Inference in non-linear time series

Erik Lindström

Centre for Mathematical Sciences
Lund University

LU/LTH & DTU
Overview

Introduction
  General
  Properties
  Estimators

Least squares

Maximum Likelihood
  MLE asymptotics
  Two theorems
  Fisher Information
  Proofs

Other estimators
We are interested in estimating a parameter vector $\theta_0$ from data $X$.

- Ad hoc or formal estimation?
- Properties
- Definition

$$\hat{\theta} = T_N(X).$$

- Interpretation
Properties

- **Bias** \( b = \theta_0 - \mathbb{E}[T_N(X)] \).
- **Asympt. bias** \( \lim_{N \to \infty} \theta_0 - \mathbb{E}[T_N(X)] \).
- **Consistency** \( \hat{\theta} \overset{p}{\to} \theta_0 \iff P(|\hat{\theta} - \theta_0| > \varepsilon) \to 0 \text{ as } N \to \infty \).
- **Strong consistency** \( \hat{\theta} \overset{a.s.}{\to} \theta_0 \).
- **Efficiency**
  \[
  \text{Var}[T_N(X)] \geq l_N^{-1}(\theta_0),
  \]
  where \( l_N^{-1}(\theta_0) \) is the Fisher information matrix.
- **Asympt. normality**
- **Convergence**
  \[
  \mathcal{N}^\alpha(\hat{\theta} - \theta_0) \overset{d}{\to} \mathcal{N}(\mu, \Sigma).
  \]
Popular estimators

How are the estimates computed?

- **Optimization**: Least squares (LS), Weighted Least Squares (WLS), Prediction Error Methods (PEM), Generalized Method of Moments (GMM) etc.

- **Solving equations**: Estimation functions (EFs), Method of moments (MM), Instrumental variable (IV) methods

- **Either**: Bayesian methods (MCMC), Maximum Likelihood, Quasi Maximum Likelihood, Approx Bayesian Computation
Least squares

- Observations $y_1, \ldots, y_N$
- Predictors $\sum_{j=1}^{J} B_j(x)\theta_j$ (Splines; Hammerstein models)
- Vector form: $Y = Z\theta + \varepsilon, \varepsilon \sim N(0, \Omega)$

(Weighted) Parameter estimation:

$$\hat{\theta}_{LS} = \operatorname{argmin}_{\theta \in \Theta} \sum_{n=1}^{N} \lambda_n \left( y_n - \sum_j B_j(x_n)\theta_j \right)^2$$  \hspace{1cm} (1)

$$= (Y - Z\theta)^T W (Y - Z\theta)$$  \hspace{1cm} (2)

Generalization (often with a Bayesian interpretation)

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + P(\theta, \theta_0)$$  \hspace{1cm} (3)
Least squares

- Observations $y_1, \ldots, y_N$
- Predictors $\sum_{j=1}^{J} B_j(x)\theta_j$ (Splines; Hammerstein models)
- Vector form: $Y = Z\theta + \varepsilon, \varepsilon \sim N(0, \Omega)$

(Weighted) Parameter estimation:

$$\hat{\theta}_{LS} = \arg\min_{\theta \in \Theta} \sum_{n=1}^{N} \lambda_n \left( y_n - \sum_{j} B_j(x_n)\theta_j \right)^2$$  \hspace{1cm} (1)

$$= (Y - Z\theta)^T W (Y - Z\theta)$$  \hspace{1cm} (2)

Generalization (often with a Bayesian interpretation)

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + P(\theta, \theta_0)$$  \hspace{1cm} (3)
WLS estimate

\[ \hat{\theta} = \left( Z^T W Z \right)^{-1} (Z W Y) \]  

Bias? No, not if the correct model is used

\[ \mathbb{E} \left[ \hat{\theta} \right] = \left( Z^T W Z \right)^{-1} (Z W (Z \theta + \varepsilon)) \]

\[ = \theta + \left( Z^T W Z \right)^{-1} (Z W \varepsilon) \]

Slightly more complicated when the model is wrong...

Variance

\[ \text{Var} \left[ \hat{\theta} \right] = \left( Z^T W Z \right)^{-1} (Z^T W \Omega W Z) (Z^T W Z)^{-1} \]

Simplifies if \( W = \Omega^{-1} \), and further if \( \Omega = \sigma^2 I_N \). We then get

\[ \text{Var} \left[ \hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1} \]
WLS estimate

\[ \hat{\theta} = \left( Z^T WZ \right)^{-1} (ZWY) \]  

Bias? No, not if the correct model is used

\[ \mathbb{E} \left[ \hat{\theta} \right] = \left( Z^T WZ \right)^{-1} (ZW(Z\theta + \varepsilon)) \]
\[ = \theta + \left( Z^T WZ \right)^{-1} (ZW\varepsilon) \]

Slightly more complicated when the model is wrong...

