Εργασία Ομάδας 10

Παναγιώτης Σαββίδης (8094)

Αντώνης Φάββας (8675)

Παύλος Φραγκιαδουλάκης (8389)

Εισαγωγή

Η τρέχουσα μελέτη βασίζεται στο κείμενο του Venus Khim-Sen Liew πάνω στην επιλογή τάξης αυτοπαλινδρόμησης (autoregression - AR) με την χρήση κάποιων βασικών κριτηρίων. Υπάρχουν διάφορες παραλλαγές επιλογής, οι πιο σημαντικές εκ των οποίων:

1. Κριτήριο Πληροφορίας Akaike (Akaike, 1973).
2. Κριτήριο Πληροφορίας Schwarz (Schwarz, 1978)
3. Κριτήριο Hannan-Quinn (Hannan Quinn, 1979)
4. Κριτήριο Τελικής Πρόβλεψης (Final Prediction) (Akaike, 1969)
5. Κριτήριο Πληροφορίας Bayes (Akaike, 1979)

Ο συγγραφέας ισχυρίζεται ότι τα μοντέλα αποδίδουν σωστά ακόμα και σε μικρά δείματα. Στις περισσότερες περιπτώσεις δεν υπάρχει ιδιαίτερος λόγος προτίμησης κάποιου κριτηρίου, όπως αυτό φαίνεται παρακάτω. Στην τρέχουσα μελέτη παρουσιάζονται οδηγίες για την επιλογή της τάξης AR διαδικασιών. Οι οδηγίες στην τρέχουσα εργασία έχουν επεκταθεί ώστε να συμπληρώνουν τα σημεία που δεν περιγράφει με λεπτομέρεια ο Liew.

Μία διαδικασία AR τάξης p αναφέρεται σε μία διακριτή χρονοσειρά τιμών, των οποίων η τρέχουσα τιμή εξαρτάται από τις p προηγούμενες (οι τιμές αυτές ονομάζεται $lags$).

$$AR(p) = y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \varepsilon_t$$

Στην πράξη το p δεν είναι γνωστό όπως και οι παράμετροι a. Τα παραπάνω κριτήρια αναφέρονται συχνά σε οικονομοτεχνικές μελέτες. Η διερεύνηση του Liew έχει περιοριστεί για $p = 3$, ενώ αγνοεί τα y_0, a_0 τα οποία (τελευταία δύο) γενικά δεν αφορούν το αποτέλεσμα της ανάλυσης. Για λόγους πληρότητας, η τρέχουσα μελέτη (Ομάδας 10) παρουσιάζει αποτελέσματα και αναλύσεις με εναλλακτικές παραμέτρους.

Μεθοδολογία

Μία διαδικασία AR(p) περιγράφεται από την παρακάτω εξίσωση:

$$AR(p) = y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \varepsilon_t$$

όπου a_oείναι ο intercept όρος, τα $a_0, a_1, a_2, ..., a_p$ είναι οι όροι αυτοπαλινδρόμησης και ε_t είναι οι όροι σφάλματος που ακολουθούν κανονική κατανομή με πεπερασμένη διακύμανση σ^2. Ο Liew δεν αναφέρει τι τιμές παίρνει για την διακύμανση αυτή. Στην τρέχουσα μελέτη επιλέχθηκαν διάφορες τιμές του σ^2.

Τα βήματα που ακολουθούνται για την επιλογή του p για κάθε κριτήριο είναι τα παρακάτω:

1. Προσομοίωση μίας σειράς συγκεκριμένου μεγέθους δείγματος (sample size) χρησιμοποιώντας τυχαίους όρους αυτοπαλινδρόμησης οι οποίοι θα πρέπει να ανήκουν στο $(-1, 1)$ για να επιτυγχάνεται εγγυημένη στασιμότητα (stationarity). Τα δείγματα επιλέγονται μεταξύ των τιμών 25, 50, 100, 200, 400, 800, 1600. Τα δείγματα προσαυξάνονται κατά 100 (ignored sample size) για την αποφυγή των αρχικών μεταβατικών φαινομένων.

2. Για κάθε p μεταξύ του 1 και 20, υπολογίζονται οι όροι αυτοπαλινδρόμησης με μεθόδους παλινδρόμησης. Ο Liew δεν αναφέρει με ποια μέθοδο πραγματοποιεί την παλινδρόμηση.

