Model Fitting
Part 1

Jean-Yves Le Boudec
Contents

1. What is model fitting?
   2. Least Squares
   3. $\ell^1$ norm minimization
   4. The bootstrap
Reminder

Mini Project definitions are due next Monday
Virus Infection Data

- We would like to capture the growth of infected hosts (explanatory model)

- An exponential model seems appropriate

- How can we fit the model, in particular, what is the value of $\alpha$?

\[ Y(t) = ae^{\alpha t} \]
Least Square Fit of Virus Infection Data

\[ \alpha = 0.5173 \]

Mean doubling time 1.34 hours

Prediction at +6 hours: 100 000 hosts
Least Square Fit of Virus Infection Data In Log Scale

\[ \alpha = 0.39 \]

Mean doubling time 1.77 hours

Prediction at +6 hours: 39 000 hosts
Compare the Two

LS fit in natural scale

LS fit in log scale
Which scale should I use?

A. The natural scale because it is the simplest (principle of parsimony)
B. The log scale because it is better adapted to exponential growth
C. I don’t know
Solution

It is not possible to answer this question simply by looking at the scale.

We need to understand *what* model fitting really is.
What is model fitting?

- Given some data $Y_i$ and a parametric model $f_i(\theta)$ we assume that:

$$Y_i = f_i(\theta) + \text{noise}$$

- The model fitting problem is to estimate the unknown parameter $\theta$

For example: the model is $Y_i = ae^{\alpha t_i}$

or

$$\log Y_i = \log a + \alpha t_i$$

These two models are equivalent; the parameter is $\theta = (a, \alpha)$
The signal processing interpretation

Given a model $Y_i = f_i(\theta) + \text{noise}$ and a score function, find $\theta$ that minimizes the score.

Classical scores

- Least Square: $\text{score} = \sum_i (\text{noise}_i)^2$
- Weighted Least Square: $\text{score} = \sum_i (w_i \times \text{noise}_i)^2$
- $\ell^1$ score: $\text{score} = \sum_i |\text{noise}_i|$
- Weighted $\ell^1$ score: $\text{score} = \sum_i w_i |\text{noise}_i|$

Which one should we use?
The statistical interpretation

The data $Y_i$ is output by a simulator, with parameter $\theta$, which we want to estimate.

This forces us to make some assumption about the *noise* term; for example, we could consider any one of the following models:

1. $Y_i = ae^{\alpha t_i} + \epsilon_i$, $\epsilon_i \text{ iid } \sim N(0, \sigma^2)$
2. $Y_i = ae^{\alpha t_i}(1 + \epsilon_i)$, $\epsilon_i \text{ iid } \sim N(0, \sigma^2)$
3. $\log Y_i = \log a + \alpha t_i + \epsilon_i$, $\epsilon_i \text{ iid } \sim N(0, \sigma^2)$

For each of these models, the parameter is now $\theta = (a, \alpha, \sigma)$.
Estimating a statistical model: maximum likelihood

- Principle: find $\theta$ that maximizes the pdf $f_Y(y|\theta)$, also called likelihood, where $y$ is the data

- This is a robust and consistent estimation method

- Example: model 1: $Y_i = ae^{at_i} + \varepsilon_i$, $\varepsilon_i$ iid $\sim N(0, \sigma^2)$

$$f_Y(y|a, \alpha, \sigma) = \frac{1}{(\sqrt{2\pi\sigma})^l} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{l}(y_i-ae^{at_i})^2}$$

the problem is to find $(a, \alpha, \sigma)$ that maximizes 

(1)
Theorem 3.1.1 (Least Squares). For the model in Eq. (3.5),

1. the maximum likelihood estimator of the parameter $(\beta, \sigma)$ is given by:

   (a) $\hat{\beta} = \arg \min_{\beta} \sum_i \left( y_i - f_i(\beta) \right)^2$

   (b) $\hat{\sigma}^2 = \frac{1}{I} \sum_i \left( y_i - f_i(\hat{\beta}) \right)^2$

i.e. the signal processing method with Least Squares score is the same as the statistical method with iid gaussian noise.
Example: model 1: \[ Y_i = ae^{\alpha t_i} + \epsilon_i, \quad \epsilon_i \text{ iid } \sim N(0, \sigma^2) \]

\[ f_Y(y|a, \alpha, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^l} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{l} (y_i - ae^{\alpha t_i})^2} \]  

(1)

the problem is to find \((a, \alpha, \sigma)\) that maximizes (1)

Thm 3.3.1⇒

\[ (\hat{a}, \hat{\alpha}) = \text{argmin}_{a, \alpha} \sum_{i=1}^{l} (y_i - ae^{\alpha t_i})^2 \]

\[ \hat{\sigma}^2 = \frac{1}{l} \sum_{i} (y_i - \hat{a} e^{\hat{\alpha} t_i})^2 \]
The residuals can help validate a fitting method

Residuals = estimate of noise terms (after estimation of parameter) – they should be consistent with the model

Model 1:
\[ Y_i = a e^{\alpha t_i} + \epsilon_i, \epsilon_i \sim iid \ N(0, \sigma^2) \]

Model 3:
\[ \log Y_i = \log a + \alpha t_i + \epsilon_i, \epsilon_i \sim iid \ N(0, \sigma^2) \]

Model 1 appears not to fit, residuals have a variance that grows with \( t_i \), model 3 appears OK; we should prefer it for fitting this data set
Given a model to be fitted

\[ Y_i = f_i(\theta) + \text{noise} \]

we have the equivalences:

<table>
<thead>
<tr>
<th>Score</th>
<th>Statistical model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least Square:</td>
<td>noise \sim iid N(0, \sigma^2)</td>
</tr>
<tr>
<td>score = \sum_i (\text{noise}_i)^2</td>
<td></td>
</tr>
<tr>
<td>Weighted Least Square:</td>
<td>noise_i \sim \text{ind } N \left(0, \frac{\sigma^2}{w_i^2}\right)</td>
</tr>
<tr>
<td>score = \sum_i (w_i \times \text{noise}_i)^2</td>
<td></td>
</tr>
</tbody>
</table>

i.e. WLS gives a weight inversely proportional to \( \sigma_i \) of noise
Which score corresponds to which model?

