

Inference in non-linear time series

Erik Lindström

Centre for Mathematical Sciences
Lund University

LU/LTH & DTU



LUND UNIVERSITY

Overview

Introduction

- General properties

- Estimators

Least squares

- Properties

Maximum Likelihood

- MLE asymptotics

- Two theorems

- Proofs

Other estimators, M / Z / Bayesian

General

We are interested in estimating a parameter vector θ_0 from data $\mathbf{X}$.

- ▶ Ad hoc or formal estimation?
- ▶ Properties
- ▶ Definition

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

- ▶ Interpretation

Properties

- ▶ Bias $b = \theta_0 - \mathbf{E}[T_N(\mathbf{X})]$.
- ▶ Asympt. bias $\lim_{N \rightarrow \infty} \theta_0 - \mathbf{E}[T_N(\mathbf{X})]$.
- ▶ Consistency $\hat{\theta} \xrightarrow{P} \theta_0 \Leftrightarrow \mathbb{P}(\|\hat{\theta} - \theta_0\| > \varepsilon) \rightarrow 0$ as $N \rightarrow \infty$.
- ▶ Strong consistency $\hat{\theta} \xrightarrow{a.s.} \theta_0$.
- ▶ Efficiency

$$\text{Var}[T_N(\mathbf{X})] \geq I_N^{-1}(\theta_0),$$

where $I_N^{-1}(\theta_0)$ is the Fisher information matrix.

- ▶ Asympt. normality
- ▶ Convergence

$$N^\alpha(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(\mu, \Sigma).$$

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 sampling*),
Maximum Likelihood,
Quasi Maximum Likelihood
Approx Bayesian Computation - likelihood free inference

Least squares

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

Full details

$$Y = \begin{bmatrix} y_1 \\ \dots \\ y_N \end{bmatrix}, \quad Z = \begin{bmatrix} B_1(x_1) & \dots & B_J(x_N) \\ \dots & & \dots \\ B_1(x_N) & \dots & B_J(x_N) \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_J \end{bmatrix} \quad (1)$$

LS criterion:

$$(Y - Z\theta)^T (Y - Z\theta) \quad (2)$$

Least squares

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

Full details

$$Y = \begin{bmatrix} y_1 \\ \dots \\ y_N \end{bmatrix}, \quad Z = \begin{bmatrix} B_1(x_1) & \dots & B_J(x_1) \\ \dots & & \dots \\ B_1(x_N) & \dots & B_J(x_N) \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_J \end{bmatrix} \quad (1)$$

LS criterion:

$$(Y - Z\theta)^T (Y - Z\theta) \quad (2)$$

Mathematical details

- (Weighted) Parameter estimation:

$$\hat{\theta}_{LS} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{n=1}^N \lambda_n \left(y_n - \sum_j B_j(x_n) \theta_j \right)^2 \quad (3)$$

$$= (Y - Z\theta)^T W (Y - Z\theta) \quad (4)$$

with $W = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$.

- Solution: $\hat{\theta} = (Z^T W Z)^{-1} (Z^T W Y)$
- Proof. Use that $\nabla_{\theta} (\theta^T A \theta + \theta^T b + c) = 2A\theta + b$

Mathematical details

- (Weighted) Parameter estimation:

$$\hat{\theta}_{LS} = \operatorname{argmin}_{\theta \in \Theta} \sum_{n=1}^N \lambda_n \left(y_n - \sum_j B_j(x_n) \theta_j \right)^2 \quad (3)$$

$$= (Y - Z\theta)^T W (Y - Z\theta) \quad (4)$$

with $W = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$.

- Solution: $\hat{\theta} = (Z^T W Z)^{-1} (Z^T W Y)$
- Proof. Use that $\nabla_{\theta} (\theta^T A \theta + \theta^T b + c) = 2A\theta + b$

Adding regularization

Generalization (often with a Bayesian interpretation)

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + P(\theta, \theta_0) \quad (5)$$

Special Cases:

- ▶ Ridge regression (L2 penalty/Gaussian prior on parameters)
- ▶ LASSO (L1 penalty/Laplace prior)
- ▶ Elastic Net (L1 & L2)
- ▶ non-convex (Adaptive LASSO/Scad/MCP etc)
- ▶ Horseshoe priors - Monte Carlo sampling does not worry about convexity

Adding regularization

Generalization (often with a Bayesian interpretation)

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + P(\theta, \theta_0) \quad (5)$$

Special Cases:

- ▶ Ridge regression (L2 penalty/Gaussian prior on parameters)
- ▶ LASSO (L1 penalty/Laplace prior)
- ▶ Elastic Net (L1 & L2)
- ▶ non-convex (Adaptive LASSO/Scad/MCP etc)
- ▶ Horseshoe priors - Monte Carlo sampling does not worry about convexity

Assume W is an arbitrary positive definite matrix (=invertible)

- WLS/GLS estimate

$$\hat{\theta} = (Z^T W Z)^{-1} (Z^T W Y) \quad (6)$$

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

$$\mathbf{E} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W (Z\theta + \varepsilon)) \quad (7)$$

$$= \theta + (Z^T W Z)^{-1} (Z^T W \varepsilon) \quad (8)$$

- Note that $A \triangleq (Z^T W Z)^{-1} (Z^T W)$ is a deterministic matrix. Thus $\mathbf{E}[A\varepsilon] = A\mathbf{E}[\varepsilon] = 0$. Q: When is A not deterministic?
- Variance

$$\text{Var} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W \Omega W Z) (Z^T W Z)^{-1} \quad (9)$$

Assume W is an arbitrary positive definite matrix (=invertible)

- WLS/GLS estimate

$$\hat{\theta} = (Z^T W Z)^{-1} (Z^T W Y) \quad (6)$$

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

$$\mathbf{E} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W (Z\theta + \varepsilon)) \quad (7)$$

$$= \theta + (Z^T W Z)^{-1} (Z^T W \varepsilon) \quad (8)$$

- Note that $A \triangleq (Z^T W Z)^{-1} (Z^T W)$ is a deterministic matrix. Thus $\mathbf{E}[A\varepsilon] = A\mathbf{E}[\varepsilon] = 0$. Q: When is A not deterministic?
- Variance

$$\text{Var} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W \Omega W Z) (Z^T W Z)^{-1} \quad (9)$$

Assume W is an arbitrary positive definite matrix (=invertible)