Variance

\[ \text{Var} \left[ \hat{\theta} \right] = (Z^T WZ)^{-1} (Z^T W\Omega WZ) (Z^T WZ)^{-1} \]

Simplifies if \( W = \Omega^{-1} \), and further if \( \Omega = \sigma^2 I_N \). We then get

\[ \text{Var} \left[ \hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}. \]
\( \hat{\theta} = \left( Z^T WZ \right)^{-1} (ZWY) \) (4)

**Bias? No, not if the correct model is used**

\[
E \left[ \hat{\theta} \right] = \left( Z^T WZ \right)^{-1} (ZW(Z\theta + \varepsilon)) = \theta + \left( Z^T WZ \right)^{-1} (ZW\varepsilon)
\] (5) (6)

Slightly more complicated when the model is wrong...

**Variance**

\[
\text{Var} \left[ \hat{\theta} \right] = (Z^T WZ)^{-1}(Z^T W\Omega WZ)(Z^T WZ)^{-1}
\] (7)

Simplifies if \( W = \Omega^{-1} \), and further if \( \Omega = \sigma^2 I_N \). We then get

\[
\text{Var} \left[ \hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}.
\] (8)
WLS estimate

\[
\hat{\theta} = \left( Z^T W Z \right)^{-1} (Z W Y) \tag{4}
\]

Bias? No, not if the correct model is used

\[
E \left[ \hat{\theta} \right] = \left( Z^T W Z \right)^{-1} (Z W (Z \theta + \varepsilon)) \tag{5}
\]

\[
= \theta + \left( Z^T W Z \right)^{-1} (Z W \varepsilon) \tag{6}
\]

Slightly more complicated when the model is wrong...

Variance

\[
\text{Var} \left[ \hat{\theta} \right] = (Z^T W Z)^{-1} (Z^T W \Omega W Z) (Z^T W Z)^{-1} \tag{7}
\]

Simplifies if \( W = \Omega^{-1} \), and further if \( \Omega = \sigma^2 I_N \). We then get

\[
\text{Var} \left[ \hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}. \tag{8}
\]
WLS estimate

\[ \hat{\theta} = \left( Z^T WZ \right)^{-1} (Z W Y) \]  

Bias? No, not if the correct model is used

\[ E \left[ \hat{\theta} \right] = \left( Z^T W Z \right)^{-1} (Z W (Z \theta + \varepsilon)) \]

\[ = \theta + \left( Z^T W Z \right)^{-1} (Z W \varepsilon) \]  

Slightly more complicated when the model is wrong...

Variance

\[ \text{Var} \left[ \hat{\theta} \right] = \left( Z^T W Z \right)^{-1} (Z^T W \Omega W Z) (Z^T W Z)^{-1} \]

Simplifies if \( W = \Omega^{-1} \), and further if \( \Omega = \sigma^2 I_N \). We then get

\[ \text{Var} \left[ \hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}. \]
F-tests

- F-tests are used to test a single of a group of parameters by computing

\[ f = \frac{RSS_R - RSS_{UR}}{|\theta_{UR}| - |\theta_R|} \]

and comparing to an \( F(|\theta_{UR}| - |\theta_R|, N - |\theta_{UR}|) \) quantile.

- Why? We estimate \( \sigma^2 = \frac{(Y - Z\hat{\theta})^T(Y - Z\hat{\theta})}{N - J} \)

- Define

\[ Q(\hat{\theta}) = (Y - Z\hat{\theta})^T(Y - Z\hat{\theta}) = Y^T(I_N - P)^T(I_N - P)Y \]

where \( P = Z(Z^T Z)^{-1}Z^T \) is the projection matrix.
We approximate $Q$ by
$$\mathbb{E}[Q] = \text{tr} \left( (I_N - P)^T (I_N - P) \text{Cov}[Y, Y] \right).$$

It holds for the standard model that $\text{Cov}[Y, Y] = \sigma^2 I_N$ and $(I_N - P)^T (I_N - P) = (I_N - P)$

It then follows that
$$\text{tr} \left( (I_N - P)^T (I_N - P) \text{Cov}[Y, Y] \right) = \sigma^2 \text{tr} (I_N - P) \quad (10)$$
$$= \sigma^2 \text{tr}(I_N) - \sigma^2 \text{tr}(P) \quad (11)$$
$$= \sigma^2 N - \sigma^2 \text{tr} \left( Z(Z^T Z)^{-1} Z^T \right) \quad (12)$$
$$= \sigma^2 (N - \text{tr} \left( (Z^T Z)^{-1} (Z^T Z) \right) = \sigma^2 (N - J) \quad (13)$$
What about the penalized asymptotics?