A. A1 B2 C3
B. A1 B3 C2
C. A2 B1 C3
D. A2 B3 C1
E. A3 B1 C2
F. A3 B2 C1
G. I don’t know

1. \( Y_i = ae^{\alpha t_i} + \epsilon_i, \quad \epsilon_i \ iid \sim N(0, \sigma^2) \)
2. \( Y_i = ae^{\alpha t_i}(1 + \epsilon_i), \quad \epsilon_i \ iid \sim N(0, \sigma^2) \)
3. \( \log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \ iid \sim N(0, \sigma^2) \)

\[
A: \sum_{i=1}^{I} (Y_i - ae^{\alpha t_i})^2 \quad B: \sum_{i=1}^{I} \left( \frac{Y_i - ae^{\alpha t_i}}{ae^{\alpha t_i}} \right)^2 \quad C: \sum_{i=1}^{I} (\log Y_i - \log a - \alpha t_i)^2
\]
Solution

1. \( Y_i = ae^{\alpha t_i} + \epsilon_i, \quad \epsilon_i \ iid \sim N(0, \sigma^2) \)
2. \( Y_i = ae^{\alpha t_i}(1 + \epsilon_i), \quad \epsilon_i \ iid \sim N(0, \sigma^2) \)
3. \( \log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \ iid \sim N(0, \sigma^2) \)

1. Corresponds to LS -> A
   \[ A: \sum_{i=1}^{l} (Y_i - ae^{\alpha t_i})^2 \]
   \[ B: \sum_{i=1}^{l} \left( \frac{Y_i - ae^{\alpha t_i}}{ae^{\alpha t_i}} \right)^2 \]

2. Can be re-written as
   \[ Y_i = ae^{\alpha t_i} + \frac{\epsilon_i}{w_i} \text{ with} \]
   \[ w_i = \frac{1}{ae^{\alpha t_i}} \rightarrow B \]

3. Corresponds to WLS in log scale -> C

Answer A
Compare models 2 and 3

1. \( Y_i = ae^{\alpha t_i} + \epsilon_i, \quad \epsilon_i \text{ iid } \sim N(0, \sigma^2) \)
2. \( Y_i = ae^{\alpha t_i}(1 + \epsilon_i), \quad \epsilon_i \text{ iid } \sim N(0, \sigma^2) \)
3. \( \log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \text{ iid } \sim N(0, \sigma^2) \)

■ Assume model 2 and take log:
\[
\log Y_i = \log a + \alpha t_i + \log(1 + \epsilon_i) \\
\approx \log a + \alpha t_i + \epsilon_i
\]

■ So we expect models 2 and 3 to be approximately equivalent.
Take-Home Message

- In order to fit a model you need to find an appropriate score function.
- One easy way to determine one is to consider an explicit model of the noise, and verify it by looking at residuals.
- A common mistake is to assume that noise has same variance when it is obviously wrong (e.g. model 1).
- Further, we may be interested in finding confidence intervals for the estimated parameters.
2. Least Squares

- Very commonly used
- Efficient solution, lots of softwares
- And gives confidence intervals

\[ Y_i = f_i(\tilde{\beta}) + \epsilon_i \text{ for } i = 1, \ldots, I \text{ with } \epsilon_i \text{ iid } \sim N_0, \sigma^2 \]

**THEOREM 3.1 (Least Squares).** *For the model in Eq.(3.5),

1. the maximum likelihood estimator of the parameter \((\hat{\beta}, \sigma)\) is given by:

\[
(a) \quad \hat{\beta} = \arg \min_{\beta} \sum_i \left( y_i - f_i(\beta) \right)^2 \\
(b) \quad \hat{\sigma}^2 = \frac{1}{I} \sum_i \left( y_i - f_i(\hat{\beta}) \right)^2
\]
2. Let $K$ be the square matrix of second derivatives (assumed to exist), defined by

$$K_{j,k} = \frac{1}{\sigma^2} \sum_i \frac{\partial f_i}{\partial \beta_j} \frac{\partial f_i}{\partial \beta_k}$$

If $K$ is invertible and if the number $I$ of data points is large, $\hat{\beta} - \bar{\beta}$ is approximately gaussian with 0 mean and covariance matrix $K^{-1}$.

