ESTIMATION ALGORITHMS

- **Solving normal equations using QR-factorization**
- **Non-linear optimization**
- **Two and multi-stage methods**
- **EM algorithm**

SOLVING NORMAL EQUATIONS USING QR FACTORIZATION

Least squares solution: $\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T Y$

Problems with direct computation of $\hat{\theta}$:

1. *Numerical condition*: $\text{cond}(\Phi^T \Phi) = \dfrac{\sigma_{\max}(\Phi^T \Phi)}{\sigma_{\min}(\Phi^T \Phi)} = \left[\dfrac{\sigma_{\max}(\Phi)}{\sigma_{\min}(\Phi)}\right]^2 = \text{cond}^2(\Phi)$

2. *No. of computations*: computing a matrix inverse takes $O(n^3)$ operations!

Idea: Factorize $\Phi = QR$, with Q orthogonal and R square upper triangular:

$$\Phi^T \Phi \hat{\theta} = \Phi^T Y \quad \Leftrightarrow \quad R^T R \hat{\theta} = R^T Q^T Y \quad \Leftrightarrow \quad R\hat{\theta} = Q^T Y$$

This can be solved in $O(n^2)$ steps, and $\text{cond}(R) = \text{cond}(\Phi) \ll \text{cond}(\Phi^T \Phi)$!

QR FACTORIZATION

Partition the matrices as:

$$\Phi = QR$$

$$[\varphi_1 \quad \Phi_2] = [q_1 \quad Q_2]\begin{bmatrix} r_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}, \quad q_1 \text{ column vector}, \quad r_{11} \in \mathbb{R}, \quad R_{22} \text{ upper triangular}$$

- q_1, Q_2 must satisfy:

$$\begin{bmatrix} q_1^T \\ Q_2^T \end{bmatrix}[q_1 \quad Q_2] = \begin{bmatrix} q_1^T q_1 & q_1^T Q_2 \\ Q_2^T q_1 & Q_2^T Q_2 \end{bmatrix} \overset{!}{=} \begin{bmatrix} 1 & 0 \\ 0 & I \end{bmatrix}$$

These quantities can be computed recursively...

QR FACTORIZATION (CONT.)

Recursive method:

$$[\varphi_1 \quad \Phi_2] = [q_1 \quad Q_2] \begin{bmatrix} r_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} = [q_1 r_{11} \quad q_1 R_{12} + Q_2 R_{22}]$$

1. Compute: $r_{11} = \|\varphi_1\|, q_1 = \varphi_1 / r_{11}$
2. Since $q_1^T \Phi_2 = q_1^T (q_1 R_{12} + Q_2 R_{22}) = R_{12}$, we have: $R_{12} = q_1^T \Phi_2$
3. As $\Phi_2 = q_1 R_{12} + Q_2 R_{22}$, Q_2 and R_{22} can be computed from the QR factorization of

$$\Phi_2 - q_1 R_{12}$$

No. of computations: $O(Nn^2)$

Alternatives: Givens rotation matrices and Householder (orthogonal) transformations

NON-LINEAR OPTIMIZATION

Quadratic PEM: $$V_N(\theta) = \frac{1}{2N}\sum_{t=1}^{N}\varepsilon_t^2(\theta)$$

Gradient: $$V_N'(\theta) = \frac{1}{N}\sum_{t=1}^{N}\psi_t(\theta)\varepsilon_t(\theta), \quad \psi_t(\theta) = \frac{\partial \varepsilon_t(\theta)}{\partial \theta}$$

Hessian: $$V_N''(\theta) = \frac{1}{N}\sum_{t=1}^{N}\psi_t(\theta)\psi_t^T(\theta) + \frac{1}{N}\sum_{t=1}^{N}\psi_t'(\theta)\varepsilon_t(\theta)$$

Minimization method: $\hat{\theta}^{(i+1)} = \hat{\theta}^{(i)} - \mu^{(i)}[R^{(i)}]^{-1}V_N'(\hat{\theta}^{(i)})$

E.g.: $R^{(i)} = V_N''(\hat{\theta}^{(i)})$ (Newton method)

