Minimum Message Length Order Selection and Parameter Estimation of Moving Average Models

Daniel F. Schmidt

Centre for MEGA Epidemiology
The University of Melbourne

Solomonoff 85th Memorial Conference, 2011

Outline

Outline

Moving Average Models (1)

- We have observed a time series $\mathbf{y}$ of n datapoints

$$\mathbf{y} = (y_1, \ldots, y_n)'$$

- A q-th order moving average model, MA(q) of $\mathbf{y}$

$$y_t = \sum_{j=1}^{q} \eta_j v_{t-j} + v_t$$

where

$$v_t \sim N(0, \tau)$$

and $\boldsymbol{\eta}_q = (\eta_1, \ldots, \eta_q)$ are the moving average coefficients
$\Rightarrow$ Models data as sum of unobserved random variables

Moving Average Models (2)

- Problem Usually only observe the data
  $\Rightarrow$ Need to estimate order q and parameters $\boldsymbol{\theta}_q = (\boldsymbol{\eta}_q, \tau)$
- Commonly done with information criteria. Let ...
  - $p(\mathbf{y}|\boldsymbol{\theta}_q)$ denote the likelihood function
  - $\hat{\boldsymbol{\theta}}_q$ denote the maximum likelihood estimates
- Estimate q by solving

$$\hat{q} = \underset{q \in \{0,\ldots,Q\}}{\arg\min} \left\{ -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_q) + \alpha_q \right\}$$

- α_q is a complexity penalty, e.g.,
  - $\alpha_q = q$ (Akaike Information Criterion)
  - $\alpha_q = (q/2)\log n$ (Bayesian Information Criterion)

Outline

Minimum Message Length (1)

- Develop a Minimum Message Length approach to solve the problem
- Practical implementation of theory of inductive inference inspired by Solomonoff's work
  - Model that yields the briefest encodingof data in a hypothetical message is optimal
- The message is composed of two-parts
  - *assertion*, statement describing a particular model $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$
  - *detail*, encoding of the data $\mathbf{y}$ using the assertion model $\boldsymbol{\theta}$

Minimum Message Length (2)

- The total length of the two-part message, $I(\boldsymbol{\theta}, \mathbf{y})$, is sum of the lengths of the assertion and the detail

$$I(\boldsymbol{\theta}, \mathbf{y}) = I(\boldsymbol{\theta}) + I(\mathbf{y}|\boldsymbol{\theta})$$

- MML advocates choosing model $\boldsymbol{\theta}$ that minimises the codelength of the hypothetical two-part message

MML87 (1)

- The Wallace–Freeman, or MML87 codelength, for model $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ and data $\mathbf{y}$ is

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}) = \underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2} \log |\mathbf{J}(\boldsymbol{\theta})| + \frac{k}{2} \log \kappa_k}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{k}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{I_{87}(\mathbf{y}|\boldsymbol{\theta})}$$

  - $p(\mathbf{y}|\boldsymbol{\theta})$ denotes the likelihood function
  - $\pi(\cdot)$ is a prior distribution over the parameter space $\Theta \subset \mathbb{R}^k$
  - $\mathbf{J}(\boldsymbol{\theta})$ is the Fisher information matrix
  - κ_k is the normalised second moment of an optimal quantising lattice in k-dimensions

Message Lengths of MA Models (1)

- Need to specify prior distributions

$$\pi(\boldsymbol{\eta}_q, \tau, q) = \pi(\boldsymbol{\eta}_q)\pi(\tau)\pi(q)$$

- We choose

$$\begin{aligned} \pi(\boldsymbol{\eta}_q) &= \frac{1}{\mathrm{vol}(\Lambda_q)}, \quad \boldsymbol{\eta}_q \in \Lambda_q \\ \pi(\tau) &\propto \frac{1}{\tau}, \quad \tau \in (\tau_0, \tau_1) \\ \pi(q) &\propto 1, \quad q \in \{0, \ldots, Q\} \end{aligned}$$

where Λ_q is the invertibility region of an MA(q) model

Message Lengths of MA Models (2)