What if the estimation is given by

\[
\hat{\theta}_{PLS} = \arg\min_{\theta \in \Theta} (Y - Z\theta)^TW(Y - Z\theta) + (\theta - \theta_0)^TD(\theta - \theta_0)
\] (14)

Derive on blackboard. [bias, degrees of freedom] Case study on Semi Parametric Lag Dependent functions

Another popular form is

\[
\hat{\theta}_{Adaptive \ LASSO} = \arg\min_{\theta \in \Theta} (Y - Z\theta)^TW(Y - Z\theta) + ||D(\theta - \theta_0)||_1
\] (15)

Easily computed using the ADMM algorithm. Case study in Matlab
What about the penalized asymptotics?

What if the estimation is given by

$$\hat{\theta}_{PLS} = \arg\min_{\theta \in \Theta} (Y - Z\theta)^T W(Y - Z\theta) + (\theta - \theta_0)^T D(\theta - \theta_0)$$  \hspace{1cm} (14)

Derive on blackboard. [bias, degrees of freedom] Case study on
Semi Parametric Lag Dependent functions

Another popular form is

$$\hat{\theta}_{Adaptive \ LASSO} = \arg\min_{\theta \in \Theta} (Y - Z\theta)^T W(Y - Z\theta) + \|D(\theta - \theta_0)\|_1$$  \hspace{1cm} (15)

Easily computed using the ADMM algorithm. Case study in Matlab
What about the penalized asymptotics?

What if the estimation is given by

$$\hat{\theta}_{PLS} = \arg\min_{\theta \in \Theta} (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D(\theta - \theta_0) \quad (14)$$

Derive on blackboard. [bias, degrees of freedom] Case study on Semi Parametric Lag Dependent functions

Another popular form is

$$\hat{\theta}_{Adaptive\ LASSO} = \arg\min_{\theta \in \Theta} (Y - Z\theta)^T W (Y - Z\theta) + ||D(\theta - \theta_0)||_1 \quad (15)$$

Easily computed using the ADMM algorithm. Case study in Matlab
Maximum Likelihood estimators

Defined as

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} p_\theta(X_1, \ldots X_N).$$

The MLE use all information in the data, and have nice properties:

- **la.** Consistency \( \hat{\theta} \xrightarrow{p} \theta_0 \).
- **lb.** Asympt. normality
- **Il.** Efficiency \( Var[T_N(X)] = I_0^{-1}(\theta_0) \),
- **Ill.** Invariant.

We have under general conditions that

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_0^{-1}(\theta_0)).$$
Two useful theorems

Assume $X_i$ iid and $\mathbb{E}[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2$

- Law of Large Numbers

$$\frac{1}{N} \sum_{i=1}^{N} X_i \xrightarrow{p/a.s.} \mu.$$ 

- Central Limit Theorem

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{X_i - \mu}{\sigma} \xrightarrow{d} Z,$$

where $Z \in \mathcal{N}(0,1)$.

These theorems can be generalized further.
Fisher information

The Fisher Information matrix is defined as

$$I(\theta) = \text{Var} \left[ \frac{\partial}{\partial \theta} \log p_\theta(X) \right],$$

which is equivalent to

$$E \left[ \left( \frac{\partial}{\partial \theta} \log p_\theta(X) \right)^2 \right]$$

and

$$-E \left[ \frac{\partial^2}{\partial \theta^2} \log p_\theta(X) \right].$$

Note $E \left[ \frac{\partial}{\partial \theta} \log p_\theta(X | \theta) \right] = 0.$
Proofs

- Ia. Kullback-Leibler and law of large number
- Ib. Second order Taylor expansion of the Likelihood.
- II. Cauchy-Schwartz inequality
- III. Direct calculations.
Consistency

The log-likelihood function is defined as

$$\ell(\theta) = \sum_{n=1}^{N} \log p_\theta(x_n|x_{1:n-1})$$  \hspace{1cm} (16)

The estimate is given by

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} \frac{1}{N} \ell(\theta)$$  \hspace{1cm} (17)

$$= \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n|x_{1:n-1})$$  \hspace{1cm} (18)

Now we rewrite this as

$$\frac{1}{N} \ell(\theta) = \frac{1}{N} \ell(\theta) \pm E[\ell(\theta)]$$  \hspace{1cm} (19)

$$= E[\ell(\theta)] + \left(\frac{1}{N} \ell(\theta) - E[\ell(\theta)]\right)$$  \hspace{1cm} (20)
Consistency