- WLS/GLS estimate

$$\hat{\theta} = (Z^T W Z)^{-1} (Z^T W Y) \quad (6)$$

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

$$\mathbf{E} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W (Z\theta + \varepsilon)) \quad (7)$$

$$= \theta + (Z^T W Z)^{-1} (Z^T W \varepsilon) \quad (8)$$

- Note that $A \triangleq (Z^T W Z)^{-1} (Z^T W)$ is a deterministic matrix.
Thus $\mathbf{E}[A\varepsilon] = A\mathbf{E}[\varepsilon] = 0$. *Q: When is A not deterministic?*

- Variance

$$\text{Var} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W \Omega W Z) (Z^T W Z)^{-1} \quad (9)$$

Assume W is an arbitrary positive definite matrix (=invertible)

- WLS/GLS estimate

$$\hat{\theta} = (Z^T W Z)^{-1} (Z^T W Y) \quad (6)$$

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

$$\mathbf{E} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W (Z\theta + \varepsilon)) \quad (7)$$

$$= \theta + (Z^T W Z)^{-1} (Z^T W \varepsilon) \quad (8)$$

- Note that $A \triangleq (Z^T W Z)^{-1} (Z^T W)$ is a deterministic matrix. Thus $\mathbf{E}[A\varepsilon] = A\mathbf{E}[\varepsilon] = 0$. Q: When is A not deterministic?
- Variance

$$\text{Var} [\hat{\theta}] = (Z^T W Z)^{-1} (Z^T W \Omega W Z) (Z^T W Z)^{-1} \quad (9)$$

cont...

- Simplifies if $W = \Omega^{-1}$ to

$$\text{Var} \left[\hat{\theta} \right] = (Z^T W Z)^{-1}, \quad (10)$$

and further if $\Omega = \sigma^2 I_N$. We then get

$$\text{Var} \left[\hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}. \quad (11)$$

- More complicated when the model is wrong...
- Model does not include the true model $\Rightarrow$ in general biased estimates
- Model includes unnecessary predictor $\Rightarrow$ nuisance variance

cont...

- Simplifies if $W = \Omega^{-1}$ to

$$\text{Var} \left[\hat{\theta} \right] = (Z^T W Z)^{-1}, \quad (10)$$

and further if $\Omega = \sigma^2 I_N$. We then get

$$\text{Var} \left[\hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}. \quad (11)$$

- More complicated when the model is wrong...
- Model does not include the true model $\Rightarrow$ in general biased estimates
- Model includes unnecessary predictor $\Rightarrow$ nuisance variance

cont...

- Simplifies if $W = \Omega^{-1}$ to

$$\text{Var} \left[\hat{\theta} \right] = (Z^T W Z)^{-1}, \quad (10)$$

and further if $\Omega = \sigma^2 I_N$. We then get

$$\text{Var} \left[\hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}. \quad (11)$$

- More complicated when the model is wrong...
- Model does not include the true model $\Rightarrow$ in general biased estimates
- Model includes unnecessary predictor $\Rightarrow$ nuisance variance

cont...

- Simplifies if $W = \Omega^{-1}$ to

$$\text{Var} \left[\hat{\theta} \right] = (Z^T W Z)^{-1}, \quad (10)$$

and further if $\Omega = \sigma^2 I_N$. We then get

$$\text{Var} \left[\hat{\theta} \right] = \sigma^2 (Z^T Z)^{-1}. \quad (11)$$

- More complicated when the model is wrong...
- Model does not include the true model $\Rightarrow$ in general biased estimates
- Model includes unnecessary predictor $\Rightarrow$ nuisance variance

F-tests

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

$$f = \frac{\frac{RSS_R - RSS_{UR}}{|\theta_{UR}| - |\theta_R|}}{\frac{RSS_{UR}}{N - |\theta_{UR}|}} \quad (12)$$

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

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

$$\mathbf{E}[Q] = \text{Tr}(\mathbf{E}[Y^T(I_N - P)^T(I_N - P)Y]) \quad (13)$$

$$= \text{Tr}((I_N - P)^T(I_N - P) \text{Cov}[Y, Y]) \quad (14)$$

- 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

$$\text{Tr}((I_N - P)^T(I_N - P) \text{Cov}[Y, Y]) = \sigma^2 \text{Tr}(I_N - P) \quad (15)$$

$$= \sigma^2 \text{Tr}(I_N) - \sigma^2 \text{Tr}(P) \quad (16)$$

$$= \sigma^2 N - \sigma^2 \text{Tr}(Z(Z^T Z)^{-1} Z^T) \quad (17)$$

$$= \sigma^2(N - \text{Tr}((Z^T Z)^{-1}(Z^T Z))) = \sigma^2(N - J) \quad (18)$$

What about the penalized asymptotics?

What if the estimation is given by

$$\hat{\theta}_{PLS} = \underset{\theta \in \Theta}{\operatorname{argmin}} (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D (\theta - \theta_0) \quad (19)$$

Derive on blackboard. [bias, degrees of freedom] *Case study on Semi Parametric Lag Dependent functions*, Lindström [2013]

Another popular form is LASSO, Tibshirani [1996]

$$\hat{\theta}_{Adaptive \text{ LASSO}} = \underset{\theta \in \Theta}{\operatorname{argmin}} (Y - Z\theta)^T W (Y - Z\theta) + ||D(\theta - \theta_0)||_1 \quad (20)$$

Easily computed using the ADMM algorithm. Case study in Matlab

What about the penalized asymptotics?

What if the estimation is given by

$$\hat{\theta}_{PLS} = \underset{\theta \in \Theta}{\operatorname{argmin}} (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D (\theta - \theta_0) \quad (19)$$

Derive on blackboard. [bias, degrees of freedom] *Case study on Semi Parametric Lag Dependent functions*, Lindström [2013]

Another popular form is LASSO, Tibshirani [1996]

$$\hat{\theta}_{Adaptive \text{ LASSO}} = \underset{\theta \in \Theta}{\operatorname{argmin}} (Y - Z\theta)^T W (Y - Z\theta) + ||D(\theta - \theta_0)||_1 \quad (20)$$

Easily computed using the ADMM algorithm. Case study in Matlab

What about the penalized asymptotics?

