Unbiased estimator*: $\mathbb{E}_\theta(T(X)) = \theta$. For example, the estimator $\hat{\sigma}_n^2$ of variance of a normal iid sample given by Theorem 2.3.1 is unbiased.
- *Consistent family of estimators*: $\mathbb{P}_\theta(|T(X) - \theta|) > \epsilon) \to 0$ when the sample size n goes to ∞. For example, the estimator $(\hat{\mu}_n, \hat{\sigma}_n^2)$ of Theorem 2.3.1 is consistent. This follows from the weak law of large numbers.

A commonly used method for deriving estimators is that of *Maximum Likelihood*. The maximum likelihood estimator is the value of θ that maximizes the likelihood $f(x_1, ..., x_n|\theta)$. This definition makes sense if the maximum exists and is unique, which is often true in practice. A formal set of conditions is the regularity condition in Definition B.2.1.

In Section B.2, we give a result that shows that the MLE for an i.i.d. sample with finite variance is asymptotically unbiased, i.e. the bias tends to 0 as the sample size increases. It is also consistent.

EXAMPLE 2.2: MLE FOR I.I.D. NORMAL DATA. Consider a sample $(x_1, ..., x_n)$ obtained from a normal i.i.d. random vector $(X_1, ..., X_n)$. The likelihood is given by Eq.(B.1). We want to maximize it, where $x_1, ..., x_n$ are given and $\mu, v = \sigma^2$ are the variables. For a given v, the maximum is reached when $\mu = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Let μ have this value and find the value of v that maximizes the resulting expression, or to simplify, the log of it. We thus have to maximize

$$-\frac{n}{2}\ln v - \frac{1}{2v}S_{x,x} + C \tag{B.2}$$

where $S_{x,x} \stackrel{\text{def}}{=} \sum_{i=1}^{n}(x - \hat{\mu}_n)^2$ and C is a constant with respect to v. This is a simple maximization problem in one variable v, which can be solved by computing the derivative. We find that there is a maximum for $v = \frac{S_{x,x}}{n}$. The maximum likelihood estimator of (μ, v) is thus precisely the estimator in Theorem 2.2.2.

We say that an estimation method *invariant by re-parametrization* if the following holds. Assume the method produces some estimator $T(X)$ for θ. Assume we re-parametrize the problem by considering that the parameter is $\phi(\theta)$, where ϕ is some invertible mapping. For example, a normal iid sample can be parametrized by $\theta = (\mu, \sigma^2)$ or by $\phi(\theta) = (\mu, \sigma)$.

QUESTION 2.8.2. *What is the mapping ϕ in this case ?* [7]

[7] $\phi(x, y) = (x, \sqrt{y})$ defined for $x \in \mathbb{R}$ and $y \geq 0$.

The maximum likelihood method *is* invariant by re-parametrization. This is because the property of being a amximum is invariant by re-parametrization. It is an important property in our context, since the model is usually not given a priori, but has to be invented by the performance analyst.

A method that provides an unbiased estimator cannot be invariant by re-parametrization, in general. For example, $(\hat{\mu}_n, \hat{\sigma}_n^2)$ of Theorem 2.3.1 is an unbiased estimator of (μ, σ^2), but $(\hat{\mu}_n, \hat{\sigma}_n)$ is a **biased** estimator of (μ, σ) (because usually $\mathbb{E}(S)^2 \neq \mathbb{E}(S^2)$ except if S is non-random). Thus, the property of being unbiased is incompatible with invariance by re-parametrization, and may thus be seen as an inadequate requirement for an estimator.

B.1.3 Efficiency and Fisher Information

The *efficiency* of an estimator $T(\vec{X})$ of the parameter θ is defined as the expected square error $\mathbb{E}_\theta(\left\|T(\vec{X}) - \theta\right\|^2)$ (here we assume that θ takes values in some space Θ where the norm is defined). The efficiency that can be reached by an estimator is captured by the concept of Fisher information, which we now define. Assume first to simplify that $\theta \in \mathbb{R}$. The *observed information* is defined by

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}$$

where $l(\theta)$ is the *log-likelihood*, defined by

$$l(\theta) = \ln \text{lik}(\theta) = \ln f(x_1, ..., x_n|\theta)$$

The *Fisher information*, or *expected information* is defined by

$$I(\theta) = \mathbb{E}_\theta(J(\theta)) = \mathbb{E}_\theta\left(-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right)$$

In general, the parameter θ is multi-dimensional, i.e., varies in an open subset Θ of $\mathbb{R}^k$. Then J and I are symmetric matrices defined by

$$[J(\theta)]_{i,j} = -\frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j}$$

and

$$[I(\theta)]_{i,j} = -\mathbb{E}_\theta \left(\frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j} \right)$$

The Cramer-Rao theorem says that the efficiency of any **unbiased** estimator is lower bounded by $\frac{1}{I(\theta)}$. Further, under the conditions in Definition 2.8.1, the MLE for an iid sample is asymptotically maximally efficient, i.e. $\mathbb{E}\left(\|T(X) - \theta\|\right) / I(\theta)$ tends to 1 as the sample size goes to infinity.

EXAMPLE 2.13: FISHER INFORMATION OF NORMAL IID MODEL. Assume $(X_i)_{i=1...n}$ is iid normal with mean μ and variance σ^2. The observed information matrix is computed from the likelihood function; we obtain:

$$J = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{2n}{\sigma^3}(\hat{\mu}_n - \mu) \\ \frac{2n}{\sigma^3}(\hat{\mu}_n - \mu) & \frac{-n}{\sigma^2} + \frac{3}{\sigma^4}\left(S_{xx} + n(\hat{\mu}_n - \mu)^2\right) \end{pmatrix}$$

and the expected information matrix (Fisher's information) is

$$I = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}$$

2.8.4 Asymptotic Confidence Intervals

Here we need to assume some regularity conditions. Assume the sample comes from an iid sequence and further, that the following regularity conditions are met.