The log-likelihood function is defined as

$$\ell(\theta) = \sum_{n=1}^{N} \log p_\theta(x_n|\mathbf{x}_{1:n-1})$$

(16)

The estimate is given by

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} \frac{1}{N} \ell(\theta)$$

(17)

$$= \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n|\mathbf{x}_{1:n-1})$$

(18)

Now we rewrite this as

$$\frac{1}{N} \ell(\theta) = \frac{1}{N} \ell(\theta) \pm \mathbb{E}[\ell(\theta)]$$

(19)

$$= \mathbb{E}[\ell(\theta)] + \left( \frac{1}{N} \ell(\theta) - \mathbb{E}[\ell(\theta)] \right)$$

(20)
Consistency

The log-likelihood function is defined as

\[ \ell(\theta) = \sum_{n=1}^{N} \log p_\theta(x_n|x_{1:n-1}) \]  

(16)

The estimate is given by

\[ \hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} \frac{1}{N} \ell(\theta) \]  

(17)

\[ = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n|x_{1:n-1}) \]  

(18)

Now we rewrite this as

\[ \frac{1}{N} \ell(\theta) = \frac{1}{N} \ell(\theta) \pm \mathbb{E}[\ell(\theta)] \]  

(19)

\[ = \mathbb{E}[\ell(\theta)] + \left( \frac{1}{N} \ell(\theta) - \mathbb{E}[\ell(\theta)] \right) \]  

(20)
This decomposition shows that

\[
\frac{1}{N} \ell(\theta) = \underbrace{E[\ell(\theta)\}}_{\text{Expected log-likelihood}} + \frac{1}{N} \ell(\theta) - E[\ell(\theta)] \underbrace{\text{Random error}}_{\text{Random error}} (21)\]

It also holds trivially that

\[
\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} E[\ell(\theta)] + \left(\hat{\theta}_{MLE} - \arg\max_{\theta \in \Theta} E[\ell(\theta)]\right) (22)\]
First we look at the random error, that is written so that we can apply the Law of Large Numbers. If

$$\lim_{N \to \infty} \frac{1}{N} \ell(\theta) - \mathbb{E}[\ell(\theta)] = 0$$  \hspace{1cm} (23)

uniformly over $\Theta$, then it follows that

$$\lim_{N \to \infty} \hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \mathbb{E}[\ell(\theta)]$$  \hspace{1cm} (24)
Finally, we denote any feasible density by $q_\theta \in Q_\theta$ and assume that the true density $p_{\theta_0} \in Q_\theta$.

The difference in expected log-likelihood between the true density and any other density is given by

$$E[\log q_\theta(X)] - E[\log p_{\theta_0}(X)] = \int (\log q_\theta(x) - \log p_{\theta_0}(x)) p_{\theta_0}(x) dx$$

$$= \int \log \left( \frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx$$

Log is a concave function. We know from Jensen’s inequality that $E[\log(\xi)] \leq \log(E[\xi])$
Finally, we denote any feasible density by $q_\theta \in Q_\theta$ and assume that the true density $p_{\theta_0} \in Q_\theta$.

The difference in expected log-likelihood between the true density and any other density is given by

$$E[\log q_\theta(X)] - E[\log p_{\theta_0}(X)] = \int (\log q_\theta(x) - \log p_{\theta_0}(x))p_{\theta_0}(x)dx$$

(25)

$$= \int \log \left( \frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x)dx$$

(26)

Log is a concave function. We know from Jensen’s inequality that $E[\log(\xi)] \leq \log(E[\xi])$.
Finally, we denote any feasible density by $q_\theta \in Q_\theta$ and assume that the true density $p_{\theta_0} \in Q_\theta$.

The difference in expected log-likelihood between the true density and any other density is given by

$$E[\log q_\theta(X)] - E[\log p_{\theta_0}(X)] = \int (\log q_\theta(x) - \log p_{\theta_0}(x))p_{\theta_0}(x)dx$$

(25)

$$= \int \log \left( \frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x)dx$$

(26)

Log is a concave function. We know from Jensen’s inequality that $E[\log(\xi)] \leq \log(E[\xi])$.
That means that
\[
\int \log \left( \frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) \, dx \leq \log \left( \int \frac{q_\theta(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) \, dx \right) = 0 \quad (27)
\]

We only get equality when \( q_\theta(x) = p_{\theta_0}(x) \) for all values of \( x \).