3. Για κάθε p υπολογίζεται κάποιο από τα κριτήρια που αναφέρονται στην επόμενη λίστα και στην

συνέχεια επιλέγεται εκείνο με την χαμηλότερη τιμή.

Τα κριτήρια που χρησιμοποιούνται είναι τα παρακάτω:

α) Κριτήριο Πληροφορίας Akaike

$$AIC_p = 2S \cdot ln(\widehat{\sigma}_p^2) + 2p$$

β) Κριτήριο Πληροφορίας Schwarz

$$SIC_p = ln(\widehat{\sigma}_p^2) + \frac{plnS}{S}$$

γ) Κριτήριο Hannan-Quinn

$$HQC_p = ln(\widehat{\sigma}_p^2) + 2 \cdot \frac{pln(ln(S))}{S}$$

δ) Κριτήριο Τελικής Πρόβλεψης

$$FPE_p = \widehat{\sigma}_p^2 \frac{(S+p)}{(S-p)}$$

ε) Κριτήριο Πληροφορίας Bayes

$$BIC_p = (S-p) \cdot ln\left[\frac{S\sigma_p^2}{S-p}\right] + S[1 + ln\sqrt{2\pi}] + pln\left[\frac{\left(\sum_{t=p}^{S} y_t^2 - S\widehat{\sigma}_p^2\right)}{p}\right]$$

όπου $\widehat{\sigma}_p^2 = \frac{\sum_{t=p}^{S} \widehat{\varepsilon}_t^2}{S-p-1}$ και S είναι το μέγεθος δείγματος

Τα στατιστικά για την αξιολόγηση κάθε κριτηρίου προκύπτουν από τα παρακάτω βήματα:

1. Τρέξιμο των προηγούμενων βημάτων (1 έως 3) για B = 1000 φορές (replications).

2. Σε κάθε τρέξιμο καταμετράμε τις μεθόδους που προβλέπουν σωστά την τάξη του p. Στο τέλος παράγονται στατιστικά σχετικών συχνοτήτων για τις περιπτώσεις σωστής εκτίμησης (success όπου $\hat{p} = p$), υποεκτίμησης (underestimate όπου $\hat{p} < p$) και υπερεκτίμησης (overestimate όπου $\hat{p} > p$).

Θεωρητικό Υπόβαθρο χαρακτηριστικών κριτηρίων

Κριτήριο Akaike

Στατιστικό Μοντέλο: $Y_j = h(t_j; q) + \varepsilon_j$,$j = 1, 2, \ldots, n$

- $h(t_j; q)$: το δείγμα παρατηρήσεων ενός μαθηματικού μοντέλου με παράμετρο χ στην χρονική στιγμή που προσδιορίζεται από την μεταβλητή t_j.
- ε_j :το σφάλμα την χρονική στιγμή t_j(τυχαία μεταβλητή).
- Y_j:παρατήρηση την χρονική στιγμή t_j(τυχαία μεταβλητή).

Κάτω από την υπόθεση ότι η μεταβλητή ε_jείναι μια ακολουθία ομοίως ανεξάρτητων κατανεμημένων τυχαίων μεταβλητών (ιδ) που λαμβάνουν τιμές με βάσει μια κανονική κατανομή $N(0, \sigma^2)$. Έχουμε:

$$g(\boldsymbol{y}|\theta) = \prod_{j=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_j - h(t_j;q))^2}{2\sigma^2} \right]$$

- g: συνάρτηση πυκνότητας πιθανότητας του $\boldsymbol{\Upsilon} = (\Upsilon_1, \Upsilon_{2,\ldots,} \Upsilon_n)^T$δεδομένης παραμέτρου θ.
- $\theta = (q, \sigma)^T$.
- $\boldsymbol{y} = (y_1, y_{2,\ldots,} y_n)^T$:είναι το δείγμα παρατηρήσεων του Ψ.