$R^{(i)} = \frac{1}{N}\sum_{t=1}^{N}\psi_t(\theta)\psi_t^T(\theta)$ (Gauss-Newton method)

The step size $\mu^{(i)}$ can be chosen as 1, or adjusted in each step (*damped Gauss-Newton*)

TWO AND MULTI-STAGE METHODS

Two-step method:

1. Obtain an initial $\sqrt{N}$-consistent estimator $\hat{\theta}_{init}$ (e.g., instrumental variables, subspace, ...)

2. Compute

$$\hat{\theta} = \hat{\theta}_{init} - [l''(\hat{\theta}_{init})]^{-1} l'(\hat{\theta}_{init})$$

An asymptotically efficient estimator without local optima problems!

TWO AND MULTI-STAGE METHODS (CONT.)

Fact. Let $\hat{\theta}_{init} = \theta_0 + O_p(N^{-1/2})$. Then, one Newton iteration of $l(\theta)$, starting from $\hat{\theta}_{init}$, gives an asymptotically efficient estimator:

$$\boxed{\hat{\theta} = \hat{\theta}_{init} - [l''(\hat{\theta}_{init})]^{-1} l'(\hat{\theta}_{init})} \quad \Rightarrow \quad \sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_F^{-1}(\theta_0))$$

Proof. Since $N^{-1/2} l'(\theta_0) \xrightarrow{d} \mathcal{N}(0, I_F(\theta_0))$ and $N^{-1} l''(\theta_0) \xrightarrow{p} -I_F(\theta_0)$:

$$\begin{aligned}
&\sqrt{N}(\hat{\theta} - \theta_0) \\
&= \sqrt{N}(\hat{\theta}_{init} - \theta_0) - \sqrt{N}[l''(\hat{\theta}_{init})]^{-1} l'(\hat{\theta}_{init}) \\
&= \sqrt{N}(\hat{\theta}_{init} - \theta_0) - N^{-1/2}[-I_F(\theta_0) + o_p(1)]^{-1}\{l'(\theta_0) + l''(\theta_0)[\hat{\theta}_{init} - \theta_0] + o_p(1)\} \\
&= \sqrt{N}(\hat{\theta}_{init} - \theta_0) - [-I_F(\theta_0) + o_p(1)]^{-1}\{N^{-1/2} l'(\theta_0) + \sqrt{N}[-I_F(\theta_0) + o_p(1)][\hat{\theta}_{init} - \theta_0] + o_p(1)\} \\
&= \sqrt{N}(\hat{\theta}_{init} - \theta_0) + I_F^{-1}(\theta_0)\{N^{-1/2} l'(\theta_0) - I_F(\theta_0)\sqrt{N}[\hat{\theta}_{init} - \theta_0]\} + o_p(1) \\
&= I_F^{-1}(\theta_0) N^{-1/2} l'(\theta_0) + o_p(1) \xrightarrow{d} \mathcal{N}(0, I_F^{-1}(\theta_0))
\end{aligned}$$

TWO AND MULTI-STAGE METHODS (CONT.)

Steiglitz-McBride method:

$$\text{OE system:} \qquad y_t = \frac{B(q)}{A(q)} u_t + e_t$$

1. Set $i \leftarrow 1$ and $F^{(1)}(q) \leftarrow 1$
2. Set $u_t^{(i)} \leftarrow F^{(i)}(q) u_t$, $y_t^{(i)} \leftarrow F^{(i)}(q) y_t$, and fit the model: $A(q) y_t^{(i)} = B(q) u_t^{(i)} + e_t$
3. Set $F^{(i+1)}(q) \leftarrow \hat{A}(q)^{-1}$ and $i \leftarrow i+1$. Go back to Step 2

- This method does *not* always converge (sometimes $\hat{A}(q)$ gets unstable!), and even if it does, the limit estimates are *not* equal to PEM (but typically are close)

- An high-order ARX pre-filtering step can be added, to handle colored noise With this step, Steiglitz-McBride is asymptotically efficient

- In frequency domain, this method is called "Sanathanan-Koerner"