The conclusion is that
\[
\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}[\ell(\theta)] \quad (28)
\]

and hence that
\[
\lim_{N \to \infty} \hat{\theta}_{MLE} = \theta_0 \quad (29)
\]
as the random error vanishes as \( N \to \infty \).
That means that

$$\int \log \left( \frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx \leq \log \left( \int \frac{q_\theta(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) dx \right) = 0 \quad (27)$$

We only get equality when $q_\theta(x) = p_{\theta_0}(x)$ for all values of $x$. The conclusion is that

$$\theta_0 = \arg\max_{\theta \in \Theta} E[\ell(\theta)] \quad (28)$$

and hence that

$$\lim_{N \to \infty} \hat{\theta}_{MLE} = \theta_0 \quad (29)$$

as the random error vanishes as $N \to \infty$. 
Proof Ib

Rewrite \( L(X) = \exp \left( \log p_\theta(X_1) + \sum_{n=2}^{N} \log p_\theta(X_n|X_1:n-1) \right) \).

Second order Taylor expansion around \( \theta_0 \). This is fine due to consistency.

Maximize

\[ \sqrt{N}(\hat{\theta} - \theta_0) \approx \ldots \]

Asymp. normality follows.
Proof I, Ext.

- Write $\ell_N(\theta) = \log L(X|\theta)$.
- This is a sum consisting of $N$ terms (think LLN and CLT!).
- Approximate $\ell_N(\theta)$ using a second order Taylor expansion around $\theta_0$.
- Thus

$$
\ell_N(\theta) = \ell_N(\theta_0) + \partial_\theta \ell_N(\theta_0)(\theta - \theta_0) + \frac{1}{2} \partial^2 \theta \ell_N(\theta_0)(\theta - \theta_0)^2 + R.
$$

- Ignore the last term, multiply $\ell_N$ with $1/\sqrt{N}$ and maximize wrt $\theta$ to obtain $\hat{\theta}$.
- We obtain $\frac{1}{\sqrt{N}} \partial_\theta \ell_N(\theta_0) + \frac{1}{N} \partial^2 \theta \ell_N(\theta_0) \sqrt{N}(\hat{\theta} - \theta_0) := 0$. 

Erik Lindström - erikl@maths.lth.se

Inference in non-linear time series
Proof I, Ext.

- We have from LLN that
  \[
  \frac{1}{N} \frac{\partial^2 \ell_N(\theta_0)}{\partial \theta^2} \to -I(\theta_0),
  \]

- and from CLT that
  \[
  \frac{1}{\sqrt{N}} \frac{\partial \ell_N(\theta_0)}{\partial \theta} \to Z,
  \]

where \( Z \in \mathcal{N}(0, I(\theta_0)) \).

- Rearranging gives \( \sqrt{N}(\hat{\theta} - \theta_0) \overset{d}{\to} \mathcal{N}(0, I^{-1}(\theta_0)) \).
Proof II

- Denote the score function $U = \frac{\partial}{\partial \theta} \log p_\theta(X)$ and another estimator by $T(X)$.
- Assume that $E[T(X)] = \theta_0$ (to rule out anomalies).
- Calculate $Cov(U, T) = E[UT] - E[U]E[T] = E[UT] - \theta_0 \cdot 0$
- Note that $E[UT] = \int T(x) \frac{\partial}{\partial \theta} \log p_{\theta_0}(x)p_{\theta_0}(x)dx$
  
  $= \frac{\partial}{\partial \theta} \int T(x)p_{\theta_0}(x)dx = \frac{\partial}{\partial \theta} \theta = 1$.

- Cauchy-Schwartz states $Cov(U, T) \leq \sqrt{Var[U]Var[T]}$.
- Thus $Var[T] \geq Var[\frac{\partial}{\partial \theta} \log p_\theta(X)]^{-1}$
- I.e. $Var[T] \geq Var[\hat{\theta}_{MLE}]$!
Proof III

- **Original problem**
  \[ \hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} p_{\theta}(X_1, \ldots, X_N). \]

- **Define** \( Y = g(X) \).
- **Calculations... Ok.**
Other estimators

- We can derive the asymptotical distribution using the same arguments for any *M-estimator*, i.e. for any estimator taking the estimate as the value that maximized/minimized a function of data.