What if the estimation is given by

$$\hat{\theta}_{PLS} = \underset{\theta \in \Theta}{\operatorname{argmin}} (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D (\theta - \theta_0) \quad (19)$$

Derive on blackboard. [bias, degrees of freedom] *Case study on Semi Parametric Lag Dependent functions*, Lindström [2013]

Another popular form is LASSO, Tibshirani [1996]

$$\hat{\theta}_{Adaptive \text{ LASSO}} = \underset{\theta \in \Theta}{\operatorname{argmin}} (Y - Z\theta)^T W (Y - Z\theta) + \|D(\theta - \theta_0)\|_1 \quad (20)$$

Easily computed using the ADMM algorithm. Case study in Matlab

Maximum Likelihood estimators

Defined as

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} p_{\theta}(X_1, \dots, X_N).$$

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

- ▶ Ia. Consistency $\hat{\theta} \xrightarrow{P} \theta_0$.
- ▶ Ib. Asympt. normality
- ▶ II. Efficiency $\operatorname{Var}[T_N(\mathbf{X})] = I_F^{-1}(\theta_0)$,
- ▶ III. Invariant under reparameterization.

We have under general conditions that

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

Two useful theorems, LLN & CLT

Assume that X_i is *iid* with $\mathbf{E}[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2$

► **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

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

and

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

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

Proofs

- $\mathbf{E} \left[\frac{\partial}{\partial \theta} \log p_{\theta}(X) \right] = \int \frac{\partial}{\partial \theta} \log p_{\theta}(x) p_{\theta}(x) dx$
- Note that $\frac{\partial}{\partial \theta} \log p_{\theta}(X) = \frac{\frac{\partial p_{\theta}(x)}{\partial \theta}}{p_{\theta}(x)}$
- Thus, the expression simplifies to

$$\int \frac{\frac{\partial p_{\theta}(x|\theta)}{\partial \theta}}{p_{\theta}(x)} p_{\theta}(x) dx = \dots = \frac{\partial}{\partial \theta} \int p_{\theta}(x) dx = \dots = 0 \quad (21)$$

Proofs

- ▶ $\mathbf{E} \left[\frac{\partial}{\partial \theta} \log p_{\theta}(X) \right] = \int \frac{\partial}{\partial \theta} \log p_{\theta}(x) p_{\theta}(x) dx$
- ▶ Note that $\frac{\partial}{\partial \theta} \log p_{\theta}(X) = \frac{\frac{\partial p_{\theta}(x)}{\partial \theta}}{p_{\theta}(x)}$
- ▶ Thus, the expression simplifies to

$$\int \frac{\frac{\partial p_{\theta}(x|\theta)}{\partial \theta}}{p_{\theta}(x)} p_{\theta}(x) dx = \dots = \frac{\partial}{\partial \theta} \int p_{\theta}(x) dx = \dots = 0 \quad (21)$$

Proofs

- ▶ $\mathbf{E} \left[\frac{\partial}{\partial \theta} \log p_{\theta}(X) \right] = \int \frac{\partial}{\partial \theta} \log p_{\theta}(x) p_{\theta}(x) dx$
- ▶ Note that $\frac{\partial}{\partial \theta} \log p_{\theta}(X) = \frac{\frac{\partial p_{\theta}(x)}{\partial \theta}}{p_{\theta}(x)}$
- ▶ Thus, the expression simplifies to

$$\int \frac{\frac{\partial p_{\theta}(x|\theta)}{\partial \theta}}{p_{\theta}(x)} p_{\theta}(x) dx = \dots = \frac{\partial}{\partial \theta} \int p_{\theta}(x) dx = \dots = 0 \quad (21)$$

- The Fisher information was defined as

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

- This is by definition

$$\mathbf{E} \left[\left(\frac{\partial}{\partial \theta} \log p_{\theta}(X) - \underbrace{\mathbf{E} \left[\frac{\partial}{\partial \theta} \log p_{\theta}(X|\theta) \right]}_{=0} \right)^2 \right]$$

- Recall the lemma $\frac{\partial}{\partial u} \frac{f(u)}{g(u)} = \frac{\partial_u f}{g} - \frac{f(u) \partial_u g(u)}{g(u)^2}$. Then

$$- \mathbf{E} \left[\frac{\partial^2}{\partial \theta^2} \log p_{\theta}(X) \right] = - \mathbf{E} \left[\frac{\partial}{\partial \theta} \frac{\partial_{\theta} p_{\theta}}{p_{\theta}} \right] \quad (22)$$

$$= - \mathbf{E} \left[\frac{\partial_{\theta}^2 p_{\theta}}{p_{\theta}} - \frac{\partial_{\theta} p_{\theta} \partial_{\theta} p_{\theta}}{p_{\theta}^2} \right] \quad (23)$$

$$= - \underbrace{\int \frac{\partial_{\theta}^2 p_{\theta}(x)}{p_{\theta}(x)} p_{\theta}(x) dx}_{=0} + \mathbf{E} \left[\left(\frac{\partial}{\partial \theta} \log p_{\theta}(X) \right)^2 \right] \quad (24)$$

- The Fisher information was defined as

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

- This is by definition

$$\mathbf{E} \left[\left(\frac{\partial}{\partial \theta} \log p_{\theta}(X) - \underbrace{\mathbf{E} \left[\frac{\partial}{\partial \theta} \log p_{\theta}(X|\theta) \right]}_{=0} \right)^2 \right]$$

- Recall the lemma $\frac{\partial}{\partial u} \frac{f(u)}{g(u)} = \frac{\partial_u f}{g} - \frac{f(u) \partial_u g(u)}{g(u)^2}$. Then

$$- \mathbf{E} \left[\frac{\partial^2}{\partial \theta^2} \log p_{\theta}(X) \right] = - \mathbf{E} \left[\frac{\partial}{\partial \theta} \frac{\partial_{\theta} p_{\theta}}{p_{\theta}} \right] \quad (22)$$

$$= - \mathbf{E} \left[\frac{\partial_{\theta}^2 p_{\theta}}{p_{\theta}} - \frac{\partial_{\theta} p_{\theta} \partial_{\theta} p_{\theta}}{p_{\theta}^2} \right] \quad (23)$$

$$= - \underbrace{\int \frac{\partial_{\theta}^2 p_{\theta}(x)}{p_{\theta}(x)} p_{\theta}(x) dx}_{=0} + \mathbf{E} \left[\left(\frac{\partial}{\partial \theta} \log p_{\theta}(X) \right)^2 \right] \quad (24)$$