Alternatively, for large $I$, an approximate confidence set at level $\gamma$ for the $j$th component $\beta_j$ of $\bar{\beta}$ is implicitly defined by

$$-2I \ln (\hat{\sigma}) + 2I \ln \left(\hat{\sigma}(\hat{\beta}_1, \ldots, \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1} \ldots \hat{\beta}_p)\right) \geq \xi_1$$

where $\hat{\sigma}^2(\bar{\beta}) = \frac{1}{I} \sum_i \left(y_i - f_i(\bar{\beta})\right)^2$ and $\xi_1$ is the $\gamma$ quantile of the $\chi^2$ distribution with 1 degree of freedom (for example, for $\gamma = 0.95$, $\xi_1 = 3.92$).
Example: model 3  \[ \log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \ iid \sim N(0, \sigma^2) \]

\( f_i = a' + \alpha t_i \) with \( \log a = a' \); let \( \beta_1 = \alpha \) and \( \beta_2 = a' \)

\( \frac{\partial f_i}{\partial \beta_1} = t_i, \quad \frac{\partial f_i}{\partial \beta_2} = 1 \)

\[ K_{1,1} = \frac{1}{\sigma^2} \sum_t t_i^2, K_{1,2} = K_{2,1} = \frac{1}{\sigma^2} \sum_t t_i, K_{2,2} = \frac{I}{\sigma^2} \]

\( f_i = a' + \alpha t_i \) with \( \log a = a' \); let \( \beta_1 = \alpha \) and \( \beta_2 = a' \)

\( \frac{\partial f_i}{\partial \beta_1} = t_i, \quad \frac{\partial f_i}{\partial \beta_2} = 1 \)

\[ K_{1,1} = \frac{1}{\sigma^2} \sum_{i} t_i^2, K_{1,2} = K_{2,1} = \frac{1}{\sigma^2} \sum_{i} t_i, K_{2,2} = \frac{I}{\sigma^2} \]

\[ K = \frac{1}{\sigma^2} \begin{bmatrix} \sum_{i} t_i^2 & \sum_{i} t_i \\ \sum_{i} t_i & I \end{bmatrix} \]
Example: model 3 \[ \log Y_i = \log \alpha + \alpha t_i + \epsilon_i, \quad \epsilon_i \ iid \sim N(0, \sigma^2) \]

2. Let \( K \) be the square matrix of second derivatives (assumed to exist), defined by

\[
K_{j,k} = \frac{1}{\sigma^2} \sum_i \frac{\partial f_i}{\partial \beta_j} \frac{\partial f_i}{\partial \beta_k}
\]

If \( K \) is invertible and if the number \( I \) of data points is large, \( \hat{\beta} - \bar{\beta} \) is approximately gaussian with 0 mean and covariance matrix \( K^{-1} \).

\[
K = \frac{1}{\sigma^2} \begin{bmatrix}
\sum_i t_i^2 & \sum_i t_i \\
\sum_i t_i & I
\end{bmatrix} = \frac{I}{\sigma^2} \begin{bmatrix}
s_t^2 + \bar{t}^2 & \bar{t} \\
\bar{t} & 1
\end{bmatrix}
\]

Matlab calls \( K \) the variance-covariance matrix

\[
K^{-1} = \frac{\sigma^2}{I s_t^2} \begin{bmatrix}
1 & -\bar{t} \\
-\bar{t} & s_t^2 + \bar{t}^2
\end{bmatrix}
\]

\[
\Rightarrow \alpha - \hat{\alpha} \sim N \left( 0, \frac{\sigma^2}{I s_t^2} \right), \quad a - \hat{a} \sim N \left( 0, \frac{\sigma^2}{I s_t^2} (\bar{t}^2 + s_t^2) \right)
\]
An approximate 95% confidence interval for $\alpha$ is

$$(\text{with } \hat{\sigma}^2 = \frac{1}{l} \sum_i (\log Y_i - \log \hat{\alpha} - \hat{\alpha} t_i)^2)$$

A. $\alpha \pm 1.96 \frac{\hat{\sigma}}{s_t}$

B. $\hat{\alpha} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{I_{st}}}$

C. $\hat{\alpha} \pm 1.96 \frac{\hat{\sigma}}{I_{st}}$

D. None of the above

E. I don’t know
Solution

The theorem says that, approximately, \( \alpha - \hat{\alpha} \sim N \left( 0, \frac{\sigma^2}{Is_t^2} \right) \)

Thus, with proba 95%:
\[ |\alpha - \hat{\alpha}| \leq 1.96 \frac{\sigma}{\sqrt{Is_t}} \]

We can approximate \( \sigma \) by \( \hat{\sigma} \)

Answer B

We find CI for \( \alpha = [0.3762 \quad 0.4066] \)

Note that the confidence interval decreases with \( \sqrt{I} \), i.e. if we have more points but the same time interval, the accuracy increases.

Also, if we keep the number of points constant but keep all time points closes, \( s_t^2 \) decreases and the CI increases: the slope of the fitted line is less accurate.
Linear Regression

- Model 3 is a special case of linear regression
- By definition, a linear regression model (with least squares) means a model of the form
  \[ Y_i = f_i(\beta) + \epsilon_i, \quad \epsilon_i \sim iid \ N(0, \sigma^2) \]
  where \( f_i \) is linear with respect to \( \beta \)
- For such models, there are closed form solutions in matrix form for the maximum likelihood estimation solution, including exact confidence intervals – see thm 3.3
THEOREM 3.3 (Linear Regression). Consider the model in Definition 3.1; let \( \vec{y} \) be the \( I \times 1 \) column vector of the data.

1. The \( p \times p \) matrix \( (X^T X) \) is invertible.

2. (Estimation) The maximum likelihood estimator of \( \vec{\beta} \) is \( \hat{\beta} = K \vec{y} \) with \( K = (X^T X)^{-1} X^T \).