IV4:

Consider the model $y_t = G(q;\theta)u_t + v_t$, where $G(q;\theta) = \dfrac{B(q;\theta)}{A(q;\theta)}$. The system can be identified using an extended instrumental variables approach:

$$\frac{1}{N}\sum_{t=1}^{N}\zeta_t L(q)[y_t - \varphi_t^T\hat{\theta}_{IV}] = 0$$

It can be shown (Ljung, pp. 486) that the C-R bound is achieved (assuming $H_0(q)$ is $AR(n_b)$) if

$$\zeta_t^T = \frac{1}{H_0(q)A_0(q)}[-G_0(q)u_{t-1} \quad \cdots \quad -G_0(q)u_{t-n_a} \quad u_{t-1} \quad \cdots \quad u_{t-n_b}]; \quad L(q) = \frac{1}{H_0(q)A_0(q)}$$

However, $G_0(q)$, $H_0(q)$ and $A_0(q)$ are *unknown*!

TWO AND MULTI-STAGE METHODS (CONT.)

IV4 method:

1. Write the model as $\hat{y}_{t|t-1} = \varphi_t^T \theta$ and estimate θ via LS $\Rightarrow$ $\hat{\theta}^{(1)}$

2. Generate:

$$\zeta_t^{(1)T} = [-x_{t-1}^{(1)} \quad \cdots \quad -x_{t-n_a}^{(1)} \quad u_{t-1} \quad \cdots \quad u_{t-n_b}], \quad x_t^{(1)} := G(q;\hat{\theta}^{(1)})u_t$$

and estimate θ via basic IV using $\zeta_t^{(1)}$ as instrument $\Rightarrow$ $\hat{\theta}^{(2)}$

3. Let $\hat{w}_t^{(2)} = A(q;\hat{\theta}^{(2)})y_t - B(q;\hat{\theta}^{(2)})u_t$, and postulate an AR model of order $n_a + n_b$:

$$L(q)\hat{w}_t^{(2)} = e_t$$

Estimate $L(q)$ via LS $\Rightarrow$ $\hat{L}(q)$

4. Let $\zeta_t^{(2)T} = \hat{L}(q)[-x_{t-1}^{(1)} \quad \cdots \quad -x_{t-n_a}^{(1)} \quad u_{t-1} \quad \cdots \quad u_{t-n_b}], \quad x_t^{(2)} := G(q;\hat{\theta}^{(2)})u_t$

and estimate θ via extended IV using $\zeta_t^{(2)T}$ and $\hat{L}(q)$

EXPECTATION MAXIMIZATION ALGORITHM

Let $X \sim p_\theta$ be observed data. Sometimes computing $\hat{\theta}_{ML}$ is very difficult, but the task can be simplified if some *hidden*/*latent* variables Z were known: e.g., the state vector in a state space model

$$
\begin{aligned}
l(\theta) &= \ln p(X;\theta) \\
&= \ln p(X,Z;\theta) - \ln p(Z \mid X;\theta), \qquad \text{for any } Z \\
&= \underbrace{\int \ln p(X,Z;\theta)\, p(Z \mid X;\theta^{(i)})\, dZ}_{Q(\theta|\theta^{(i)})} - \underbrace{\int \ln p(Z \mid X;\theta)\, p(Z \mid X;\theta^{(i)})\, dZ}_{-H(\theta|\theta^{(i)})}
\end{aligned}
$$

Idea: Instead of maximizing $l(\theta)$, let's maximize $Q(\theta \mid \theta^{(i)})$!

EM algorithm:

1. *E-step*: Compute $Q(\theta \mid \theta^{(i)}) = \int \ln p(X,Z;\theta)\, p(Z \mid X;\theta^{(i)})\, dZ$
2. *M-step*: Maximize $Q(\theta \mid \theta^{(i)})$ with respect to θ, and denote $\theta^{(i+1)}$ the maximizer

EXPECTATION MAXIMIZATION ALGORITHM (CONT.)