- The Fisher information was defined as

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

- This is by definition

$$\mathbf{E} \left[\left(\frac{\partial}{\partial \theta} \log p_{\theta}(X) - \underbrace{\mathbf{E} \left[\frac{\partial}{\partial \theta} \log p_{\theta}(X|\theta) \right]}_{=0} \right)^2 \right]$$

- Recall the lemma $\frac{\partial}{\partial u} \frac{f(u)}{g(u)} = \frac{\partial_u f}{g} - \frac{f(u) \partial_u g(u)}{g(u)^2}$. Then

$$- \mathbf{E} \left[\frac{\partial^2}{\partial \theta^2} \log p_{\theta}(X) \right] = - \mathbf{E} \left[\frac{\partial}{\partial \theta} \frac{\partial_{\theta} p_{\theta}}{p_{\theta}} \right] \quad (22)$$

$$= - \mathbf{E} \left[\frac{\partial_{\theta}^2 p_{\theta}}{p_{\theta}} - \frac{\partial_{\theta} p_{\theta} \partial_{\theta} p_{\theta}}{p_{\theta}^2} \right] \quad (23)$$

$$= - \underbrace{\int \frac{\partial_{\theta}^2 p_{\theta}(x)}{p_{\theta}(x)} p_{\theta}(x) dx}_{=0} + \mathbf{E} \left[\left(\frac{\partial}{\partial \theta} \log p_{\theta}(X) \right)^2 \right] \quad (24)$$

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

$$\ell(\theta) = \sum_{n=1}^N \log p_{\theta}(x_n | x_{1:n-1}) \quad (25)$$

The estimate is given by

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta) = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \ell(\theta) \quad (26)$$

$$= \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^N \log p_{\theta}(x_n | x_{1:n-1}) \quad (27)$$

Now we rewrite this as

$$\frac{1}{N} \ell(\theta) = \frac{1}{N} \ell(\theta) \pm \mathbf{E}[\ell(\theta)] \quad (28)$$

$$= \mathbf{E}[\ell(\theta)] + \left(\frac{1}{N} \ell(\theta) - \mathbf{E}[\ell(\theta)] \right) \quad (29)$$

Consistency

The log-likelihood function is defined as

$$\ell(\theta) = \sum_{n=1}^N \log p_{\theta}(x_n | x_{1:n-1}) \quad (25)$$

The estimate is given by

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta) = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \ell(\theta) \quad (26)$$

$$= \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^N \log p_{\theta}(x_n | x_{1:n-1}) \quad (27)$$

Now we rewrite this as

$$\frac{1}{N} \ell(\theta) = \frac{1}{N} \ell(\theta) \pm \mathbf{E}[\ell(\theta)] \quad (28)$$

$$= \mathbf{E}[\ell(\theta)] + \left(\frac{1}{N} \ell(\theta) - \mathbf{E}[\ell(\theta)] \right) \quad (29)$$

Consistency

The log-likelihood function is defined as

$$\ell(\theta) = \sum_{n=1}^N \log p_{\theta}(x_n | x_{1:n-1}) \quad (25)$$

The estimate is given by

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta) = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \ell(\theta) \quad (26)$$

$$= \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^N \log p_{\theta}(x_n | x_{1:n-1}) \quad (27)$$

Now we rewrite this as

$$\frac{1}{N} \ell(\theta) = \frac{1}{N} \ell(\theta) \pm \mathbf{E}[\ell(\theta)] \quad (28)$$

$$= \mathbf{E}[\ell(\theta)] + \left(\frac{1}{N} \ell(\theta) - \mathbf{E}[\ell(\theta)] \right) \quad (29)$$

Consistency cont.

This decomposition shows that

$$\frac{1}{N}\ell(\theta) = \underbrace{\mathbf{E}[\ell(\theta)]}_{\text{Expected log-likelihood}} + \underbrace{\frac{1}{N}\ell(\theta) - \mathbf{E}[\ell(\theta)]}_{\text{Random error}} \quad (30)$$

It also holds trivially that

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} \mathbf{E}[\ell(\theta)] + \left(\hat{\theta}_{MLE} - \operatorname{argmax}_{\theta \in \Theta} \mathbf{E}[\ell(\theta)] \right) \quad (31)$$

First we look at the random error, that is written so that we can apply the *Law of Large Numbers*. If

$$\lim_{N \rightarrow \infty} \frac{1}{N} \ell(\theta) - \mathbf{E}[\ell(\theta)] = 0 \quad (32)$$

uniformly over Θ , then it follows that

$$\lim_{N \rightarrow \infty} \hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} \mathbf{E}[\ell(\theta)] \quad (33)$$

- 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

$$\mathbf{E}[\log q_\theta(X)] - \mathbf{E}[\log p_{\theta_0}(X)] = \int (\log q_\theta(x) - \log p_{\theta_0}(x)) p_{\theta_0}(x) dx \quad (34)$$

$$= \int \log \left(\frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx \quad (35)$$

- Log is a concave function. We know from Jensen's inequality that $\mathbf{E}[\log(\xi)] \leq \log(\mathbf{E}[\xi])$

- 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

$$\mathbf{E}[\log q_\theta(X)] - \mathbf{E}[\log p_{\theta_0}(X)] = \int (\log q_\theta(x) - \log p_{\theta_0}(x)) p_{\theta_0}(x) dx \quad (34)$$

$$= \int \log \left(\frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx \quad (35)$$

- Log is a concave function. We know from Jensen's inequality that $\mathbf{E}[\log(\xi)] \leq \log(\mathbf{E}[\xi])$

- ▶ 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

$$\mathbf{E}[\log q_\theta(X)] - \mathbf{E}[\log p_{\theta_0}(X)] = \int (\log q_\theta(x) - \log p_{\theta_0}(x)) p_{\theta_0}(x) dx \quad (34)$$

$$= \int \log \left(\frac{q_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx \quad (35)$$

- ▶ Log is a concave function. We know from Jensen's inequality that $\mathbf{E}[\log(\xi)] \leq \log(\mathbf{E}[\xi])$

That means that

$$\int \log \left(\frac{q_{\theta}(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx \leq \log \left(\int \frac{q_{\theta}(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) dx \right) = 0 \quad (36)$$

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

The conclusion is that

$$\theta_0 = \operatorname{argmax}_{\theta \in \Theta} \mathbf{E}[\ell(\theta)] \quad (37)$$

and hence that

$$\lim_{N \rightarrow \infty} \hat{\theta}_{MLE} = \theta_0 \quad (38)$$

as the random error vanishes as $N \rightarrow \infty$.