3. (Standardized Residuals) Define the \( i \)th residual as \( e_i = \left( \vec{y} - X \hat{\beta} \right)_i \). The residuals are zero-mean gaussian but are correlated, with covariance matrix \( \sigma^2 (I_{d_I} - H) \), where \( H = X (X^T X)^{-1} X^T \).

Let \( s^2 = \frac{1}{I - p} ||e||^2 = \frac{1}{I - p} \sum_i e_i^2 \) (rescaled sum of squared residuals). \( s^2 \) is an unbiased estimator of \( \sigma^2 \).

The standardized residuals defined by \( r_i := \frac{e_i}{s \sqrt{1 - H_{i,i}}} \) have unit variance and \( r_i \sim t_{I - p - 1} \).

This can be used to test the model by checking that \( r_i \) are approximately normal with unit variance.

4. (Confidence Intervals) Let \( G = (X^T X)^{-1} = KK^T \); the distribution of \( \hat{\beta} \) is gaussian with mean \( \vec{\beta} \) and covariance matrix \( \sigma^2 G \), and \( \hat{\beta} \) is independent of \( e \).

In particular, assume we want a confidence interval for a (non-random) linear combination of the parameters \( \gamma = \sum_{j=1}^p u_j \beta_j \); \( \hat{\gamma} = \sum_j u_j \hat{\beta}_j \) is our estimator of \( \gamma \). Let \( g = \sum_{j,k} u_j G_{j,k} u_k = \sum_k \left( \sum_j u_j K_{j,k} \right)^2 \) (\( g \) is called the variance bias). Then \( \frac{\hat{\gamma} - \gamma}{\sqrt{gs}} \sim t_{I - p} \).

This can be used to obtain a confidence interval for \( \gamma \).
Example: Joe’s shop

\[ Y_i = (a + bx_i)1_{x_i \leq \xi} + (c + dx_i)1_{x_i > \xi} + \epsilon_i, \epsilon_i \sim \text{iid } N(0, \sigma^2) \]

with \( a + b\xi = c + d\xi \)
Is this a linear regression model?

\[ Y_i = (a + bx_i)1_{x_i \leq \xi} + (c + dx_i)1_{x_i > \xi} + \epsilon_i, \epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2) \]

with \( a + b\xi = c + d\xi \)

A. Yes
B. Yes if we assume \( \xi \) is known
C. No, even if \( \xi \) is not known
D. I don’t know
Assume $\xi$ is known. The function $f_i$ is given by

$$f_i = a + bx_i \text{ if } x_i \leq \xi$$
$$f_i = c + dx_i \text{ if } x_i > \xi$$

The points $x_i$ are known (and non random in this model)

The dependence on $(a, b, c, d)$ is linear and the noise is iid gaussian, therefore it is a linear regression model.

But if we assume that $\xi$ needs to be fitted as well, it is a non-linear model.

Answer B
Solution assuming $\xi = 70$

\begin{align*}
  a & \quad 0.978 \pm 0.609 \\
  b & \quad 0.0915 \pm 0.0137 \\
  c & \quad 15.8 \pm 2.99 \\
  d & \quad -0.121 \pm 0.037
\end{align*}
Validate the Assumptions with Residuals

We also computed the residuals $e_i$ (crosses) and standardized residuals $r_i$ (circles). There is little difference between both types of residuals. They appear reasonably normal, but one might criticize the model in that the variance appears smaller for smaller values of $x$. The normal qqplot of the residuals also shows approximate normality (the qqplot of standardized residuals is similar and is not shown).
2. Let $K$ be the square matrix of second derivatives (assumed to exist), defined by

$$K_{j,k} = \frac{1}{\sigma^2} \sum_i \frac{\partial f_i}{\partial \beta_j} \frac{\partial f_i}{\partial \beta_k}$$

If $K$ is invertible and if the number $I$ of data points is large, $\hat{\beta} - \bar{\beta}$ is approximately gaussian with 0 mean and covariance matrix $K^{-1}$.

Alternatively, for large $I$, an approximate confidence set at level $\gamma$ for the $j$th component $\beta_j$ of $\bar{\beta}$ is implicitly defined by

$$-2I \ln (\hat{\sigma}) + 2I \ln \left( \hat{\sigma}(\hat{\beta}_1, \ldots, \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1} \ldots \hat{\beta}_p) \right) \geq \xi_1$$

where $\hat{\sigma}^2(\bar{\beta}) = \frac{1}{I} \sum_i \left( y_i - f_i(\bar{\beta}) \right)^2$ and $\xi_1$ is the $\gamma$ quantile of the $\chi^2$ distribution with 1 degree of freedom (for example, for $\gamma = 0.95$, $\xi_1 = 3.92$).
Example 3.8: Joe’s Shop - Beyond the Linear Case - Estimation of $\xi$. In Example 3.6 we assumed that the value $\xi$ after which there is congestion collapse is known in advance. Now we relax this assumption. Our model is now the same as Eq.(3.9), except that $\xi$ is also now a parameter to be estimated.

To do this, we apply maximum likelihood estimation. We have to maximize the log-likelihood $l_\hat{y}(a, b, d, \xi, \sigma)$, where $\hat{y}$, the data, is fixed. For a fixed $\xi$, we know the value of $(a, b, d, \sigma)$ that achieves the maximum, as we have a linear regression model. We plot the value of this maximum versus $\xi$ (Figure 3.2) and numerically find the maximum. It is for $\xi = 77$.