- Similar arguments can also be used for *Z-estimators*, i.e. estimator taking the estimate as the value that solves a system of equations depending on data.
M-estimators

- Take any loss function $J(\theta)$, such that
  \[
  \hat{\theta} = \arg\max_{\theta \in \Theta} J(\theta). \tag{30}
  \]
- E.g. PEM: $J(\theta) = \sum_{i=1}^{N} \varepsilon_i(\theta)^2$.
- The corresponding problem is
  \[
  \frac{1}{\sqrt{N}} \partial_{\theta} J(\theta_0) + \frac{1}{N} \partial^2_{\theta} J(\theta_0) \sqrt{N} (\hat{\theta} - \theta_0) := 0.
  \]
- Assume that $\mathbb{E}[\frac{1}{\sqrt{N}} \partial_{\theta} J(\theta_0)] = 0$ (Why?),
  $\text{Var}[\frac{1}{\sqrt{N}} \partial_{\theta} J(\theta_0)] = \mathcal{W}$ and $\mathbb{E}[\frac{1}{N} \partial^2_{\theta} J(\theta_0)] = \mathcal{V}$. Applying the limit theorems as in the Maximum likelihood setup gives the asymptotics.
- Result: $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}^{-1} \mathcal{W}(\mathcal{V}^{-1})^T)$. 

Erik Lindström - erikl@maths.lth.se

Inference in non-linear time series
It is generally recommended to use to *Sandwich estimator* introduced by Halbert White in an econometrica paper in 1982.

\[ \sqrt{N}(\hat{\theta} - \theta_0) \stackrel{d}{\rightarrow} \mathcal{N}(0, V^{-1} W (V^{-1})^T). \]

as it is nearly impossible to know that you are using the correct model...
Z-estimators

- Take any loss function $G(\theta)$, such that
  \[ \hat{\theta} = \arg\max_{\theta \in \Theta} G(\theta) = 0. \quad (31) \]

- E.g. Estimation functions $G(\theta) = \sum g_i(\theta)(X_i - E[X_i])$

- Compute a first order Taylor expansion around $\theta_0$, and solve wrt $\theta$.

- Thus $G(\theta_0) + \partial_\theta G(\theta_0)(\hat{\theta} - \theta_0) := 0$.

- Multiply by $1/\sqrt{N}$ and apply limit theorems.

- Let $E[\frac{1}{\sqrt{N}} G(\theta_0)] = 0$ (why?), $\text{Var}[\frac{1}{\sqrt{N}} G(\theta_0)] = W$ and $E[\frac{1}{N} \partial_\theta G(\theta_0)] = V$.

- Then $\sqrt{N} (\hat{\theta} - \theta_0) \overset{d}{\to} \mathcal{N}(0, V^{-1} W (V^{-1})^T)$. 
The likelihood framework can be used to test the fit of models.

- Confidence intervals
- Likelihood ratio tests
Confidence intervals.

- Assume $\sqrt{N}(\hat{\theta} - \theta_0) \overset{d}{\to} \mathcal{N}(0, \Sigma)$.
- Then $l_{\theta_0} \approx \hat{\theta} \pm 2\sqrt{\frac{\text{diag}(\Sigma)}{N}}$.

**Question:** What happens if $\hat{\theta}$ has a heavy tailed distribution?

**Better accuracy:** Use profile likelihood!
Confidence intervals.

- Assume $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.
- Then $I_{\theta_0} \approx \hat{\theta} \pm 2\sqrt{\frac{\text{diag}(\Sigma)}{N}}$.
- Question: What happens if $\hat{\theta}$ has a heavy tailed distribution?
- Better accuracy: Use profile likelihood!
Likelihood based tests.

- Likelihood ratio tests

\[ \Lambda = \frac{\{ \sup L(X|\theta), \theta \in \Theta^F \}}{\{ \sup L(X|\theta), \theta \in \Theta^R \}} \]

where \( \Theta^R \subset \Theta^F \). Then

\[ 2 \log(\Lambda) \in \chi^2(k), \]

\( k \) being the degrees of freedom.

- LR tests are optimal, cf. Neyman-Pearson.

- and are closely related to AIC, BIC etc.
Likelihood ratios

- Compare \( L(X|\hat{\theta}) \) to \( L(X|\theta_0) \).
- Taylor expand

\[
2 \log(\Lambda) = 2 \left( \ell_N(\hat{\theta}) - \ell_N(\theta_0) \right).
\]

- We have that

\[
\ell_N(\theta) = \ell_N(\theta_0) + \partial_\theta \ell_N(\theta_0)(\theta - \theta_0) + \frac{1}{2} \partial^2_\theta \ell_N(\theta_0)(\theta - \theta_0)^2 + R.
\]

- The estimate must solve

\[
\partial_\theta \ell_N(\theta_0) + \partial^2_\theta \ell_N(\theta_0)(\hat{\theta} - \theta_0) = 0.
\]

- Thus

\[
2 \log(\Lambda) = 2 \left( \frac{1}{2} \partial_\theta \ell_N(\theta_0)(\hat{\theta} - \theta_0) \right).
\]
Likelihood ratios, cont.