THEOREM B.2.1. *Under the conditions in Definition B.2.1, the MLE exists, converges almost surely to the true value. Further* $I(\theta)^{\frac{1}{2}}(\hat{\theta}-\theta)$ *converges in distribution towards a standard normal distribution, as* n *goes to infinity. It follows that, asymptotically:*

1. *the distribution of* $\hat{\theta}-\theta$ *can be approximated by* $N\left(0, I(\hat{\theta})^{-1}\right)$ *or* $N\left(0, J(\hat{\theta})^{-1}\right)$
2. *the distribution of* $2\left(l(\hat{\theta})-l(\theta)\right)$ *can be approximated by* χ^2_k *(where* k *is the dimension of* Θ*).*

The quantity $2\left(l(\hat{\theta})-l(\theta)\right)$ is called the *likelihood ratio statistic*.

COROLLARY B.2.1 (Asymptotic Confidence Intervals). *When n is large, approximate confidence intervals can be obtained as follows:*

1. *For the ith coordinate of θ, the interval is: $\hat{\theta}_i \pm \eta\sqrt{\left[I(\hat{\theta})^{-1}\right]_{i,i}}$ or $\hat{\theta} \pm \eta\sqrt{\left[J(\hat{\theta})^{-1}\right]_{i,i}}$, where $N_{0,1}(\eta) = \frac{1+\gamma}{2}$ (for example, with $\gamma = 0.95$, $\eta = 1.96$).*
2. *If θ is in $\mathbb{R}$: the interval can be defined implicitly as $\{\theta : l(\hat{\theta}) - \frac{\xi}{2} \leq l(\theta) \leq l(\hat{\theta})\}$, where $\chi^2_1(\xi) = \gamma$. For example, with $\gamma = 0.95$, $\xi = 3.84$.*

EXAMPLE 2.15: LAZY NORMAL IID. Assume our data comes from an iid normal model X_i, $i = 1, \ldots n$. We compare the exact confidence interval for the mean (from Theorem 2.3.1) to the approximate ones given by the corollary.

The MLE of (μ, σ) is $(\hat{\mu}_n, s_n)$. The exact confidence interval is

$$\hat{\mu}_n \pm \eta' \frac{\hat{\sigma}_n}{\sqrt{n}}$$

with $\hat{\sigma}^2_n = S_{x,x}/(n-1)$ and $t_{n-1}(\eta') = \frac{1+\gamma}{2}$.

Now we compute the approximate confidence interval obtained from the Fisher information. We have

$$I(\mu,\sigma)^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}$$

thus the distribution of $(\mu - \hat{\mu}_n, \sigma - s_n)$ is approximately normal with 0 mean and covariance matrix $\begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}$. It follows that $\mu - \hat{\mu}_n$ is approximately $N(0, \frac{s_n^2}{n})$, and an approximate confidence interval is

$$\hat{\mu}_n \pm \eta \frac{s_n}{\sqrt{n}}$$

with $s_n = s_{x,x}/n$ and $N_{0,1}(\eta) = \frac{1+\gamma}{2}$.

Thus the use of Fisher information gives the same asymptotic interval for the mean as Theorem 2.3.2. This is quite general: the use of Fisher information is the generalization of the large sample asymptotic of Theorem 2.3.2.

We can also compare the approximate confidence interval for σ. The exact interval is given by Theorem 2.3.1: with probability γ we have

$$\frac{\xi_2}{n-1} \leq \frac{\hat{\sigma}_2^n}{\sigma^2} \leq \frac{\xi_1}{n-1}$$

With Fisher information, we have that $\sigma - s_n$ is approximately $N_{0,\frac{\sigma^2}{2n}}$ Thus with probability γ

$$|\sigma - s_n| \leq \eta \frac{\sigma}{\sqrt{2n}}$$

with $N_{0,1}(\eta) = \frac{1+\gamma}{2}$.

Divide by σ and obtain, after some algebra, that with probability γ:

$$\frac{1}{1+\frac{\eta}{\sqrt{2n}}} \leq \frac{\sigma}{s_n} \leq \frac{1}{1-\frac{\eta}{\sqrt{2n}}}$$

Fisher Method for CI for σ

n	30	60	120
Exact	$0.7964 - 1.3443$	$0.8476 - 1.2197$	$0.8875 - 1.1454$
Fisher	$0.7847 - 1.3162$	$0.8411 - 1.2077$	$0.8840 - 1.1401$

2.8.5 CONFIDENCE INTERVAL IN PRESENCE OF NUISANCE PARAMETERS

In many cases, the parameter has the form $\theta = (\mu, \nu)$, and we are interested only in μ (for example, for a normal model: the mean) while the remaining element ν, that still need to be estimated, is considered a nuisance (for example: the variance). In such cases, we can use the following theorem

THEOREM B.3.1 ([32]). *Under the conditions in Definition B.2.1, assume that $\Theta = M \times N$, where M, N are open subsets of $\mathbb{R}^p, \mathbb{R}^q$. Thus the parameter is $\theta = (\mu, \nu)$ with $\mu \in M$ and $\nu \in N$ (p is the "dimension", or number of degrees of freedom, of μ).*
For any μ, let $\hat{\nu}_\mu$ be the solution to

$$l(\mu, \hat{\nu}_\mu) = \max_\nu l(\mu, \nu)$$

and define the profile log likelihood pl by

$$pl(\mu) \stackrel{\text{def}}{=} \max_\nu l(\mu, \nu) = l(\mu, \hat{\nu}_\mu)$$

Let $(\hat{\mu}, \hat{\nu})$ be the MLE. If (μ, ν) is the true value of the parameter, the distribution of $2\left(pl(\hat{\mu}) - pl(\mu)\right)$ tends to χ^2_p.
An approximate confidence region for μ at level γ is

$$\{\mu \in M : pl(\mu) \geq pl(\hat{\mu}) - \frac{1}{2}\xi\}$$

where $\chi^2_p(\xi) = \gamma$.

