Model robust inference and model selection issues for stationary Gaussian time series

Gudmund Hermansen
based on joint work with Nils Lid Hjort

Department of Mathematics at University of Oslo and (sfi)^2

August 19, 2010

Outline

Introduction and summary: Density estimation for iid. observations

Given iid. observations where $Y_i \sim h^\circ$. We want to estimate the density and try to fit members of the class $\mathcal{H}_\theta = \{h_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$, where h° might not be in $\mathcal{H}_\theta$, see Claeskens & Hjort (2008).

- Under reasonable conditions the ML-estimate $\hat{\theta}_n$ converges at rate *root-n* to the *least false parameter* value θ_0, i.e.

$$\hat{\theta}_n = \arg\max_\theta \ell_n(h_\theta) \to_{P_{h^\circ}} \arg\min_\theta KL(h^\circ, h_\theta) = \theta_0.$$

- The Kullback-Leibler distance between the true and estimated model is given by

$$\begin{aligned} \mathrm{KL}(h^\circ, h_{\hat{\theta}_n}) &= \int h^\circ(\omega) \log \frac{h^\circ(\omega)}{h_{\hat{\theta}_n}(\omega)} \, d\omega \\ &= \int h^\circ(\omega) \log h^\circ(\omega) \, d\omega - \int h^\circ(\omega) \log h_{\hat{\theta}_n}(\omega) \, d\omega. \end{aligned}$$

- In addition one can prove that

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to_d N_p(0, J^{-1} K J^{-1})$$

where J and K will be defined later. If the model is correctly specified $J = K$.

Introduction and summary: Density estimation for iid. observations

- The Akaike's Information Criterion (AIC) of a candidate model h_θ is defied by

$$\mathrm{AIC}(h_\theta) = 2\ell_{n,\max}(h_\theta) - 2\dim(\theta) = 2\ell_n(h_{\hat{\theta}_n}) - 2p_\theta.$$

- The AIC formula is related to the K-L distance. In the sense that it can be viewed as a scaled first order bias corrected versions of the naive estimate of the model specific part, i.e. we wish to estimate

$$\int h^\circ(\omega) \log h_{\hat{\theta}_n}(\omega)\, d\omega \quad \text{by} \quad n^{-1} \sum_{i=1}^{n} \log h_{\hat{\theta}_n}(Y_i).$$

- By choosing the model with the highest AIC score we are choosing the model with minimal estimated KL-distance.
- There is a model robust version of AIC, known as AIC* or TIC, given by

$$\mathrm{AIC}^*(h_\theta) = 2\ell_{n,\max}(h_\theta) - 2\,\mathrm{tr}(\hat{J}_n^{-1}\hat{K}_n),$$

and again if $J = K$ we get the usual AIC formula since $\mathrm{tr}(J^{-1}K) = \mathrm{tr}(I_{p_\theta}) = p_\theta$.

Introduction and summary: Time series models

- There exists several types of model selection criteria for stationary time series. They all aims at choosing the *best* model among a set of candidate model.
- Note that *best* does not have the same interpretation for the different criteria.
- Examples of such are AIC AIC$_c$, FPE, C$_p$, HQ, etc., see McQuarrie and Tsai (1998).
- Some of these, like the AIC, aims at estimating a distance between the assumed true model and a given candidate model.
- In the literature these criteria are often studied and justified under the assumption that the true model is within the same family (such as AR) as the candidate models.
- The goal is to derive a version of AIC* within large class of models without assuming that the true model necessarily is included among the candidate models.

Introduction and summary: Autoregressive models

- As an example, consider a sample from a standard AR(p_0) with unknown order p_0. The object is to choose the "optimal" order for the estimated AR(p) model.
- By assuming that the true model is also autoregressive, conditioning on the first p observations and under the assumption that $p > p_0$ a corrected version of AIC is given by

$$\mathrm{AIC}_c(p) = -m \log \hat{\sigma}_p^2 - m(m+p)/(m-p-2)$$

  where $\hat{\sigma}_p = (m^{-1} \sum_j (y_j - \hat{y}_j)^2)^{1/2}$ and $m = n - p$, see Clifford & Tsai (1989) or Claeskens & Hjort (2008).
- Another choice is to use $\mathrm{AIC}_u(p)$ where $\hat{\sigma}_p$ is replaced by the unbiased estimate $\hat{s}_p$, see McQuarrie & Tsai (1998).