To find a confidence interval, we use the asymptotic result in Theorem B.3.1. It says that a 95% confidence interval is obtained by solving $l(\hat{\xi}) - l(\xi) \leq 1.9207$, which gives $\xi \in [73, 80]$.
3. $\ell^1$- Norm Minimization

Laplace Noise

Recall that the pdf of the exp($\lambda$) distribution is $f(x) = \lambda e^{-\lambda x} 1_{x \geq 0}$

Consider the following sampling method:

- Draw $Y \sim \text{exp}(\lambda)$
- With probability 0.5, let $X = Y$, with probability 0.5 let $X = -Y$

The output $X$ is called Laplace Noise

What is its pdf?
The PDF of Laplace noise is ...

A. \( f(x) = \frac{\lambda}{2} e^{\lambda x} + \frac{\lambda}{2} e^{-\lambda x} \)

B. \( f(x) = \frac{\lambda}{2} e^{-\lambda |x|} \)

C. \( f(x) = \frac{\lambda}{2} e^{-\lambda |x|} + \frac{\lambda}{2} e^{-\lambda x} \)

D. \( f(x) = \frac{\lambda}{2} e^{-\lambda |x|} + \frac{\lambda}{2} e^{\lambda |x|} \)

E. I don’t know
Solution

Method 1 (heuristic) Let $Q = 1$ when we choose the first alternative (i.e. we do $X = Y$) and $Q = 0$ otherwise.

Let $x > 0$: the event $X \in [x, x + dx]$ is equal to $(Q = 1)$ and $(Y \in [x, x + dx])$. Its probability is $\frac{1}{2} \times \lambda e^{-\lambda x} \, dx + o(dx)$. Thus $f_X(x) = \frac{1}{2} \times \lambda e^{-\lambda x}$ when $x > 0$

Let $x < 0$: the event $X \in [x, x + dx]$ is equal to $(Q = 0)$ and $(Y \in [-x - dx, -x])$. Its probability is $\frac{1}{2} \times \lambda e^{-\lambda |x|} \, dx + o(dx)$. Thus $f_X(x) = \frac{1}{2} \times \lambda e^{-\lambda |x|}$ when $x < 0$

Answer B
Method 2 (rigorous)

Solution

For an arbitrary test function $\phi$, let us compute $E(\phi(X))$

If we can write it under the form $E(\phi(X)) = \int_{-\infty}^{+\infty} \phi(x)f_X(x)dx$

this will uniquely define $f_X(x)$

$$E(\phi(X)|Q = 1) = E(\phi(Y)) = \int_{0}^{+\infty} \phi(x)\lambda e^{-\lambda x} dx$$

$$= \int_{0}^{+\infty} \phi(x)\lambda e^{-\lambda |x|} dx$$

$$E(\phi(X)|Q = 0) = E(\phi(-Y))) = \int_{0}^{+\infty} \phi(-x)\lambda e^{-\lambda x} dx$$

$$= \int_{-\infty}^{0} \phi(x)\lambda e^{\lambda x} dx = \int_{-\infty}^{0} \phi(x)\lambda e^{-\lambda |x|} dx$$
Method 2 (rigorous)  

Solution

For any test function $\varphi$, let us compute $E(\varphi(X))$

\[
E(\varphi(X)) = \frac{1}{2} E(\varphi(X)|Q = 1) + \frac{1}{2} E(\varphi(X)|Q = 0)
\]

\[
= \frac{1}{2} \int_0^{+\infty} \varphi(x) \lambda e^{-\lambda |x|} \, dx + \frac{1}{2} \int_0^0 \varphi(x) \lambda e^{-\lambda |x|} \, dx
\]

\[
= \int_{-\infty}^{+\infty} \varphi(x) \left( \frac{1}{2} e^{-\lambda |x|} \right) \, dx = \int_{-\infty}^{+\infty} \varphi(x) f_X(x) \, dx
\]

thus $f_X(x) = \frac{1}{2} \lambda e^{-\lambda |x|}$
Laplace Noise

- The distribution with pdf $f_X(x) = \frac{1}{2} \lambda e^{-\lambda |x|}$ is called the Laplace distribution, or 2-sided exponential distribution.
- Its mean is 0 (by symmetry of the pdf) and its variance is $\frac{2}{\lambda^2}$.
- It can be used to model noise that has a heavier tail than gaussian noise; very large values are not too unlikely.

![Graph comparing PDF and 1-CDF of Normal(0,1) and Laplace(1.414) distributions](image)

$P(X > x)$
\[ Y_i = f_i(\beta) + \epsilon_i \text{ with } \epsilon_i \text{ iid } \sim \text{Laplace}(\lambda) \]

**THEOREM 3.2 (Least Deviation).** For the model in Eq. (3.7), the maximum likelihood estimator of the parameter \((\beta, \lambda)\) is given by:

1. \[ \hat{\beta} = \arg \min_\beta \sum_i \left| y_i - f_i(\hat{\beta}) \right| \]

2. \[ \frac{1}{\hat{\lambda}} = \frac{1}{I} \sum_i \left| y_i - f_i(\hat{\beta}) \right| \]

- i.e. the signal processing method with \(\ell^1\) score is the same as the statistical method with iid Laplace noise.