EXAMPLE 2.16: LAZY NORMAL IID REVISITED. Consider the log of the data in Figure 2.3, which appears to be normal. The model is $Y_i \sim iidN_{\mu,\sigma^2}$ where Y_i is the log of the data. Assume we would like to compute a confidence interval for μ but are too lazy to apply the exact student statistic in Theorem 2.3.1.

For any μ, we estimate the nuisance parameter σ, by maximizing the log-likelihood:

$$l(\mu,\sigma) = -\frac{1}{2}\left(n\ln\sigma^2 + \frac{1}{\sigma^2}\sum_i (Y_i-\mu)^2\right)$$

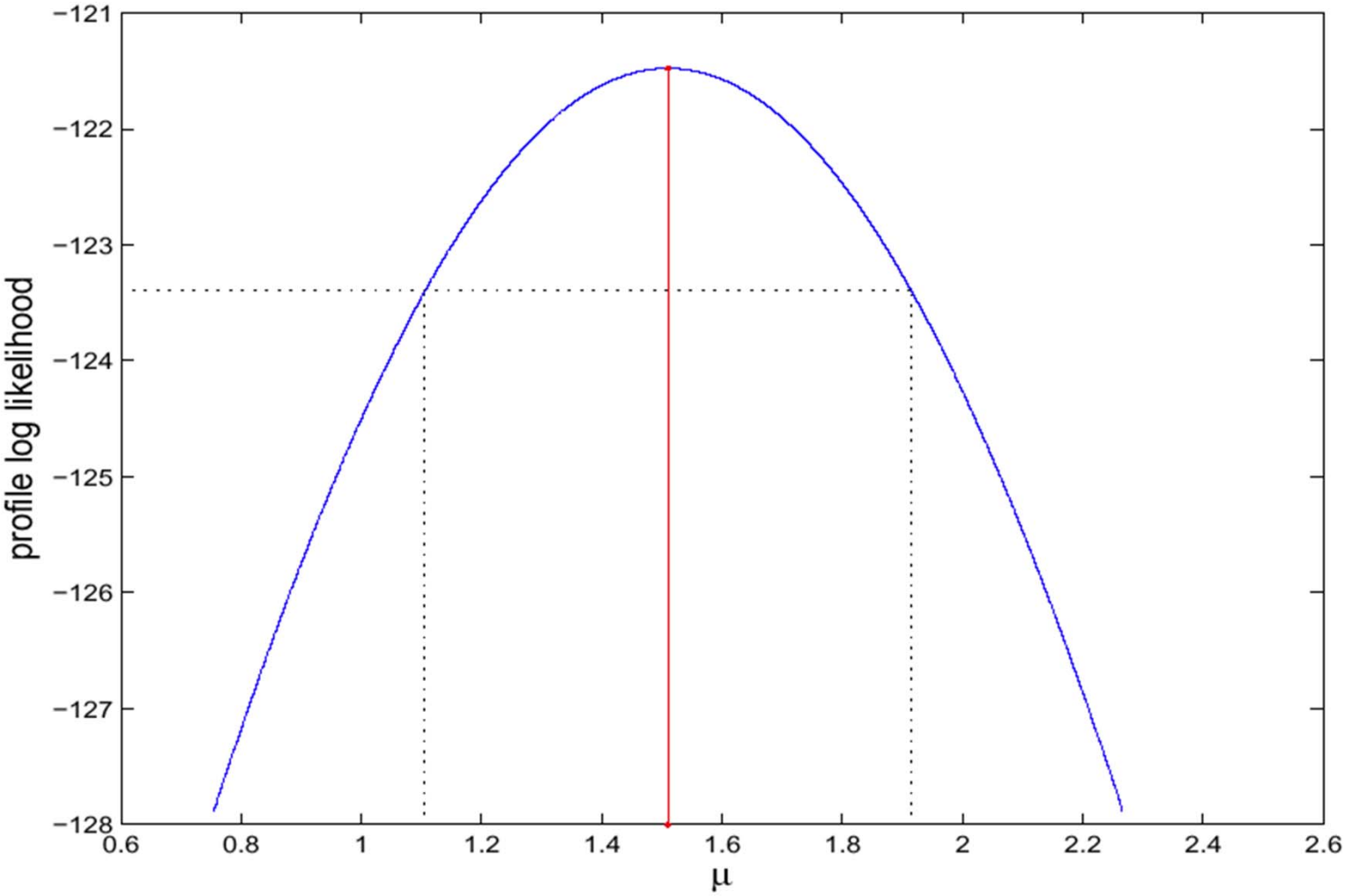
It comes

$$\hat{\sigma}^2_\mu = \frac{1}{n}\sum_i (Y_i-\mu)^2 = \frac{1}{n}S_{YY} + (\bar{Y}-\mu)^2$$