Derivation of AIC*: Conditions

Consider a smooth regular parametric model with ML-estimate $\hat{\theta}_n$ that converges at *root-n* rate to the least false parameter value θ_0.

i) Suppose we have the process convergence

$$H_n(s) = \ell_n(\theta_0 + sn^{-1/2}) - \ell_n(\theta_0) \to_d H(s) = s^t U - 2^{-1} s^t J s,$$

where $U \sim N_p(0, K)$.

- For the classical model we have $H_n(s) = sU_n - 2^{-1} s^t J_n s + o_P(1)$, where

$$U_n = n^{-1/2} \ell_n'(\theta_0) \to_d N_p(0, K) \text{ and } J_n = -n^{-1} \ell_n''(\theta_0) \to_P J.$$

ii) Under some extra conditions we have that

- $Z_n = \arg\max(H_n) = n^{1/2}(\hat{\theta}_n - \theta_0) \to_d Z = \arg\max(H) = J^{-1}U.$
- $\max H_n = H_n(Z_n) = \ell_n(\hat{\theta}_n) - \ell_n(\theta_0) \to_d \max H = H(Z) = 2^{-1} U^t J^{-1} U.$

Derivation of AIC*

The goal is to estimate the model specific part of the KL-distance, $A(\hat{\theta}_n) = E_{h^\circ} \log h(Y, \hat{\theta}_n)$, since

$$A_n(\theta) = n^{-1}\ell_n(\theta) \to_{P_{h^\circ}} A(\theta) = E_{h^\circ} \log h(Y, \theta)$$

we can start with $A_n(\hat{\theta}_n)$ as an estimate for $A(\hat{\theta}_n)$.

- This estimator overshoots its target. By a Taylor expansion we obtain in the classical case

  $$A(\hat{\theta}_n) \doteq A(\theta_0) - 2^{-1}(\hat{\theta}_n - \theta_0)^t J(\hat{\theta}_n - \theta_0) = A(\theta_0) - 2^{-1}n^{-1}Z_n^t J Z_n$$

  and secondly

  $$A_n(\hat{\theta}_n) = n^{-1}\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} + A_n(\theta_0) = n^{-1}H_n(Z_n) + A_n(\theta_0).$$

- Then

  $$A_n(\hat{\theta}_n) - A(\hat{\theta}_n) \doteq \epsilon_n + n^{-1}\{H_n(Z_n) + 2^{-1}Z_n^t J Z_n\} = \epsilon_n + n^{-1}W_n,$$

  where $\epsilon_n = A_n(\theta_0) - A(\theta_0)$ has mean zero and

  $$W_n = H_n(Z_n) + 2^{-1}Z_n^t J Z_n \to_d H(Z) + 2^{-1}Z^t J Z = U^t J^{-1} U.$$

- Now the AIC* is given by

  $$\text{AIC}^* = 2\{\ell_n(\hat{\theta}_n) - \hat{p}^*\}$$

  where $\hat{p}^*$ is a consistent estimate of $p^* = EU^t J^{-1} U = \text{tr}(J^{-1}K)$.

Model and model assumptions

- We consider real, zero mean, stationary Gaussian time series models, $\{Y_t\}$ with dependency structure defined by the spectral density g, then

$$\mathrm{Cov}(Y_t,\, Y_{t+h}) = C_g(h) = 2\int_0^\pi \cos(\omega h) g(\omega)\, d\omega \text{ and}$$

$$g(\omega) = \frac{C_g(0)}{2\pi} + \frac{1}{\pi}\sum_{h=1}^{\infty} \cos(\omega h) C_g(h),$$

  see Priestley (1981) or Brockwell & Davis (1991).

- The true series is assumed to be of short memory, especially that

$$\sum_h |h|^2 |C_g(h)|^2 < \infty, \quad \sum_h |h||C_g(h)| < \infty$$

  and $0 < g(\omega) < \infty$, for $0 \le \omega \le \pi$. Note that the first two conditions are related to the continuity and smoothness of the spectral densities.