Plugging in

\[(\hat{\theta} - \theta_0) = - (\partial^2_{\theta}\ell_N(\theta_0))^{-1} \partial_{\theta}\ell_N(\theta_0)\]

gives

\[2\log(\Lambda) = - \partial_{\theta}\ell_N(\theta_0)(\partial^2_{\theta}\ell_N(\theta_0))^{-1} \partial_{\theta}\ell_N(\theta_0)\]

Or nicer written

\[2\log(\Lambda) = \frac{1}{\sqrt{N}} \partial_{\theta}\ell_N(\theta_0)(-\frac{1}{N} \partial^2_{\theta}\ell_N(\theta_0))^{-1} \frac{1}{\sqrt{N}} \partial_{\theta}\ell_N(\theta_0).\]

This is a quadratic form, distributed as \(\chi^2(k)\), \(k\) being the degrees of freedom.
What about Bayesian methods

- Advances in MCMC has lead to an increased interest in these methods!
- Adaptive MCMC is a must! Read any recent (post 2008+) overview paper
- Moreover, suppose you only have an unbiased estimate of the likelihood. You can then run an augmented MCMC chain, simulating the quantities of interest and the random variables used to approximate the likelihood. The algorithm is called the *Pseudo Marginal MCMC* algorithm.
- The resulting chain will, quite surprisingly, target the correct distribution!, but the mixing will be worse.
- Results on optimal noise levels exists.

**Case:** Fitting a SETAR model.
What about Bayesian methods

- Advances in MCMC has lead to an increased interest in these methods!
- Adaptive MCMC is a must! Read any recent (post 2008+) overview paper

Moreover, suppose you only have an unbiased estimate of the likelihood. You can then run an augmented MCMC chain, simulating the quantities of interest and the random variables used to approximate the likelihood. The algorithm is called the Pseudo Marginal MCMC algorithm.

The resulting chain will, quite surprisingly, target the correct distribution!, but the mixing will be worse.

Results on optimal noise levels exists.

Case: Fitting a SETAR model.
What about Bayesian methods

- Advances in MCMC has lead to an increased interest in these methods!
- Adaptive MCMC is a must! Read any recent (post 2008+) overview paper
- Moreover, suppose you only have an unbiased estimate of the likelihood. You can then run an augmented MCMC chain, simulating the quantities of interest and the random variables used to approximate the likelihood. The algorithm is called the Pseudo Marginal MCMC algorithm
- The resulting chain will, quite surprisingly, target the correct distribution!, but the mixing will be worse.
- Results on optimal noise levels exists.

Case: Fitting a SETAR model.
What about Bayesian methods

- Advances in MCMC has lead to an increased interest in these methods!
- Adaptive MCMC is a must! Read any recent (post 2008+) overview paper
- Moreover, suppose you only have an unbiased estimate of the likelihood. You can then run an augmented MCMC chain, simulating the quantities of interest and the random variables used to approximate the likelihood. The algorithm is called the *Pseudo Marginal MCMC* algorithm
- The resulting chain will, quite surprisingly, target the correct distribution!, but the mixing will be worse.
- Results on optimal noise levels exists.

Case: Fitting a SETAR model.

Erik Lindström - erikl@maths.lth.se
What about Bayesian methods

- Advances in MCMC has lead to an increased interest in these methods!
- Adaptive MCMC is a must! Read any recent (post 2008+) overview paper
- Moreover, suppose you only have an unbiased estimate of the likelihood. You can then run an augmented MCMC chain, simulating the quantities of interest and the random variables used to approximate the likelihood. The algorithm is called the Pseudo Marginal MCMC algorithm.
- The resulting chain will, quite surprisingly, target the correct distribution!, but the mixing will be worse.
- Results on optimal noise levels exists.

Case: Fitting a SETAR model.
What if you don’t know the likelihood function? Assume you have

- **Data**, $y$
- Some prior belief on the parameters $p(\theta)$, e.g. they are positive and fairly small...
- and importantly, that you can *simulate* from the model, $y^* \sim p(y|\theta)$
Simple ABC algorithm

Draw $K$ samples from

- Simulate $\theta^k \sim p(\theta)$
- Generate $y^k \sim p(y|\theta^k)$
- Keep $\theta^k$ if $y^k == y$.