and thus

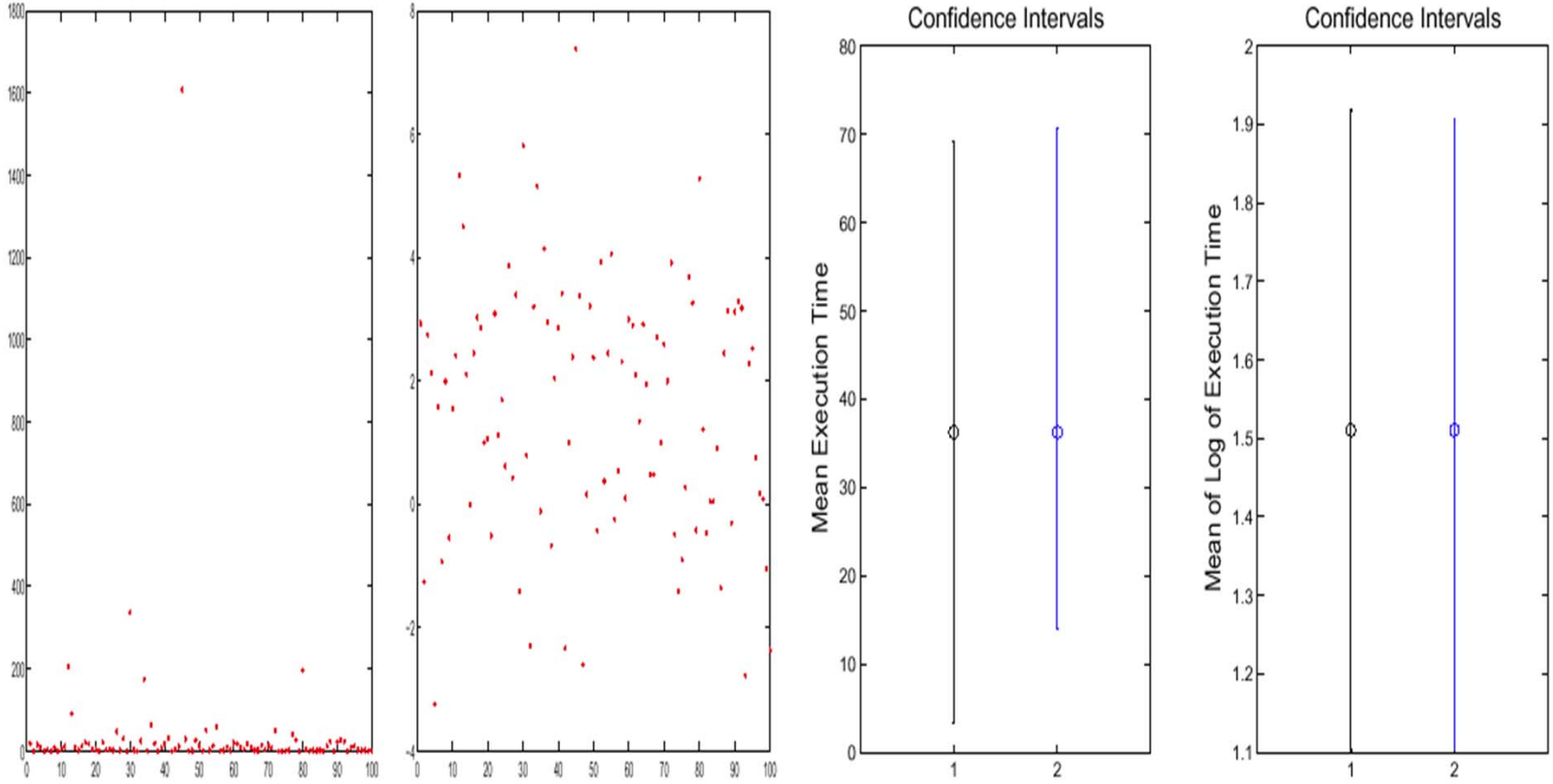
$$pl(\mu) := l(\mu,\hat{\sigma}_\mu) = -\frac{n}{2}(\ln\hat{\sigma}^2_\mu + 1)$$

Profile Log Likelihood Method



On Figure 2.12 we plot $pl(\mu)$. We find $\hat{\mu} = 1.510$ as the point that maximizes $pl(\mu)$. A 95%-confidence interval is obtained as the set $\{pl(\mu) \geq pl(\hat{\mu}) - \frac{1}{2}3.84\}$. We obtain the interval $[1.106, 1.915]$. Compare to the exact confidence interval obtained with Theorem 2.3.1, which is equal to $[1.103, 1.918]$: the difference is negligible.

Find the Box-Cox Exponent



(a) (Data) (b) (Log of data) (c) (Confidence Intervals)

EXAMPLE 2.17: RE-SCALING. Consider the data in Figure 2.3, which does not appear to be normal in natural scale, and for which we would like to do a Box-Cox transformation. We would like a confidence interval for the exponent of the transformation.

The transformed data is $Y_i = b_s(X_i)$, and the model now assumes that Y_i is iid $\sim N_{\mu,\sigma^2}$. We take the unknown parameter to be $\theta = (\mu, \sigma, s)$. The distribution of X_i, under θ is:

$$f_{X_i}(x|\theta) = b_s'(x) f_{Y_i}\left(b_s(x)|\mu,\sigma\right) = x^{s-1} h(b_s(x)|\mu,\sigma^2)$$

where $h(x|\mu,\sigma^2)$ is the density of the normal distribution with mean μ and variance σ^2.