Derivation of AIC* for stationary time series

- Our candidate models are given by the set of spectral densities $\mathcal{F}_\theta$, where $\theta \in \Theta \subset \mathbb{R}^p$ and all $f_\theta \in \mathcal{F}_\theta$ are assumed to satisfy:
  - i) there exist $m, M \in \mathbb{R}$ such that $0 < m \leq f_\theta(\omega) \leq M < \infty$
  - ii) and $f_\theta(\omega)$ is smooth and continuous in both ω and θ,

  for $0 \leq \omega \leq \pi$ and $\theta \in \Theta$.
- Note that the true spectral density g is not necessary included in $\mathcal{F}_\theta$.
- The joint Gaussian log-density:

$$\ell_n(f_\theta) = -2^{-1}[n \log 2\pi + \log |\Sigma_{f_\theta}| + y^t \Sigma_{f_\theta}^{-1} y]$$

- The Whittle approximation:

$$\ell_n^w(f_\theta) = -\frac{n}{2}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log(2\pi f_\theta(\omega))\, d\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_n(\omega)}{f_\theta(\omega)}\, d\omega + \log(2\pi)\right]$$

  where $I_n(\omega) = (2\pi n)^{-1} |\sum_j Y_j \exp\{-i\omega j\}|^2$ is the periodogram.
- The connections between l_n and l_n^w are throughly studied under the assumption that the model is correctly specified, see Dzhaparidze (1986) and Davis (1973).

Derivation of AIC* for stationary time series

- In density estimation the KL-distance can be motivated as the limit of the following scaled log-likelihood differences. i.e.

$$n^{-1}[\ell_n(h^\circ) - \ell_n(h_\theta)] \to_{P_{h^\circ}} \mathrm{KL}(h^\circ, h_\theta), \text{ for given } \theta.$$

- Inspired from this we construct the following discrepancy measure

$$\begin{aligned} d_n(g, f_\theta) &= n^{-1}[\ell_n(g) - \ell_n(f_\theta)] \\ &= n^{-1}[\ell_n^w(g) - \ell_n^w(f_\theta)] + n^{-1}\{[\ell_n(g) - \ell_n^w(g)] - [\ell_n(f_\theta) - \ell_n^w(f_\theta)]\} \\ &\to_{P_g} \frac{1}{2\pi}\int_0^\pi \left[-\log(g(\omega)/f_\theta(\omega)) + \left(\frac{g(\omega)}{f_\theta(\omega)} - 1\right)\right] d\omega \\ &= d(g, f_\theta) \end{aligned}$$

  where $d(g, f_\theta) \geq 0$ and $d(g, f_\theta) = 0$ if and only if $g = f_\theta$ almost everywhere. This limit is sometimes referred to as the *asymptotic K-L discrepancy information*.

- This idea was mentioned in Clifford & Tsai (1989) but only used under the assumption that the true model was included in the set of candidate models.

Derivation of AIC* for stationary time series

- There are essential two parallel stories, with two main points each. Let

$$\hat{\theta}_n^w = \arg\max_\theta \ell_n^w(f_\theta) \text{ and } \hat{\theta}_n = \arg\max_\theta \ell_n(f_\theta)$$

  then we would like to show for both $\hat{\theta}_n^w$ and $\hat{\theta}_n$:
  - i) consistency and asymptotic normality for the normalized estimator and
  - ii) construct a version of AIC*.
- Following the work of Taniguchi (1987), (2000) and especially Dahlhaus and Wefelmeyer (1996) let D be a discrepancy or distance function and define

$$\theta_0 = \arg\min_\theta D(\theta, g) \text{ and } \theta(\hat{g}_n) = \arg\min_\theta D(\theta, \hat{g}_n).$$

  The main objective in these articles is to determine under what circumstances $n^{1/2}(\theta(\hat{g}_n) - \theta_0)$ has a Gaussian limit distribution.
- Note that if $\hat{g}_n = I_n$ (periodogram) and

$$D(\theta, g) = (4\pi)^{-1} \int_{-\pi}^{\pi} \{\log(f_\theta) + g(\omega)/f_\theta(\omega)\}\, d\omega.$$

- Result i) above, for both $\hat{\theta}_n^w$ and $\hat{\theta}_n$, is proved in Dahlhaus and Wefelmeyer (1996).