That means that

$$\int \log \left(\frac{q_{\theta}(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x) dx \leq \log \left(\int \frac{q_{\theta}(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) dx \right) = 0 \quad (36)$$

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

$$\theta_0 = \operatorname{argmax}_{\theta \in \Theta} \mathbf{E}[\ell(\theta)] \quad (37)$$

and hence that

$$\lim_{N \rightarrow \infty} \hat{\theta}_{MLE} = \theta_0 \quad (38)$$

as the random error vanishes as $N \rightarrow \infty$.

Proof Ib

- ▶ Rewrite $L(\mathbf{X}) = \exp \left(\log p_{\theta}(X_1) + \sum_{n=2}^N \log p_{\theta}(X_n | X_{1:n-1}) \right)$.
- ▶ Second order Taylor expansion around θ_0 . This is fine due to consistency.
- ▶ Maximize wrt θ
- ▶ $\sqrt{N}(\hat{\theta} - \theta_0) \approx \dots$
- ▶ Asymp. normality follows. **Note:** Finite sample properties may be different!

Proof Ib

- ▶ Rewrite $L(\mathbf{X}) = \exp \left(\log p_{\theta}(X_1) + \sum_{n=2}^N \log p_{\theta}(X_n | X_{1:n-1}) \right)$.
- ▶ Second order Taylor expansion around θ_0 . This is fine due to consistency.
- ▶ Maximize wrt θ
- ▶ $\sqrt{N}(\hat{\theta} - \theta_0) \approx \dots$
- ▶ Asymp. normality follows. **Note:** Finite sample properties may be different!

Proof I, Ext.

- ▶ Write $\ell_N(\theta) = \log L(\mathbf{X}|\theta)$.
- ▶ This is a sum consisting of N terms (think LLN and CLT!).
- ▶ Approximate $\ell_N(\theta)$ using a second order Taylor expansion around θ_0 .
- ▶ Thus, with a mild abuse of notation

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

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

Proof I, Ext.

- ▶ Write $\ell_N(\theta) = \log L(\mathbf{X}|\theta)$.
- ▶ This is a sum consisting of N terms (think LLN and CLT!).
- ▶ Approximate $\ell_N(\theta)$ using a second order Taylor expansion around θ_0 .
- ▶ Thus, with a mild abuse of notation

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

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

Proof I, Ext.

- ▶ We have from LLN that

$$\frac{1}{N} \partial_{\theta}^2 \ell_N(\theta_0) \rightarrow -I(\theta_0),$$

- ▶ and from CLT that

$$\frac{1}{\sqrt{N}} \partial_{\theta} \ell_N(\theta_0) \rightarrow Z,$$

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

- ▶ Rearranging in combination with Slutsky's theorem gives

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0))$$

.

Proof II

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

$$\begin{aligned}\mathbf{E}[UT] &= \int T(x) \frac{\partial}{\partial \theta} \log p_{\theta_0}(x) p_{\theta_0}(x) dx \\ &= \frac{\partial}{\partial \theta} \int T(x) p_{\theta_0}(x) dx = \frac{\partial}{\partial \theta} \theta = 1.\end{aligned}$$

- ▶ Cauchy-Schwartz states $\text{Cov}(U, T) \leq \sqrt{\text{Var}[U] \text{Var}[T]}$.
- ▶ Thus $\text{Var}[T] \geq \text{Var}[\frac{\partial}{\partial \theta} \log p_{\theta}(X)]^{-1}$
- ▶ This implies that $\text{Var}[T] \geq \text{Var}[\hat{\theta}_{MLE}]$!

Proof III

- ▶ Original problem

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta \in \Theta} p_{\theta}(X_1, \dots, X_N).$$

- ▶ 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 many similar estimators taking the estimate as the value that maximized/minimizes 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.

Q: Identical when...

Other estimators

- ▶ We can derive the asymptotical distribution using the same arguments for any *M-estimator*, i.e. for many similar estimators taking the estimate as the value that maximized/minimizes 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.

Q: Identical when...

M-estimators

- Take any loss function $J(\theta)$, such that

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} J(\theta). \quad (39)$$

- E.g. PEM: $J(\theta) = \sum_{i=1}^N \varepsilon_i(\theta)^2$.
- The corresponding problem is

$$\frac{1}{\sqrt{N}} \partial_{\theta} J(\theta_0) + \frac{1}{N} \partial_{\theta}^2 J(\theta_0) \sqrt{N}(\hat{\theta} - \theta_0) := 0.$$

- Assume that $\mathbf{E}[\frac{1}{\sqrt{N}} \partial_{\theta} J(\theta_0)] = 0$ (Why?), $\operatorname{Var}[\frac{1}{\sqrt{N}} \partial_{\theta} J(\theta_0)] = W$ and $\mathbf{E}[\frac{1}{N} \partial_{\theta}^2 J(\theta_0)] = V$. Applying the limit theorems as in the Maximum likelihood setup gives the asymptotics.
- Result: $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V^{-1} W (V^{-1})^T)$.

Sandwich estimator

It is generally recommended to use the *Sandwich estimator* introduced by Halbert White in an econometrica paper in 1982.

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, V^{-1}W(V^{-1})^T\right).$$

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

Z-estimators

- Take any loss function $G(\theta)$, such that

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} G(\theta) = 0. \quad (40)$$

- E.g. Estimation functions $G(\theta) = \sum g_i(\theta)(X_i - \mathbf{E}[X_i])$
- Compute a first order Taylor expansion around θ_0 , and solve wrt θ .
- Thus $G(\theta_0) + \partial_{\theta} G(\theta_0)(\hat{\theta} - \theta_0) := 0$.
- Multiply by $1/\sqrt{N}$ and apply limit theorems.
- Let $\mathbf{E}[\frac{1}{\sqrt{N}} G(\theta_0)] = 0$ (why?), $\operatorname{Var}[\frac{1}{\sqrt{N}} G(\theta_0)] = W$ and $\mathbf{E}[\frac{1}{N} \partial_{\theta} G(\theta_0)] = V$.
- Then $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V^{-1}W(V^{-1})^T)$.

The likelihood framework can be used to test the fit of models.