That algorithm samples from $p(\theta|y^*)$. 
Simple ABC algorithm

Draw $K$ samples from

- Simulate $\theta^k \sim p(\theta)$
- Generate $y^k \sim p(y|\theta^k)$
- Keep $\theta^k$ if $y^k == y$

That algorithm samples from $p(\theta|y^*)$. 
Consider the joint distribution of the algorithm

\[ f(\theta, y) \propto p(y|\theta)p(\theta)1_{\{y=y^*\}} \]  \hspace{1cm} (32)

Marginalize over \( y \) leading to

\[ \int f(y, \theta)dy \propto p(y^*|\theta)p(\theta) = p(\theta|y^*) \]  \hspace{1cm} (33)
Practical considerations, continuous data

That algorithm is very inefficient!

- Replace \( 1_{\{y=y^*\}} \) by \( y \approx y^* \) This leads to bias, cf. kernel smoothing, but higher acceptance rates.
- Replace \( y \approx y^* \) by \( S(y) \approx S(y^*) \) where \( S() \) are summary statistics. Solve the curse of dimensionality problem.
That algorithm is very inefficient!

- Replace $1_{\{y = y^*\}}$ by $y \approx y^*$ This leads to bias, cf. kernel smoothing, but higher acceptance rates.

- Replace $y \approx y^*$ by $S(y) \approx S(y^*)$ where $S()$ are summary statistics. Solve the curse of dimensionality problem.
That algorithm is very inefficient!

- Replace $1_{\{y=y^*\}}$ by $y \approx y^*$ This leads to bias, cf. kernel smoothing, but higher acceptance rates.
- Replace $y \approx y^*$ by $S(y) \approx S(y^*)$ where $S()$ are summary statistics. Solve the curse of dimensionality problem.
Even better, ABC-MCMC

Include the ABC framework in an MCMC algorithm. This leads to so-called likelihood-free methods.

- Let $J_\epsilon(y, y^*)$ measure the distance between the data and the simulations. Typically exponentially quadratic loss.
- The augmented posterior is then given by
  \[
  \pi(\theta, y^*|y) \propto J_\epsilon(y^*, y)p(y^*|\theta)p(\theta)
  \] (34)
  while proposals are drawn $(\theta', y')$ are drawn from
  \[
  q(\theta', \theta^*)p(y'|\theta')
  \] (35)
- Then accept draws with probability
  \[
  \alpha = \min \left(1, \frac{J_\epsilon(y^*, y)p(\theta^*)q(\theta'|\theta^*)}{J_\epsilon(y', y)p(\theta')q(\theta^*|\theta')}\right)
  \] (36)
  for some prior accepted data $y'$ and parameters $\theta'$. Note that the likelihood term cancels out between the proposal and the augmented posterior!
Even better, ABC-MCMC

Include the ABC framework in an MCMC algorithm. This leads to so called likelihood free methods.

- Let $J_\epsilon(y, y^*)$ measure the distance between the data and the simulations. Typically exponentially quadratic loss.
- The augmented posterior is then given by

$$
\pi(\theta, y^*|y) \propto J_\epsilon(y^*, y)p(y^*|\theta)p(\theta) \quad (34)
$$

while proposals are drawn $(\theta', y')$ are drawn from

$$
q(\theta', \theta^*)p(y'|\theta') \quad (35)
$$

- Then accept draws with probability

$$
\alpha = \min \left(1, \frac{J_\epsilon(y^*, y)p(\theta^*)q(\theta'|\theta^*)}{J_\epsilon(y', y)p(\theta')q(\theta^*|\theta')}\right) \quad (36)
$$

for some prior accepted data $y'$ and parameters $\theta'$. Note that the likelihood term cancels out between the proposal and the augmented posterior!
Even better, ABC-MCMC

Include the ABC framework in an MCMC algorithm. This leads to so called likelihood free methods.

- Let $J_\epsilon(y, y^*)$ measure the distance between the data and the simulations. Typically exponentially quadratic loss.

- The augmented posterior is then given by

$$
\pi(\theta, y^*|y) \propto J_\epsilon(y^*, y)p(y^*|\theta)p(\theta)
$$

while proposals are drawn $(\theta', y')$ are drawn from

$$
q(\theta', \theta^*)p(y'|\theta')
$$

- Then accept draws with probability

$$
\alpha = \min \left( 1, \frac{J_\epsilon(y^*, y)p(\theta^*)q(\theta'|\theta^*)}{J_\epsilon(y', y)p(\theta')q(\theta^*|\theta')} \right)
$$

for some prior accepted data $y'$ and parameters $\theta'$. Note that the likelihood term cancels out between the proposal and the augmented posterior!
Send feedback, questions and information about typos to erik.lindstrom@matstat.lu.se.