Derivation of AIC* for stationary time series

- It follows now that

$$\hat{\theta}_n^w = \arg\max_\theta \ell_n^w(f_\theta) \to_{P_g} \theta_0 = \arg\min_\theta d(g, f_\theta)$$

and

$$n^{1/2}(\hat{\theta}_n^w - \theta_0) \to_d J^{-1}U =_d N_p(0, J^{-1}KJ^{-1})$$

where J and K are defined below and the same result is also true for $\hat{\theta}_n$.

- We further have the process convergence

$$H_n^w(s) = \ell_n^w(f_{\theta_0 + sn^{-1/2}}) - \ell_n^w(f_{\theta_0}) \to_d H(s) = s^t U - 2^{-1} s^t J s,$$

where $U \sim N_p(0, K)$,

$$K = \frac{1}{2\pi} \int_0^\pi \dot{\Psi}_{\theta_0}(\omega) \dot{\Psi}_{\theta_0}(\omega)^t [g(\omega)/f_{\theta_0}(\omega)]^2 \, d\omega$$

and

$$J = \frac{1}{2\pi} \int_0^\pi \dot{\Psi}_{\theta_0}(\omega) \dot{\Psi}_{\theta_0}(\omega)^t g(\omega)/f_{\theta_0}(\omega) + \ddot{\Psi}_{\theta_0}(\omega)(f_{\theta_0}(\omega) - g(\omega))/f_{\theta_0}(\omega) \, d\omega,$$

and where $\Psi_\theta = \log f_\theta$, $\dot{\Psi}_\theta = \partial/\partial\theta \log f_\theta$ and $\ddot{\Psi}_\theta = \partial^2/\partial\theta\partial\theta^t \log f_\theta$.

- Note that $J = K$ if the model is correctly specified.

The J and K matrices with estimates

- By extending and using results from Ibragimov (1963) and Brillinger (1975) we obtain consistent estimates for K and J from

$$\frac{2\pi}{n}\sum_{j=1}^{\lfloor n/2\rfloor} h(\omega_j)I_n(\omega_j) \to_{P_g} \int_0^\pi h(\omega)g(\omega)\,d\omega \text{ and}$$

$$\frac{2\pi}{n}\sum_{j=1}^{\lfloor n/2\rfloor} k_j h(\omega_j)I_n(\omega_j)^2 \to_{P_g} \int_0^\pi h(\omega)g(\omega)^2\,d\omega$$

  where $\omega_j = 2\pi j/n$ and $k_j = 1/3$ if $\omega_j = 0 \bmod \pi$ else $k_j = 1/2$, where $j = 0, \ldots, n-1$.
- It is also possible to construct estimates based on integrals of I_n, see Taniguchi (1980).