The log-likelihood is

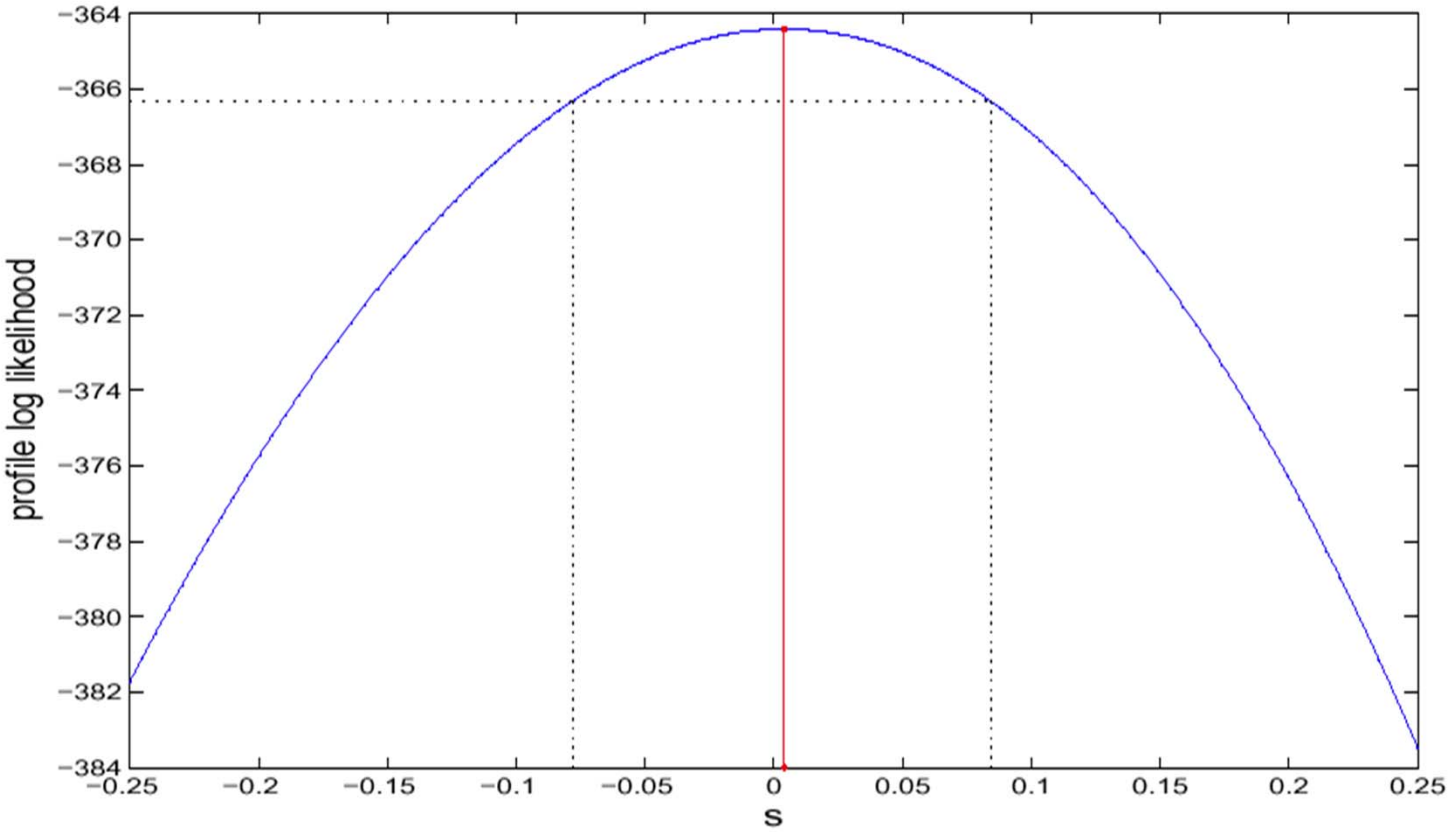
$$l(\mu,\sigma,s) = C - n\ln\sigma + \sum_i \left((s-1)\ln x_i - \frac{(b_s(x_i)-\mu)^2}{2\sigma^2} \right)$$

where C is some constant (independent of the parameter). For a fixed s it is maximized by the MLE for a Gaussian sample

$$\hat{\mu}_s = \frac{1}{n}\sum_i b_s(x_i)$$

$$\hat{\sigma}_s^2 = \frac{1}{n}\sum_i (b_s(x_i) - \hat{\mu})^2$$

CI for Box-Cox Exponent



We can use a numerical estimation to find the value of s that maximizes $l(\hat{\mu}_s, \hat{\sigma}_s, s)$; see Figure 2.13 for a plot. The estimated value is $\hat{s} = 0.0041$, which gives $\hat{\mu} = 1.5236$ and $\hat{\sigma} = 2.0563$.

We now give a confidence interval for s, using the asymptotic result in Theorem 2.8.2. A 95% confidence interval is readily obtained from Figure 2.13, which gives the interval $[-0.0782, 0.0841]$.

Fisher Information Matrix for (μ, σ, s)

$$J = \begin{pmatrix} 23.7 & 0 & -77.1 \\ 0 & 47.3 & -146.9 \\ 77.1 & -146.9 & 1291.1 \end{pmatrix}$$

$$J^{-1} = \begin{pmatrix} 0.0605 & 0.0173 & 0.0056 \\ 0.0173 & 0.0377 & 0.0053 \\ 0.0056 & 0.0053 & 0.0017 \end{pmatrix}$$

$$\hat{s} \pm 1.96\sqrt{0.0017} = [-0.0770, 0.0852]$$

EXAMPLE: JOE'S SHOP - ESTIMATION OF ξ. In Example 4.3 on page 94 we assumed that the value ξ after which there is congestion collapse is known in advance. Now we relax this assumption. Our model is now the same as Equation (4.8), except that ξ is also now a parameter to be estimated.

EXAMPLE 4.3: JOE'S SHOP AGAIN, FIGURE 1.1(B). We assume that there is a threshold ξ beyond which the throughput collapses (we take $\xi = 70$). The statistical model is

$$Y_i = (a + bx_i)1_{x_i \leq \xi} + (c + d\xi)1_{\{x_i > \xi\}} + \epsilon_i \tag{4.8}$$

where we impose

$$a + b\xi = c + d\xi \tag{4.9}$$

To do this, we apply maximum likelihood estimation. We have to maximize the log-likelihood $l_{\vec{y}}(a, b, d, \xi, \sigma)$, wher $\vec{y}$, the data, is fixed. For a fixed ξ, we know the value of (a, b, d, σ) that achieves the maximum, as we have a linear regression model. We plot the value of this maximum versus ξ (Figure **??**) and numerically find the maximum. It is for $\xi = 77$.

To find a confidence interval, we use the asymptotic result in Theorem 2.8.2. It says that a 95% confidence interval is obtained by solving $l(\hat{\xi}) - l(\xi) \leq 1.9207$, which gives $\xi \in [73, 80]$.