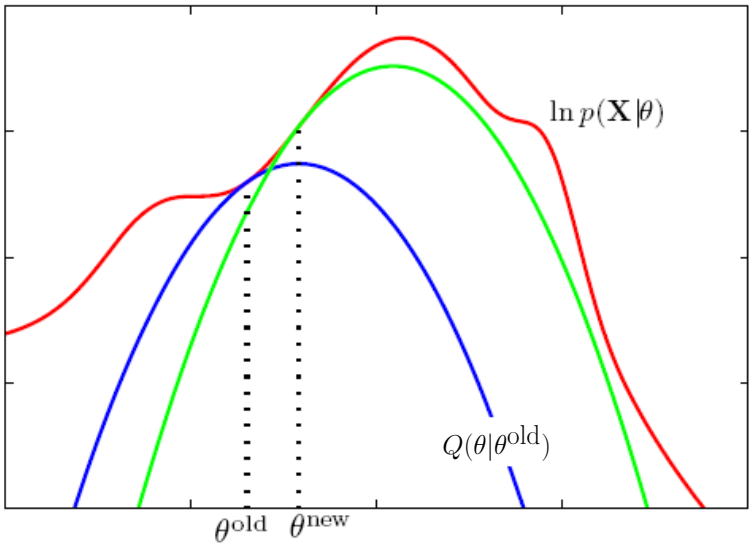
Why would it work?

$$H(\theta \mid \theta^{(i)}) - H(\theta^{(i)} \mid \theta^{(i)}) = -\int \ln\left[\frac{p(Z \mid X;\theta)}{p(Z \mid X;\theta^{(i)})}\right] p(Z \mid X;\theta^{(i)})\, dZ$$

$$\geq -\ln \int \frac{p(Z \mid X;\theta)}{p(Z \mid X;\theta^{(i)})}\, p(Z \mid X;\theta^{(i)})\, dZ \quad \text{(Jensen's inequality)}$$

$$= 0$$

Therefore, $Q(\theta \mid \theta^{(i)})$ is a *lower bound* of $l(\theta)$! (modulo $H(\theta^{(i)} \mid \theta^{(i)})$)



EXPECTATION MAXIMIZATION ALGORITHM (CONT.)

Applications

1. Gaussian mixtures / clustering (see Wikipedia entry on “EM algorithm”)

2. Missing data (Goodwin&Feuer, 1998)

3. ML estimation of state space models (Shumway & Stoffer, 1982)

4. Identification of nonlinear state space models (Schön, Wills & Ninness, 2011)

EXPECTATION MAXIMIZATION ALGORITHM (CONT.)

Example (courtesy of T. Schön)

$$x_{t+1} = \theta x_t + v_t$$
$$y_t = 0.5x_t + e_t, \qquad [v_t \; e_t]^T \sim \mathcal{N}(0, 0.1I)$$

Consider data $X := \{y_1, \ldots, y_N\}$, and $Z := \{x_1, \ldots, x_{N+1}\}$ as hidden variables. Then,

$$p(X, Z) = p_\theta(x_1) \prod_{t=1}^{N} p_\theta(x_{t+1}, y_t \mid x_t)$$

where

$$[x_{t+1}^T \quad y_t]^T \mid x_t \sim \mathcal{N}\left(\begin{bmatrix} \theta \\ 0.5 \end{bmatrix} x_{t+1}, 0.1I \right)$$

EXPECTATION MAXIMIZATION ALGORITHM (CONT.)

Example (cont.)

The Q function is

$$Q(\theta \mid \theta^{(i)}) = \int \ln p(X,Z;\theta)\, p(Z \mid X;\theta^{(i)})\, dZ$$

$$= \int \ln p_\theta(x_1)\, p(Z \mid X;\theta^{(i)})\, dZ + \sum_{t=1}^{N} \int \ln p_\theta(x_{t+1}, y_t \mid x_t)\, p(Z \mid X;\theta^{(i)})\, dZ$$

$$= \text{const} - \frac{1}{0.2} E\{x_1^2 \mid X;\theta^{(i)}\} - \frac{1}{0.2} \sum_{t=1}^{N} E\{(x_{t+1} - \theta x_t)^2 + (y_t - 0.5x_t)^2 \mid X;\theta^{(i)}\}$$