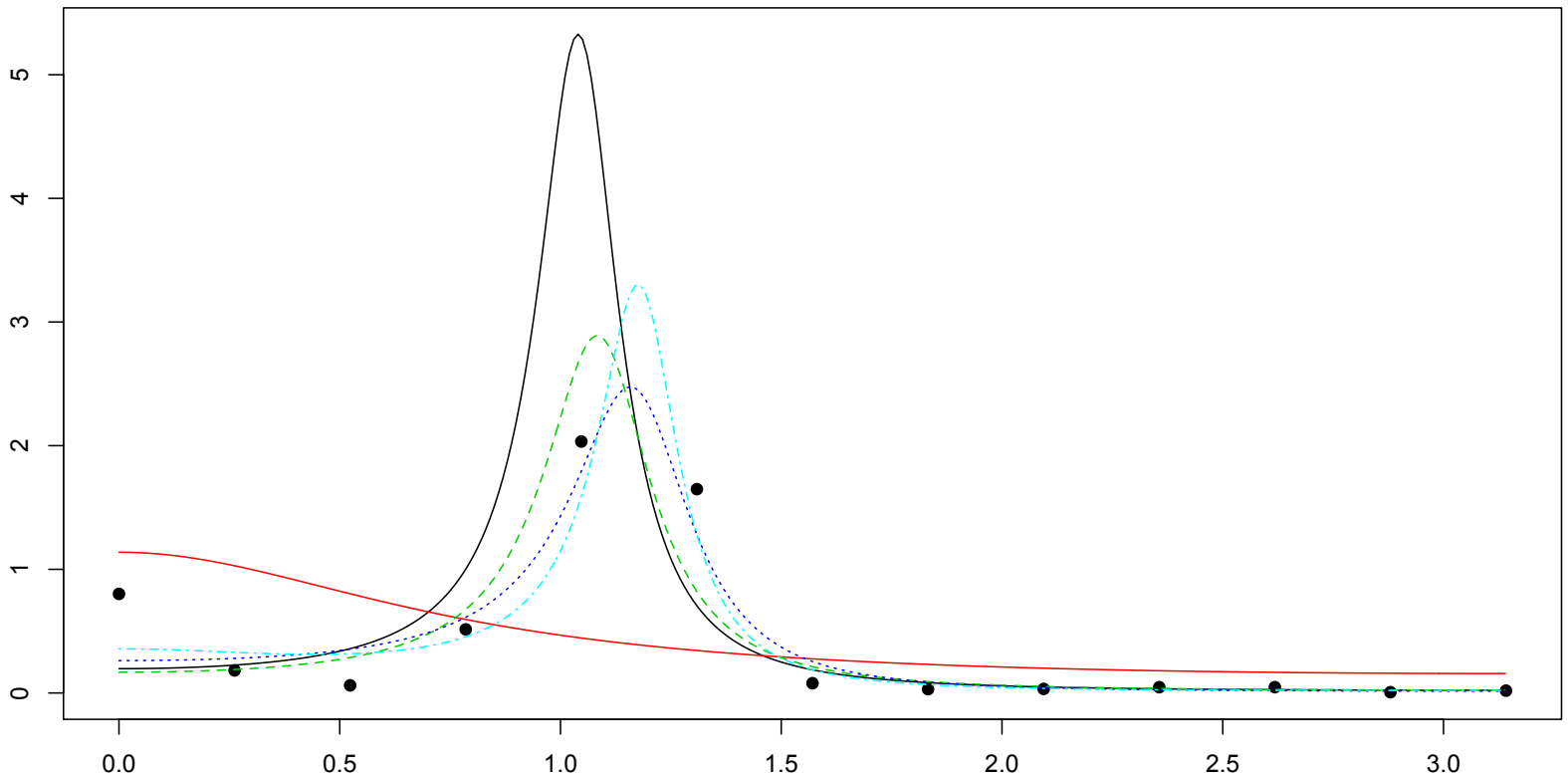
Combining all of this we get a version of the AIC* formula, SIC (Spectral Information Criterion) given by

$$\mathrm{SIC}(f_\theta) = 2\ell_n(f_{\hat\theta_n}) - 2\,\mathrm{tr}(\hat J_n^{-1}\hat K_n\ )$$

as an alternative one can also use $l_n^w(f_{\hat\theta_n^w})$.

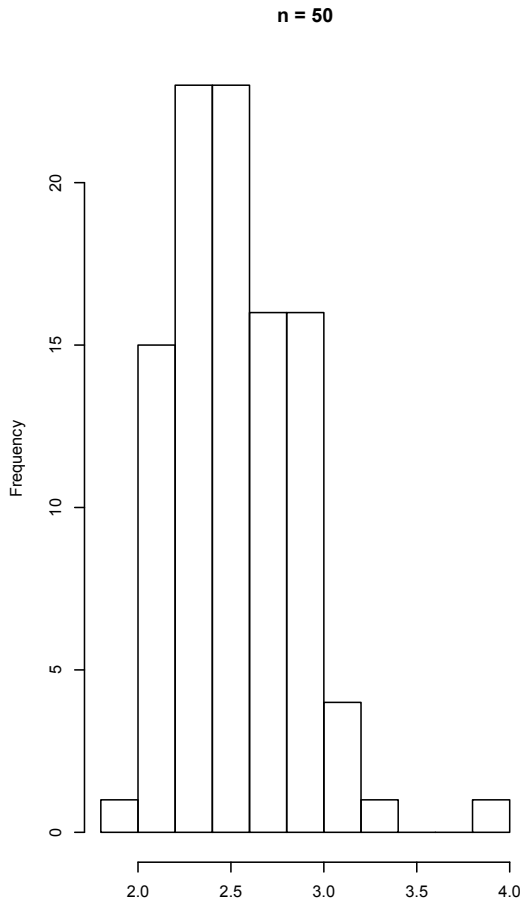
Examples 1: $\hat{p}^*$ as a penalization