The \(\ell^1\) norm of a sequence \(z = (z_1, \ldots, z_n)\) is \(\|z\|_1 = \sum_{i=1}^{n} |z_i|\)
Given a model to be fitted

\[ Y_i = f_i(\theta) + \text{noise} \]

we have the equivalences:

<table>
<thead>
<tr>
<th>Score</th>
<th>Statistical model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least Square:</td>
<td></td>
</tr>
<tr>
<td>( \text{score} = \sum_i (\text{noise}_i)^2 )</td>
<td>( \text{noise \sim iid } N(0, \sigma^2) )</td>
</tr>
<tr>
<td>Weighted Least Square:</td>
<td></td>
</tr>
<tr>
<td>( \text{score} = \sum_i (w_i \times \text{noise}_i)^2 )</td>
<td>( \text{noise}_i \sim \text{ind } N \left(0, \frac{\sigma^2}{w_i^2}\right) )</td>
</tr>
<tr>
<td>( \ell^1 \text{ score} = \sum_i</td>
<td>\text{noise}_i</td>
</tr>
<tr>
<td>( \ell^1 \text{ score} = \sum_i w_i</td>
<td>\text{noise}_i</td>
</tr>
</tbody>
</table>
Virus propagation example: both fits give the same result in log-scale
$\ell^1$-norm minimization is more robust to « Outliers »

Figure 3.1: Fitting an exponential growth model to the data in Example 3.1, showing the fits obtained with least square (plain) and with $\ell^1$ norm minimization (dashed). First panel: original data; both fits are the same; Second panel: data corrupted by one outlier; the fit with $\ell^1$ norm minimization is not affected, whereas the least square fit is.

<table>
<thead>
<tr>
<th></th>
<th>Least Square</th>
<th>$\ell^1$ norm minimization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rate</td>
<td>prediction</td>
</tr>
<tr>
<td>no outlier</td>
<td>0.3914</td>
<td>30300</td>
</tr>
<tr>
<td>with one outlier</td>
<td>0.3325</td>
<td>14500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>rate</td>
<td>prediction</td>
</tr>
<tr>
<td>no outlier</td>
<td>0.3938</td>
<td>32300</td>
</tr>
<tr>
<td>with one outlier</td>
<td>0.3868</td>
<td>30500</td>
</tr>
</tbody>
</table>
To understand the difference between LS and $\ell^1$ consider the simple example $y_i = m + \text{noise}$ i.e. fit a cloud of data to a single value

**Least Square**
- Model:
  \[ y_i = m + \varepsilon_i, \varepsilon_i \text{ iid } \sim N(0, \sigma^2) \]
- What is $m$?
- Confidence interval?

**L1 Norm Minimization**
- Model:
  \[ y_i = m + \varepsilon_i, \varepsilon_i \text{ iid } \sim \text{Laplace}(\lambda) \]
- What is $m$?
- Confidence interval?
What is the LS fit of $m$ in $y_i = m + \text{noise}$?

A. $m = \text{median of } y_i$

B. $m = \bar{y}$ where $\bar{y} = \text{mean of } y_i$

C. $m = \bar{y} + \frac{\hat{\sigma}^2}{y}$ where $\hat{\sigma}^2 = \frac{1}{l} \sum_i (y_i - \bar{y})^2$

D. $m = \bar{y} + \frac{\hat{\sigma}^2}{y}$ where $\hat{\sigma}^2 = \frac{1}{l-1} \sum_i (y_i - \bar{y})^2$

E. I don’t know
Solution

For LS fit, $m$ minimizes $\sum_{i=1}^{l} (y_i - m)^2$; now

$$\sum_{i} (y_i - m)^2 = \sum_{i} (y_i - \bar{y})^2 + I(m - \bar{y})^2$$

Thus the minimum is for $m = \bar{y}$.

Answer B
What is the $\ell^1$ fit of $m$ in $y_i = m + \text{noise}$?

A. $m = y$ where $y = \text{mean of } y_i$

B. $m = \text{median of } y_i$

C. $m = \bar{y} + \frac{\hat{\sigma}^2}{y}$ where $\hat{\sigma}^2 = \frac{1}{l} \sum_{i=1}^{l} (y_i - \bar{y})^2$

D. $m = \bar{y} + \frac{\hat{\sigma}^2}{y}$ where $\hat{\sigma}^2 = \frac{1}{l-1} \sum_{i=1}^{l} (y_i - \bar{y})^2$

E. I don’t know
For $\ell^1$ fit, $m$ minimizes
$$\sum_{i=1}^{I} |y_i - m|;$$
Consider the mapping
$$m \mapsto f(m) = \sum_{i=1}^{I} |y_i - m|$$
e.g. when $y_{(1)} < m < y_{(2)}$
$$f(m) = (m - y_{(1)}) + (y_{(2)} - m) + (y_{(3)} - m) + (y_{(4)} - m) + (y_{(5)} - m)$$
and $f'(m) = 1 - 4 = -3$
In general when $m$ is not one of the $y_i$
$$f'(m) = \text{nb points below } m - \text{nb points above } m; \text{ when } I \text{ is odd:}$$
$f$ decreases for $m < \text{median and } f$ increases for $m > \text{median}$
Answer B
Least Square

- Model: \( y_i = m + \epsilon_i, \epsilon_i \text{ iid } \sim N(0, \sigma^2) \)
- \( m \) is the mean
- \( m = \bar{y} \pm \eta \frac{\hat{\sigma}}{\sqrt{I}} \)

L1 Norm Minimization

- Model: \( y_i = m + \epsilon_i, \epsilon_i \text{ iid} \)
- \( m \) is the median
- Confidence interval given by order statistics