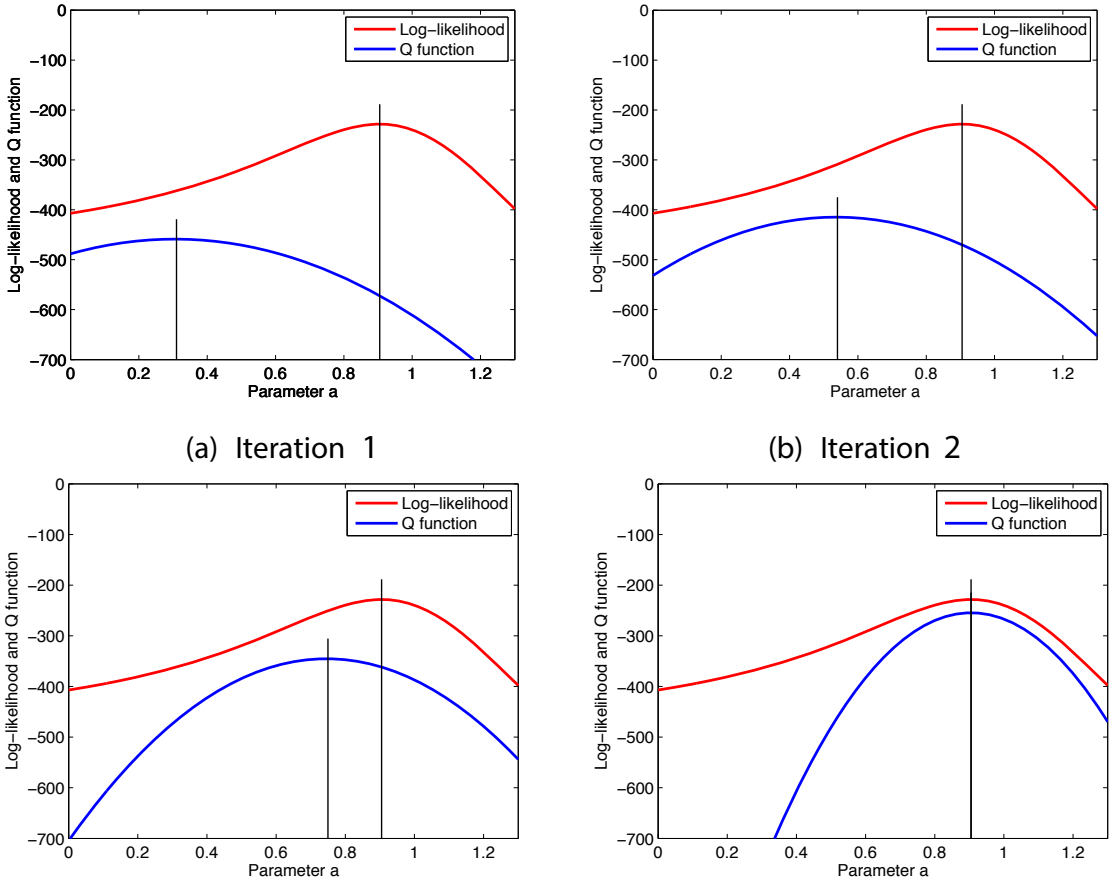
$$= \text{const} - 5\theta^2 \underbrace{\sum_{t=1}^{N} E\{x_t^2 \mid X;\theta^{(i)}\}}_{\varphi} + 10\theta \underbrace{\sum_{t=1}^{N} E\{x_t x_{t+1} \mid X;\theta^{(i)}\}}_{\psi}$$

φ and ψ can be computed using Kalman smoothing theory, giving

$$\theta^{(i+1)} = \arg\max_\theta Q(\theta \mid \theta^{(i)}) = \psi / \varphi$$

EXPECTATION MAXIMIZATION ALGORITHM (CONT.)

Example (cont.)



(a) Iteration 1

(b) Iteration 2

(c) Iteration 3

(d) Iteration 11