- ▶ Confidence intervals
- ▶ Likelihood ratio tests

Confidence intervals.

- ▶ Assume $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.
- ▶ 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!

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

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

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

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

- ▶ We have that

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

- ▶ The estimate must solve

$$\partial_{\theta} \ell_N(\theta_0) + \partial_{\theta}^2 \ell_N(\theta_0)(\hat{\theta} - \theta_0) = 0.$$

- ▶ Thus

$$2 \log(\Lambda) = 2 \left(\frac{1}{2} \partial_{\theta}^2 \ell_N(\theta_0)(\hat{\theta} - \theta_0) \right).$$

Likelihood ratios, cont.

- Plugging in

$$(\hat{\theta} - \theta_0) = -(\partial_{\theta}^2 \ell_N(\theta_0))^{-1} \partial_{\theta} \ell_N(\theta_0)$$

gives



$$2 \log(\Lambda) = -\partial_{\theta} \ell_N(\theta_0) (\partial_{\theta}^2 \ell_N(\theta_0))^{-1} \partial_{\theta} \ell_N(\theta_0)$$

- Or nicer written, and using Slutsky's theorem

$$2 \log(\Lambda) = \frac{1}{\sqrt{N}} \partial_{\theta} \ell_N(\theta_0) \left(-\frac{1}{N} \partial_{\theta}^2 \ell_N(\theta_0) \right)^{-1} \frac{1}{\sqrt{N}} \partial_{\theta} \ell_N(\theta_0).$$

- This is a quadratic form, distributed as $\chi^2(k)$, k being the degrees of freedom.

What about Bayesian methods

- ▶ Advances in MCMC and computational resources 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, Andrieu et al. [2009]
- ▶ The resulting chain will, quite surprisingly, *target the correct distribution!*, but the *mixing will be worse*.
- ▶ Results on optimal noise levels exists, balancing computational efforts vs mixing rate.

What about Bayesian methods

- ▶ Advances in MCMC and computational resources 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, Andrieu et al. [2009]
- ▶ The resulting chain will, quite surprisingly, *target the correct distribution!*, but the *mixing will be worse*.
- ▶ Results on optimal noise levels exists, balancing computational efforts vs mixing rate.

What about Bayesian methods

- ▶ Advances in MCMC and computational resources 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, Andrieu et al. [2009]
- ▶ The resulting chain will, quite surprisingly, *target the correct distribution!*, but the *mixing will be worse*.
- ▶ Results on optimal noise levels exists, balancing computational efforts vs mixing rate.

What about Bayesian methods

- ▶ Advances in MCMC and computational resources 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, Andrieu et al. [2009]
- ▶ The resulting chain will, quite surprisingly, *target the correct distribution!*, but the *mixing will be worse*.
- ▶ Results on optimal noise levels exists, balancing computational efforts vs mixing rate.

What about Bayesian methods

- ▶ Advances in MCMC and computational resources 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, Andrieu et al. [2009]
- ▶ The resulting chain will, quite surprisingly, *target the correct distribution!*, but the *mixing will be worse*.
- ▶ Results on optimal noise levels exists, balancing computational efforts vs mixing rate.

Crash course on MCMC

- ▶ Standard Monte Carlo generates *iid* samples from some distribution $p(x)$.
- ▶ *Markov Chain* Monte Carlo generates dependent samples, where the stationary distribution for the Markov Chain is $p(x)$.
- ▶ Example. Say that we are interested in sampling from a Gaussian $N(\mu, \sigma^2)$.
- ▶ The AR(1) process

$$X_n - \mu = a(X_{n-1} - \mu) + \sqrt{1 - a^2}\sigma e_n, \quad e_n \sim N(0, 1) \quad (41)$$

has the correct stationary distribution for any $|a| < 1$.

- ▶ **Q:** What is the optimal value for a ?

Crash course on MCMC

- ▶ Standard Monte Carlo generates *iid* samples from some distribution $p(x)$.
- ▶ *Markov Chain* Monte Carlo generates dependent samples, where the stationary distribution for the Markov Chain is $p(x)$.
- ▶ Example. Say that we are interested in sampling from a Gaussian $N(\mu, \sigma^2)$.
- ▶ The AR(1) process

$$X_n - \mu = a(X_{n-1} - \mu) + \sqrt{1 - a^2}\sigma e_n, \quad e_n \sim N(0, 1) \quad (41)$$

has the correct stationary distribution for any $|a| < 1$.

- ▶ Q: What is the optimal value for a ?

Crash course on MCMC

- ▶ Standard Monte Carlo generates *iid* samples from some distribution $p(x)$.
- ▶ *Markov Chain* Monte Carlo generates dependent samples, where the stationary distribution for the Markov Chain is $p(x)$.
- ▶ Example. Say that we are interested in sampling from a Gaussian $N(\mu, \sigma^2)$.
- ▶ The AR(1) process

$$X_n - \mu = a(X_{n-1} - \mu) + \sqrt{1 - a^2}\sigma e_n, \quad e_n \sim N(0, 1) \quad (41)$$

has the correct stationary distribution for any $|a| < 1$.

- ▶ **Q:** What is the optimal value for a ?

Metropolis-Hastings algorithm, crash course

Define **proposal distribution** $g(u'|u_{n-1})$

► Example: Gaussian random walk $u' = u_{n-1} + \sigma e_n$

►
$$g(u'|u_{n-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(u' - u_{n-1})^2}{\sigma^2}}$$

Define the **acceptance probability** $A(u'|u_{n-1})$

$$A(u'|u_{n-1}) = \min \left(1, \frac{p(u')}{p(u_{n-1})} \frac{g(u_{n-1}|u')}{g(u'|u_{n-1})} \right) \quad (42)$$

Metropolis-Hastings algorithm, crash course

Define **proposal distribution** $g(u'|u_{n-1})$

► Example: Gaussian random walk $u' = u_{n-1} + \sigma e_n$

► $g(u'|u_{n-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(u' - u_{n-1})^2}{\sigma^2}}$

Define the **acceptance probability** $A(u'|u_{n-1})$

$$A(u'|u_{n-1}) = \min \left(1, \frac{p(u')}{p(u_{n-1})} \frac{g(u_{n-1}|u')}{g(u'|u_{n-1})} \right) \quad (42)$$

Algorithm

- ▶ Initialize u_0 .
- ▶ For $n = 1 : N$ do
 - ▶ $u' \sim g(u|u_{n-1})$
 - ▶ Select u_n according to

$$u_n = \begin{cases} u' & \text{with probability } A(u'|u_{n-1}) \\ u_{n-1} & \text{with probability } 1 - A(u'|u_{n-1}) \end{cases} \quad (43)$$

Practical considerations, use a burnin

Case: Generate samples from Exponential distribution.