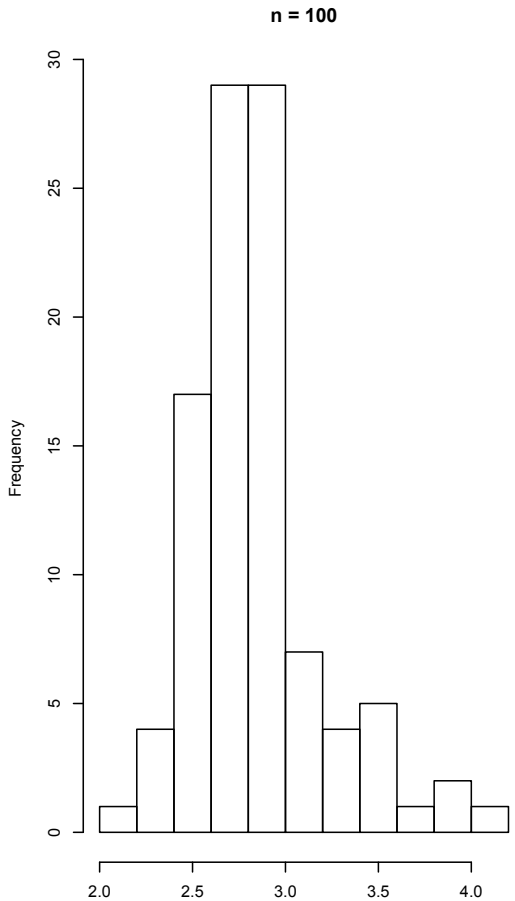
$n = 24$, $\rho = (0.9, -0.8)$ and $\sigma = 1$.

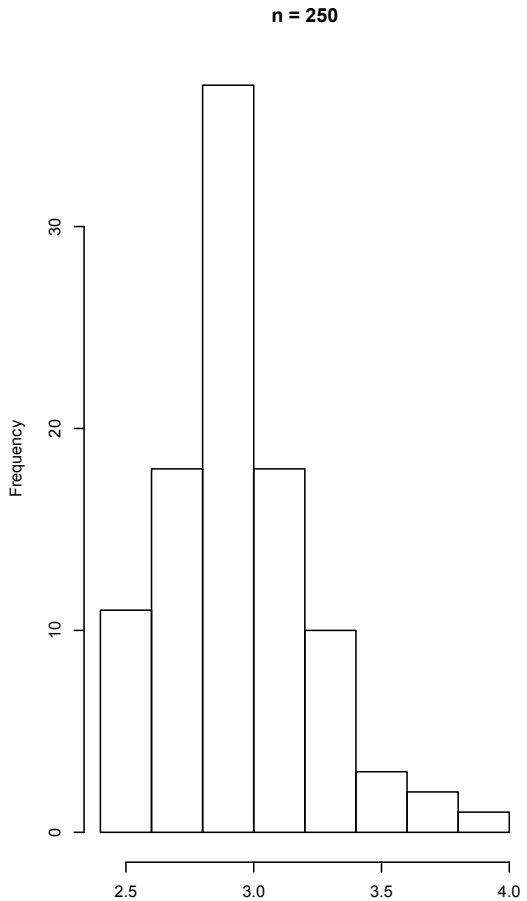


$p^* = (5.54, 3.98, 5.98, 7.62)$ and $\hat{p}^* = (2.03, 2.37, 3.24, 3.64)$.

Example 2: Estimate of p^* in a Matérn inspired model







Example 3: AR(p) models

	Selected model order						
Criterion	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
AIC_u	4	94	2	0	0	0	0
AIC_c	1	87	8	2	2	0	0
AIC	0	66	9	4	6	6	9
SIC	1	70	11	10	6	2	0

$n = 24$, $\rho = (0.9, -0.8)$ and $\sigma = 1$.

	Selected model order						
Criterion	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
AIC_u	27	9	4	52	5	2	1
AIC_c	16	7	4	57	6	5	5
AIC	14	7	4	56	6	7	6
SIC	5	6	4	57	14	6	8

$n = 84$, $\rho = (0.3, -0.2, 0.2, -0.3)$ and $\sigma = 1$.

Concluding remarks and further work

- Better estimate of K and J, especially in small samples.
- A more complete simulation study.
- Include a trend component.
- A version of FIC, Claeskens & Hjort (2008).
- Relax the model assumptions, for example non-Gaussian models.