Take Home message for LS vs \( \ell^1 \) scores in general

\( \ell^1 \) is more robust than LS, less sensitive to outliers
Confidence intervals are simple with LS – but meaningful only if normal assumption holds
Linear Regression with $\ell^1$ norm minimization

= $\ell^1$ norm minimization + linear dependency on parameter

More robust than LS linear regression

Less traditional

**Definition 3.2 (Linear Regression Model with Laplace Noise).**

$$Y_i = (X\tilde{\beta})_i + \epsilon_i \text{ for } i = 1, \ldots, I \text{ with } \epsilon_i \text{ iid } \sim \text{Laplace}(\lambda)$$
THEOREM 3.3.1. Consider the model in Definition 3.2.1; let $\tilde{y}$ be the $I \times 1$ column vector of the data. The maximum likelihood estimator of $\tilde{\beta}$ is obtained by solving the linear program:

$$\text{minimize} \quad \sum_{i=1}^{I} u_i$$

over \quad $\tilde{\beta} \in \mathbb{R}^p$, $u \in \mathbb{R}^I$

subject to the constraints

$$u_i \geq y_i - \left( X \tilde{\beta} \right)_i$$

$$u_i \geq -y_i + \left( X \tilde{\beta} \right)_i$$

The maximum likelihood of the noise parameter $\lambda$ is

$$\left( \frac{1}{I} \sum_{i=1}^{I} \left| y_i - \left( X \tilde{\beta} \right)_i \right| \right)^{-1}.$$
Figure 3.3: Modelling congestion collapse in Joe’s shop with a piecewise linear function and $\ell^1$ norm minimization of the errors.
Confidence Intervals

- No closed form
- We can use an approximate solution based on simulation ("Bootstrap")

<table>
<thead>
<tr>
<th></th>
<th>Achieved Throughput (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.32 ± 0.675</td>
</tr>
<tr>
<td>b</td>
<td>0.0791 ± 0.0149</td>
</tr>
<tr>
<td>c</td>
<td>11.7 ± 3.24</td>
</tr>
<tr>
<td>d</td>
<td>-0.0685 ± 0.0398</td>
</tr>
</tbody>
</table>
4. Bootstrap : Computation of CI

- We have a parametric model with parameters \( \beta, \lambda \) and want to find a confidence interval \([U, V]\) for, say, \( \beta_1 \); let \( \hat{\beta}_1 \) be our estimator of \( \beta_1 \).

- Recall that \( U, V \) is a function of the data (hence is random); we want to have \( P_{\beta, \lambda}(U \leq \beta_1 \leq V) \geq 0.95 \) for any \( \beta, \lambda \).

- Assume someone tells us the true value of \( \beta, \lambda \); in theory, we can compute the quantiles of the distribution of \( \hat{\beta}_1 \); let

  \[ U_{\beta, \lambda} = 2.5\% \text{ quantile of } \hat{\beta}_1 \text{ and } V_{\beta, \lambda} = 97.5\% \text{ quantile of } \hat{\beta}_1 \]

  we have

  \[ P_{\beta, \lambda}(U_{\beta, \lambda} \leq \hat{\beta}_1 \leq V_{\beta, \lambda}) = 0.95 \]

- We have computed an estimator \( \hat{\beta}, \hat{\lambda} \); we take as approximate confidence interval \( U = U_{\hat{\beta}, \hat{\lambda}} \) and \( V = V_{\hat{\beta}, \hat{\lambda}} \).
Parametric Bootstrap Computation of CI (cont’d)

- Let $U_{\beta, \lambda} = 2.5\%$ quantile of $\hat{\beta}_1$ and $V_{\beta, \lambda} = 97.5\%$ quantile of $\hat{\beta}_1$

- We take as approximate confidence interval $U = U_{\hat{\beta}, \hat{\lambda}}$ and $V = V_{\hat{\beta}, \hat{\lambda}}$

- How can we practically compute $U_{\beta, \lambda}, V_{\beta, \lambda}$?
  By Monte-Carlo simulation:

  - Assume you know $\beta, \lambda$
    - Do $r = 1: R$
      - simulate $Y_i, i = 1 \ldots I$ from the model
      - obtain an estimate $\hat{\beta}_i^r$
    end

    - $U_{\beta, \lambda} = 2.5\%$ quantile of $\hat{\beta}_i^r$, $V_{\beta, \lambda} = 97.5\%$ quantile of $\hat{\beta}_i^r$
      e.g. with $R = 999$, $U_{\beta, \lambda} = \hat{\beta}_i^{(25)}$ and $V_{\beta, \lambda} = \hat{\beta}_i^{(975)}$

- Do the above with $(\beta, \lambda) = (\hat{\beta}, \hat{\lambda})$ – this is called **Parametric Bootstrap**
Example

- We have obtained $\hat{\lambda} = 1.0055$
- For $r = 1:999$ we draw $I$ residuals $\sim$ Laplace(1.0055) and simulate $n$ artificial data points $y_i^r$, $i = 1 \ldots I$ from the fitted model
- For each replay experiment we estimate $a, b, c, d$ and obtain $R = 999$ estimates $\hat{a}^r, \hat{b}^r, \hat{c}^r, \hat{d}^r$
- A confidence interval for $d$ is $[\hat{d}^{(25)}, \hat{d}^{(975)}]$

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.49 ± 0.601</td>
</tr>
<tr>
<td>b</td>
<td>0.079 ± 0.0143</td>
</tr>
<tr>
<td>c</td>
<td>11.2 ± 2.96</td>
</tr>
<tr>
<td>d</td>
<td>-0.062 ± 0.0368</td>
</tr>
</tbody>
</table>
Bootstrap with Resample From Residuals

- Instead of simulating the noise from the model (by drawing from \( \text{Laplace}(\lambda) \)) assume we draw \( I \) noise samples from the residuals \textit{with replacement}.
- This method is called \textit{Bootstrap with Resampling from Residuals}.
Algorithm 2 The Bootstrap with Re-Sampling From Residuals. The goal is to compute a confidence interval for some function $\varphi(\hat{\beta})$ of the parameter of the model in Definition 3.2.1. $r_0$ is the algorithm’s accuracy parameter.