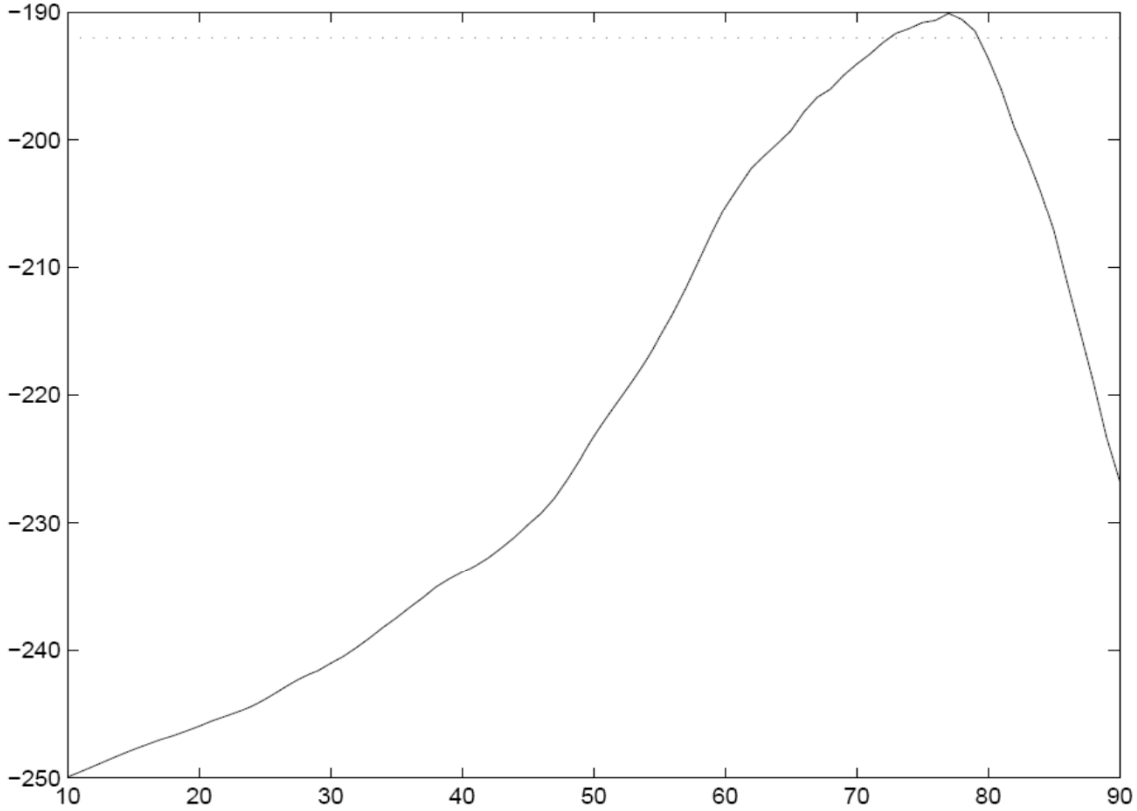
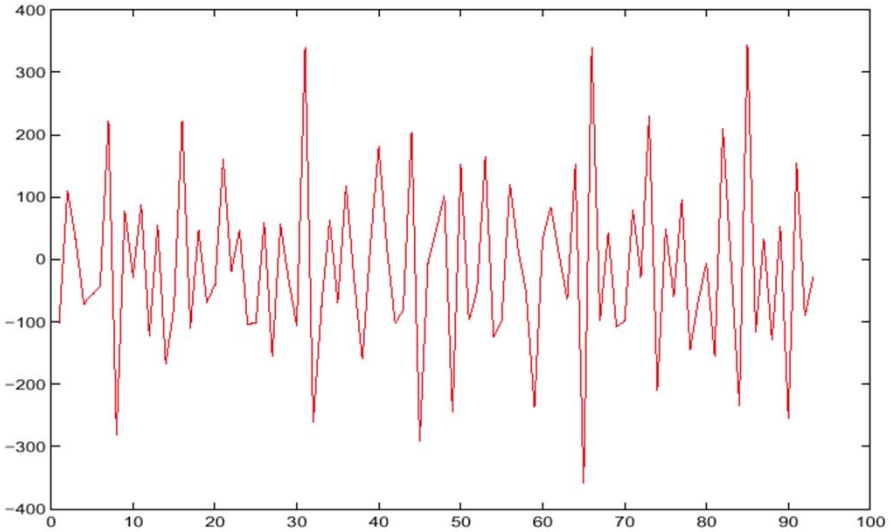


Figure 4.1: Log likelihood for Joes' shop as a function of ξ.

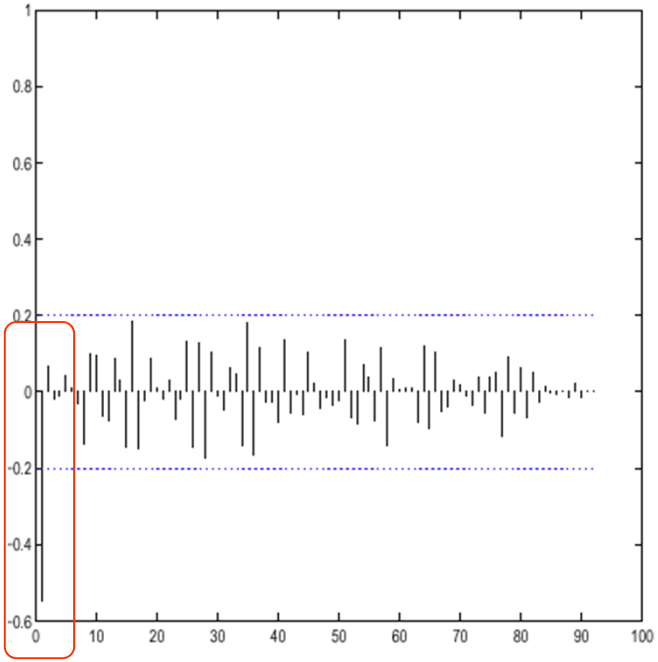
EXAMPLE 2.4: JOE'S BALANCE DATA. Joe's shop sells online access to visitors who download electronic content on their smartphones. At the end of day $t-1$, Joe's employee counts the amount of cash c_{t-1} present in the cash register and puts it into the safe. In the morning of day t, the cash amount c_{t-1} is returned to the cash register. The total amount of service sold (according to bookkeeping data) during day t is s_t. During the day, some amount of money b_t is sent to the bank. At the end of day t, we should have $c_t = c_{t-1} + s_t - b_t$. However, there are always small errors in counting the coins, in bookkeeping and in returning change. Joe computes the balance $Y_t = c_t - c_{t-1} - s_t + b_t$ and would like to know whether there is a systematic source of errors (i.e. Joe's employee is losing money, maybe because he is not honest, or because some customers are not paying for what they take). The data for Y_t is shown on Figure 2.10. The sample mean is $\mu = -13.95$, which is negative. However, we need a confidence interval for μ before risking any conclusion.