Case: Fitting a SETAR model.

Algorithm

- ▶ Initialize u_0 .
- ▶ For $n = 1 : N$ do
 - ▶ $u' \sim g(u|u_{n-1})$
 - ▶ Select u_n according to

$$u_n = \begin{cases} u' & \text{with probability } A(u'|u_{n-1}) \\ u_{n-1} & \text{with probability } 1 - A(u'|u_{n-1}) \end{cases} \quad (43)$$

Practical considerations, use a burnin

Case: Generate samples from Exponential distribution.

Case: Fitting a SETAR model.

Algorithm

- ▶ Initialize u_0 .
- ▶ For $n = 1 : N$ do
 - ▶ $u' \sim g(u|u_{n-1})$
 - ▶ Select u_n according to

$$u_n = \begin{cases} u' & \text{with probability } A(u'|u_{n-1}) \\ u_{n-1} & \text{with probability } 1 - A(u'|u_{n-1}) \end{cases} \quad (43)$$

Practical considerations, use a burnin

Case: Generate samples from Exponential distribution.

Case: Fitting a SETAR model.

Approximate Bayesian Computation

What if you don't know the likelihood function (say large, mechanistic model)? Assume you have

- ▶ Data, y^D
- ▶ 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 $y^{*,k} == y^D$.

That algorithm samples from $p(\theta|y^D)$, Diggle and Gratton [1984].

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 $y^{*,k} == y^D$.

That algorithm samples from $p(\theta|y^D)$, Diggle and Gratton [1984].

Proof

- Consider the joint distribution of the algorithm

$$f(\theta, y) \propto p(y|\theta)p(\theta)\mathbf{1}_{\{y=y^D\}} \quad (44)$$

- Marginalize over y leading to

$$\int f(y, \theta) dy \propto p(y^D|\theta)p(\theta) = p(\theta|y^D) \quad (45)$$

Practical considerations, continuous data

That algorithm is very inefficient!

- ▶ Replace $\mathbf{1}_{\{y^*=y^D\}}$ by $y^* \approx y^D$, formally as $\rho(y^*, y^D) < \epsilon$. This leads to bias, cf. kernel smoothing, but higher acceptance rates, Tavaré et al. [1997]
- ▶ Replace $y \approx y^*$ by $S(y) \approx S(y^*)$ where $S()$ are summary statistics, thus $\rho(S(y^*), S(y^D)) < \epsilon$. This solves the curse of dimensionality problem, Pritchard et al. [1999]

Case: ABC for Gaussian data. Conclusion: $\epsilon \rightarrow 0$ yields $p(\theta|y^D)$, $\epsilon \rightarrow \infty$ yields $p(\theta)$.

Matlab code ABCtest

Practical considerations, continuous data

That algorithm is very inefficient!

- ▶ Replace $\mathbf{1}_{\{y^*=y^D\}}$ by $y^* \approx y^D$, formally as $\rho(y^*, y^D) < \epsilon$. This leads to bias, cf. kernel smoothing, but higher acceptance rates, Tavaré et al. [1997]
- ▶ Replace $y \approx y^*$ by $S(y) \approx S(y^*)$ where $S()$ are summary statistics, thus $\rho(S(y^*), S(y^D)) < \epsilon$. This solves the curse of dimensionality problem, Pritchard et al. [1999]

Case: ABC for Gaussian data. Conclusion: $\epsilon \rightarrow 0$ yields $p(\theta|y^D)$, $\epsilon \rightarrow \infty$ yields $p(\theta)$.

Matlab code ABCtest

Practical considerations, continuous data

That algorithm is very inefficient!

- ▶ Replace $\mathbf{1}_{\{y^*=y^D\}}$ by $y^* \approx y^D$, formally as $\rho(y^*, y^D) < \epsilon$. This leads to bias, cf. kernel smoothing, but higher acceptance rates, Tavaré et al. [1997]
- ▶ Replace $y \approx y^*$ by $S(y) \approx S(y^*)$ where $S()$ are summary statistics, thus $\rho(S(y^*), S(y^D)) < \epsilon$. This solves the curse of dimensionality problem, Pritchard et al. [1999]

Case: ABC for Gaussian data. Conclusion: $\epsilon \rightarrow 0$ yields $p(\theta|y^D)$, $\epsilon \rightarrow \infty$ yields $p(\theta)$.

Matlab code ABCtest

Practical considerations, continuous data

That algorithm is very inefficient!

- ▶ Replace $\mathbf{1}_{\{y^*=y^D\}}$ by $y^* \approx y^D$, formally as $\rho(y^*, y^D) < \epsilon$. This leads to bias, cf. kernel smoothing, but higher acceptance rates, Tavaré et al. [1997]
- ▶ Replace $y \approx y^*$ by $S(y) \approx S(y^*)$ where $S()$ are summary statistics, thus $\rho(S(y^*), S(y^D)) < \epsilon$. This solves the curse of dimensionality problem, Pritchard et al. [1999]

Case: ABC for Gaussian data. Conclusion: $\epsilon \rightarrow 0$ yields $p(\theta|y^D)$, $\epsilon \rightarrow \infty$ yields $p(\theta)$.

Matlab code ABCtest

Even better, ABC-MCMC

Integrate the ABC framework in an MCMC algorithm. This leads to so called *likelihood free methods*, Marjoram et al. [2003].

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

$$J_\epsilon(y^D, y^*) \propto e^{-\frac{1}{2\epsilon} \|S(y^*) - S(y^D)\|_2^2}$$

- ▶ "Recall" the joint distribution $f(\theta, y)$

$$f(\theta, y) \propto p(y|\theta)p(\theta)J_\epsilon(y^D, y)$$

- ▶ Parameters and states (θ', y') are drawn from

$$p(y'|\theta')g(\theta', \theta_{n-1}) \tag{46}$$

Even better, ABC-MCMC

Integrate the ABC framework in an MCMC algorithm. This leads to so called *likelihood free methods*, Marjoram et al. [2003].