1: $R = \left\lceil \frac{2r_0}{1 - \gamma} \right\rceil - 1$  \hspace{1cm} $\triangleright$ For example $r_0 = 25$, $\gamma = 0.95$, $R = 999$
2: estimate $\hat{\beta}$ using Theorem 3.3.1; obtain $\hat{\beta}$
3: compute the residuals $e_i = y_i - \left( X \hat{\beta} \right)_i$
4: $\textbf{for } r = 1 : R \textbf{ do}$ \hspace{1cm} $\triangleright$ Re-sample from residuals
5: \hspace{1cm} draw $I$ numbers with replacement from the list $(e_1, \ldots, e_I)$ and call them $E^r_1, \ldots, E^r_I$
6: \hspace{1cm} generate the bootstrap replicate $Y^r_1, \ldots, Y^r_I$ from the estimated model:
7: \hspace{1cm} $Y^r_i = \left( X \hat{\beta} \right)_i + E^r_i$ for $i = 1 \ldots I$
8: \hspace{1cm} re-estimate $\tilde{\beta}$, using $Y^r_i$ as data, using Theorem 3.3.1; obtain $\tilde{\beta}^r$
9: $\textbf{end for}$
10: $(\varphi(1), \ldots, \varphi(R)) = \text{sort} \left( \varphi(\tilde{\beta}^1), \ldots, \varphi(\tilde{\beta}^R) \right)$
11: confidence interval for $\varphi(\hat{\beta})$ is $[\varphi(r_0); \varphi(R+1-r_0)]$
\[ \xi = 69 \]

### CI with Parametric Bootstrap, 999 replays

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.49 ± 0.601</td>
</tr>
<tr>
<td>b</td>
<td>0.079 ± 0.0143</td>
</tr>
<tr>
<td>c</td>
<td>11.2 ± 2.96</td>
</tr>
<tr>
<td>d</td>
<td>−0.062 ± 0.0368</td>
</tr>
</tbody>
</table>

### CI with Bootstrap, Resampling from Residuals, 999 replays

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1.32 ± 0.675</td>
</tr>
<tr>
<td>b</td>
<td>0.0791 ± 0.0149</td>
</tr>
<tr>
<td>c</td>
<td>11.7 ± 3.24</td>
</tr>
<tr>
<td>d</td>
<td>−0.0685 ± 0.0398</td>
</tr>
</tbody>
</table>
Other Uses of the Bootstrap

- The idea of the bootstrap is to use the data itself as estimate of the unknown distribution of the data.

- Can be used to obtain confidence or predictions intervals in a very simple way --- tends to underestimate.

- Example: CI for the mean of a data set $y_i$

  The model is $Y_i \sim \text{iid } F()$ and the problem is to estimate the mean of $F()$

  With the bootstrap we replace the unknown $F()$ by the empirical distribution of the data itself, i.e. the distribution that puts probability $\frac{1}{I}$ at every $y_i$, $i = 1 \ldots I$. 

Which algorithm is a correct implementation of the bootstrap for a 95% CI of the mean of $y_1, \ldots, y_I$?

A. A  
B. B  
C. Both  
D. None  
E. I don’t know

Algorithm A
for $r = 1: 999$
  for $n = 1: I$
    draw one integer $K \sim \text{unif}\{1, \ldots, I\}$
    $x_{n,r} = y_K$
  end
end
for $r = 1: 999$
  $m_r = \text{mean}(x_{1,r}, \ldots, x_{I,r})$
end
$\text{CI} = [m_{(25)}, m_{(975)}]$ 

Algorithm B
for $r = 1: 999$
  draw a random permutation $\sigma$ of $\{1, \ldots, I\}$
  $x_{n,r} = y_{\sigma(n)}$, $n = 1 \ldots I$
end
for $r = 1: 999$
  $m_r = \text{mean}(x_{1,r}, \ldots, x_{I,r})$
end
$\text{CI} = [m_{(25)}, m_{(975)}]$
Solution

Answer A
The bootstrap replaces the model \( Y_i \sim \text{iid } F() \) by \( Y_i \sim \text{iid } \hat{F}() \)
where \( \hat{F}() \) is the empirical distribution of the \( I \) data points \( y_i \).
One replay experiment simulates the data from the model \( Y_i \sim \hat{F}() \), i.e. draws \( I \) independent samples from \( \hat{F}(\ ) \)

Q: How does one draw one sample \( Y \) from \( \hat{F}(\ ) \)?
A: Draw an index \( K \in \{1, ..., I\} \) uniformly at random and let \( Y = y_K \)

Algorithm A does this, algorithm B does not (not iid). Note that we need to resample with replacement.
Example: Compiler Options, CI for mean

- Assuming data is normal (Thm 2.3)
- Assuming $n$ is large and variance is finite (Thm 2.2)

- Bootstrap