If we would assume that the errors Y_t are iid, then a confidence interval would be given by Theorem 2.2.2 and we find approximately $[-43, 15]$. Thus, with the iid model, we cannot conclude that there is a fraud.

Non-IID Data: Joe's Cash Register

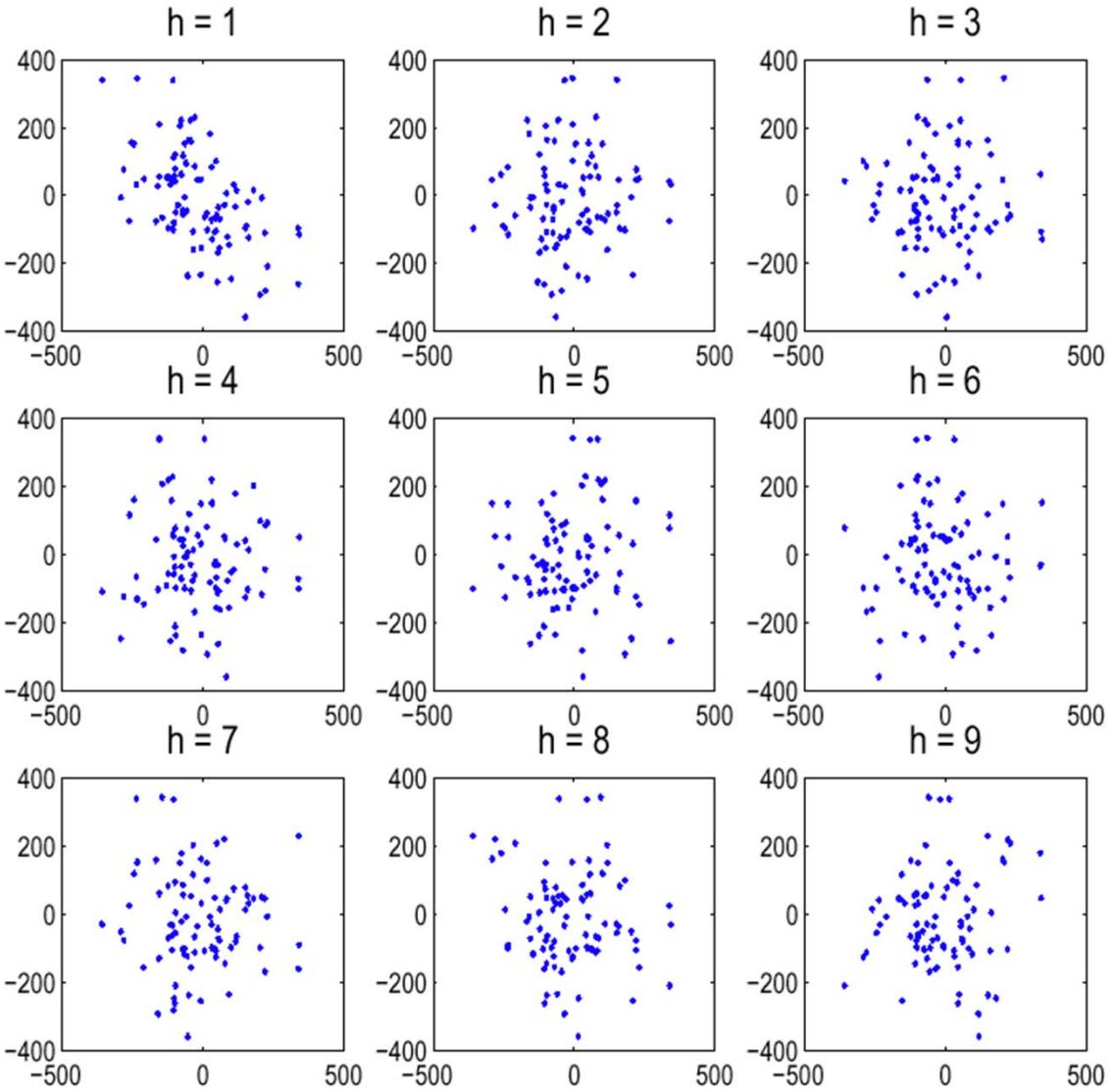


(a) Data



(c) Autocorrelation

Lag Plots show Some Correlation at lag 1



CONFIDENCE INTERVAL FOR BALANCE ASSUMING IID MODEL. If we would assume that the errors Y_t are iid, then a confidence interval would be given by Theorem 2.3.2. In fact, the qqplot indicates that the data looks normal so we can use the student statistic in Theorem 2.3.1: the sample standard deviation is $S = 141.6$, so the 95%-confidence interval is $-13.95 \pm \eta S/\sqrt{n} \approx [-43, 15]$, where n is the sample size and $\eta = 1.986$. Thus, with the iid model, we cannot conclude that there is a fraud.