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

$$J_\epsilon(y^D, y^*) \propto e^{-\frac{1}{2\epsilon} \|S(y^*) - S(y^D)\|_2^2}$$

- ▶ "Recall" the joint distribution $f(\theta, y)$

$$f(\theta, y) \propto p(y|\theta)p(\theta)J_\epsilon(y^D, y)$$

- ▶ Parameters and states (θ', y') are drawn from

$$p(y'|\theta')g(\theta', \theta_{n-1}) \tag{46}$$

Even better, ABC-MCMC

Integrate the ABC framework in an MCMC algorithm. This leads to so called *likelihood free methods*, Marjoram et al. [2003].

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

$$J_\epsilon(y^D, y^*) \propto e^{-\frac{1}{2\epsilon} \|S(y^*) - S(y^D)\|_2^2}$$

- ▶ "Recall" the joint distribution $f(\theta, y)$

$$f(\theta, y) \propto p(y|\theta)p(\theta)J_\epsilon(y^D, y)$$

- ▶ Parameters and states (θ', y') are drawn from

$$p(y'|\theta')g(\theta', \theta_{n-1}) \tag{46}$$

- The acceptance probability is the given by ($u = \{y, \theta\}$)

$$\begin{aligned}\alpha &= \min \left(1, \frac{p(u')}{p(u_{n-1})} \frac{g(u_{n-1}|u')}{g(u'|u_{n-1})} \right) \\ &= \min \left(1, \frac{p(y'|\theta')p(\theta')J_{\epsilon}(y^D, y')}{p(y_{n-1}|\theta_{n-1})p(\theta_{n-1})J_{\epsilon}(y^D, y_{n-1})} \frac{p(y_{n-1}|\theta_{n-1})g(\theta_{n-1}, \theta')}{p(y'|\theta')g(\theta', \theta_{n-1})} \right)\end{aligned}$$

- Then acceptance probability is then given by

$$\alpha = \min \left(1, \frac{p(\theta')J_{\epsilon}(y^D, y')}{p(\theta_{n-1})J_{\epsilon}(y^D, y_{n-1})} \frac{g(\theta_{n-1}, \theta')}{g(\theta', \theta_{n-1})} \right) \quad (47)$$

for some data y' and parameters θ' . *Note that the likelihood terms $p(y|\theta)$ cancels out between the proposal and the augmented posterior!*

- The acceptance probability is the given by ($u = \{y, \theta\}$)

$$\begin{aligned}\alpha &= \min \left(1, \frac{p(u')}{p(u_{n-1})} \frac{g(u_{n-1}|u')}{g(u'|u_{n-1})} \right) \\ &= \min \left(1, \frac{p(y'| \theta') p(\theta') J_{\epsilon}(y^D, y')}{p(y_{n-1} | \theta_{n-1}) p(\theta_{n-1}) J_{\epsilon}(y^D, y_{n-1})} \frac{p(y_{n-1} | \theta_{n-1}) g(\theta_{n-1}, \theta')}{p(y' | \theta') g(\theta', \theta_{n-1})} \right)\end{aligned}$$

- Then acceptance probability is then given by

$$\alpha = \min \left(1, \frac{p(\theta') J_{\epsilon}(y^D, y')}{p(\theta_{n-1}) J_{\epsilon}(y^D, y_{n-1})} \frac{g(\theta_{n-1}, \theta')}{g(\theta', \theta_{n-1})} \right) \quad (47)$$

for some data y' and parameters θ' . *Note that the likelihood terms $p(y|\theta)$ cancels out between the proposal and the augmented posterior!*

ABC-MCMC discussion

- ▶ Note that the ABC-MCMC still targets the joint (θ, y) distribution, regardless of the proposal distribution.
- ▶ The "naive choice" of the proposal density cancelled out all likelihood terms - *important when the likelihood cannot be found (easily)*
- ▶ Further simplifications are obtained for symmetric proposals, $g(|)$ for the parameters, e.g. for random walk dynamics.
- ▶ Thus, the ABC-MCMC algorithm is surprisingly simple to implement!
 - ▶ Simulate a parameter candidate
 - ▶ Simulate a data candidate
 - ▶ accept or reject the candidates

ABC-MCMC discussion

- ▶ Note that the ABC-MCMC still targets the joint (θ, y) distribution, regardless of the proposal distribution.
- ▶ The "naive choice" of the proposal density cancelled out all likelihood terms - *important when the likelihood cannot be found (easily)*
- ▶ Further simplifications are obtained for symmetric proposals, $g(|)$ for the parameters, e.g. for random walk dynamics.
- ▶ Thus, the ABC-MCMC algorithm is surprisingly simple to implement!
 - ▶ Simulate a parameter candidate
 - ▶ Simulate a data candidate
 - ▶ accept or reject the candidates

ABC-MCMC discussion

- ▶ Note that the ABC-MCMC still targets the joint (θ, y) distribution, regardless of the proposal distribution.
- ▶ The "naive choice" of the proposal density cancelled out all likelihood terms - *important when the likelihood cannot be found (easily)*
- ▶ Further simplifications are obtained for symmetric proposals, $g(|)$ for the parameters, e.g. for random walk dynamics.
- ▶ Thus, the ABC-MCMC algorithm is surprisingly simple to implement!
 - ▶ Simulate a parameter candidate
 - ▶ Simulate a data candidate
 - ▶ accept or reject the candidates

ABC-MCMC discussion

- ▶ Note that the ABC-MCMC still targets the joint (θ, y) distribution, regardless of the proposal distribution.
- ▶ The "naive choice" of the proposal density cancelled out all likelihood terms - *important when the likelihood cannot be found (easily)*
- ▶ Further simplifications are obtained for symmetric proposals, $g(|)$ for the parameters, e.g. for random walk dynamics.
- ▶ Thus, the ABC-MCMC algorithm is surprisingly simple to implement!
 - ▶ Simulate a parameter candidate
 - ▶ Simulate a data candidate
 - ▶ accept or reject the candidates

References I

- Christophe Andrieu, Gareth O Roberts, et al. The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.
- Peter J Diggle and Richard J Gratton. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):193–212, 1984.
- Erik Lindström. Semiparametric lag dependent functions. *Applied Mathematical Sciences*, 7(12):551–566, 2013.
- Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

References II

- Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.
- Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Feedback

Send feedback, questions and information about typos to
`erik.lindstrom@matstat.lu.se`.