- Exact Fisher information matrx

$$\mathbf{J}_n(\boldsymbol{\theta}_q^*) = -\mathrm{E}_{\boldsymbol{\theta}_q^*}\left[\left.\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_q^*}\right]$$

- Moving average exact Fisher violates MML87 assumptions $\Rightarrow |\mathbf{J}_n(\boldsymbol{\theta}_q^*)| \to 0$ as $\boldsymbol{\theta}_q^*$ approaches boundary of Λ_q
- We use asymptotic Fisher information matrix

$$\mathbf{J}(\boldsymbol{\theta}_q^*) = \lim_{n\to\infty}\left\{\frac{\mathbf{J}_n(\boldsymbol{\theta}_q^*)}{n}\right\}$$

Message Lengths of MA Models (3)

- The invertibility region Λ_q is complex for $q > 2$
- Partial autocorrelations $\boldsymbol{\rho}(\boldsymbol{\eta}_q) : (-1, 1)^q \to \Lambda_q$
  $\Rightarrow$ simpler invertibility region
- Perform numerical minimisation using partial autocorrelations
- Yields simple expression for asymptotic Fisher information

$$|\mathbf{J}(\boldsymbol{\eta}_q, \tau)| = \left(\frac{n^{q+1}}{2\tau^2}\right) \left(\prod_{j=1}^{q} \frac{1}{(1 - \rho_j^2(\boldsymbol{\eta}_q))^j}\right)$$

Message Lengths of MA Models (4)

- Find estimates $(\hat{q}^{87}, \hat{\boldsymbol{\eta}}_q^{87}, \hat{\tau}^{87})$ that minimises

$$\begin{aligned} I(\mathbf{y}, \boldsymbol{\theta}_q, q) &= -\log p_q(\mathbf{y}|\boldsymbol{\eta}_q, \tau) - \frac{1}{2}\sum_{j=1}^{q} j \log(1-\rho_j^2(\boldsymbol{\eta}_q)) \\ &\quad + \frac{q}{2}\log n + \log \mathrm{vol}(\Lambda_q) + c(q+1) + \mathrm{const} \end{aligned}$$

  where $c(q+1)$ are terms dependent only on q
- Message length is clearly BIC plus
  - term dependent on coefficients, $\boldsymbol{\eta}_q$, and
  - terms dependent on the order, q
- Coefficient dependent term acts as regularisation

Properties of MML87 estimates

- Let q^* be the "true" order of the MA model that generated $\mathbf{y}$
- We can rewrite message length as

$$-\log p_q(\mathbf{y}|\boldsymbol{\eta}_q, \tau) + \frac{q}{2}\log n + O(1)$$

- Then ...
  - $\hat{q}^{87}$ is consistent if $Q \geq q^*$ does not grow with n
  - $(\hat{\boldsymbol{\eta}}_q^{87}, \hat{\tau}^{87})$ are consistent when $q \geq q^*$
- Easy to show that

$$||\boldsymbol{\rho}(\hat{\boldsymbol{\eta}}_q^{87})||_\infty < 1$$

$\Rightarrow$ MML87 estimates do not suffer from the "pile-up" effect

Discussion (1)

- Performance of MML87 was examined through simulation
- Parameter estimation
  - Compared to maximum likelihood and modified Durbin algorithm
  - MML87 performed well
    $\Rightarrow$ less tendency to overestimate effect strengths
- Order selection
  - Compared to AIC, BIC, KIC
  - MML87 performed well for prediction
  - Not always best at order selection
- Details in paper

Discussion (2)

- Some possible future work ...
  - Explore robustness to prior assumptions
  - Extend to autoregressive moving average (ARMA) models
  - Extend to regression with (AR)MA noise models