However, we need to verify the iid assumption before giving an interpretation. The data appears to be stationary (no trend or seasonal behaviour) thus we can use the ACF diagram. Figure **??** shows that there is a strong correlation at lag 1. This is confirmed by the lag plot. Thus, we can conclude that the iid assumption does not hold for this data set.

CONFIDENCE INTERVAL WITH MOVING AVERAGE MODEL. To go further, we need a valid model. Assume that the coin counting and bookkeeping processes have random, independent errors:

$$C_t = c_t + \epsilon_t \tag{2.27}$$

$$S_t - r_t = s_t - r_t + \epsilon'_t \tag{2.28}$$

where upper case if for reported (observed) values and lower case for the true (non observed) values. Also assume that there is an external flow of money $\mu + \epsilon_t$" every day (a negative μ is a loss of money). Assume that all ϵs are iid and independent of each other. Then we have

$$Y_t := C_t - C_{t-1} = c_t - c_{t-1} - s_t + r_t + \epsilon_t - \epsilon_{t-1} - \epsilon'_t$$

and

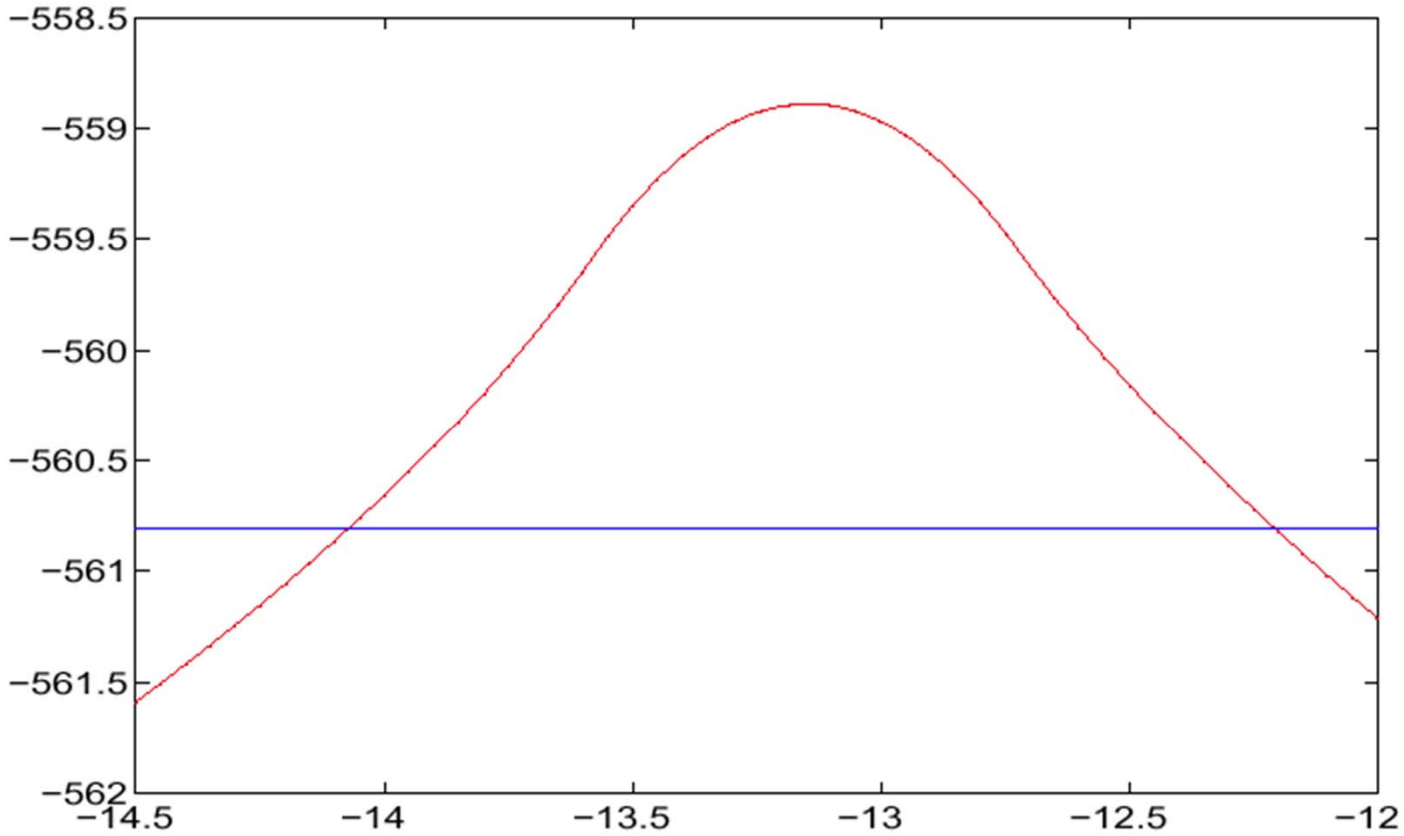
$$c_t = c_{t-1} - s_t + r_t + \mu + \epsilon_t"$$

It follows that

$$Y_t = \mu + \epsilon_t" + \epsilon_t - \epsilon_{t-1} - \epsilon'_t$$

lags 0 and 1 are said to be moving average processes of order 1 (in short, MA(1)). Thus, $Y_t - \mu$ is an MA(1) process.

The general method of maximum likelihood estimation in Section 2.8 applies, as we see now. We are interested in obtaining a confidence interval for μ. We use the MLE asymptotic in Theorem 2.8.2 on Page 44.



$pl(\mu))$ is approximately χ^2_1. Figure 2.15 shows a plot of $l(\mu)$. It follows that $\hat{\mu} = -13.2$ and an approximate 95%-confidence interval is $[-14.1, -12.2]$. Contrary to the iid model, this suggests that there *is* a loss of money, in average 13€ per day.