Kullback-Leibler Information

Γενική ιδέα: Μέρος της πληροφορίας χάνεται όταν προσπαθούμε να εκτιμήσουμε τα πραγματικά δεδομένα με ένα μοντέλο εκτίμησης. Για την ανάλυση γίνεται η υπόθεση ότι φ, γ είναι συναρτήσεις κατανομών μια συνεχής τυχαίας μεταβλητής.

$$I(f, g(\cdot|\theta)) = \int_{\Omega} f(\boldsymbol{x}) log(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x}|\theta)}) d\boldsymbol{x}$$

$$= \int_{\Omega} f(\boldsymbol{x}) log(f(\boldsymbol{x})) d\boldsymbol{x} - \int_{\Omega} f(\boldsymbol{x}) log(g(\boldsymbol{x}|\theta)) d\boldsymbol{x}$$

- f: πραγματικό μοντέλο κατανομής πιθανότητας
- g: μοντέλο εκτίμησης κατανομής πιθανότητας
- θ: παράμετρος διάνυσμα στο μοντέλο εκτίμηση γ.

Παρατηρήσεις:

1. $I(f,g) \neq I(g,f)$ το οποίο συνεπάγεται πως το Κ-Λ ινφορματιον δεν είναι η πραγματική "απόσταση" των 2 κατανομών.
2. $I(f,g) \geq 0$
3. $I(f,g) = 0$ ανν $f = g$ σχεδόν παντού.

Έχοντας ως βάση όλα τα παραπάνω ορίζεται:

- Η πραγματική συνάρτηση κατανομή πιθανότητας f είναι άγνωστη.
- Η παράμετρος θ στην συνάρτηση g πρέπει να εκτιμηθεί από τα εμπειρικά δεδομένα $\{y_j\}_{j=1}^{n}$.
- $\boldsymbol{y} = (y_1, y_{2,\ldots,}y_n)^T$ είναι ένα τυχαίο δείγμα παρατηρήσεων από την κατανομή πυκνότητας πιθανότητας $f(y)$της τυχαίας μεταβλητής Y.
- $\theta^*(\boldsymbol{\Upsilon})$: εκτιμητής του ϑ. Είναι τυχαία μεταβλητή.

Επειδή η απόσταση $I(f, g(\cdot|\theta^*(\boldsymbol{Y})))$ είναι τυχαία μεταβλητή, πρέπει να ληφθεί την ανεμενόμενη τιμή αυτής ως προς $\boldsymbol{Y}$.

$$E_{\boldsymbol{Y}}[I(f, g(\cdot|\theta^*(\boldsymbol{Y})))] = \int_{\Omega} f(\boldsymbol{x}) log(f(\boldsymbol{x}))d\boldsymbol{x} - \int_{\Omega} f(\boldsymbol{y}) \left[\int_{\Omega} f(\boldsymbol{x}) log(g(\boldsymbol{x}|\theta^*(\boldsymbol{y})))d\boldsymbol{x}\right] d\boldsymbol{y}.$$

Όπως είναι εύκολα κατανοητό εφόσον η παραπάνω συνάρτηση πρόκειται για την αναμενόμενη τιμή της διαφοράς ανάμεσα στο πραγματικό μοντέλο και το μοντέλο εκτίμησης, θέλουμε να ελαχιστοποιηθεί αυτή την διαφορά. Άρα θέλουμε να μεγιστοποιήσουμε τον όρο

$$\int_{\Omega} f(\boldsymbol{y}) \left[\int_{\Omega} f(\boldsymbol{x}) log(g(\boldsymbol{x}|\theta^*(\boldsymbol{y})))d\boldsymbol{x}\right] d\boldsymbol{y} = E_{\boldsymbol{Y}} E_{\boldsymbol{X}}[log(g(X|\theta * (Y)))].$$

Ένας μη-πολωμένος εκτιμητής της $E_{\boldsymbol{Y}} E_{\boldsymbol{X}}[log(g(X|\theta * (Y)))]$ για μεγάλα δείγματα είναι

$$log(\mathcal{L}(\hat{\theta}|\boldsymbol{y}) - k$$

- $\left(\mathcal{L}(\hat{\theta}|\boldsymbol{y}\right)$: αναπαριστά την πιθανοφάνεια του $\hat{\theta}$ με δεδομένα τα $\{y_j\}_{j=1}^{n}$και είναι $\mathcal{L}(\hat{\theta}|\boldsymbol{y}) = g(\boldsymbol{y}|\hat{\theta})$.
- $\hat{\theta}$: εκτιμητής μέγιστης πιθανοφάνειας του θ με δεδομένα τα $\{y_j\}_{j=1}^{n}$ και είναι $\hat{\theta} = \theta^*(\boldsymbol{y})$.
- k: είναι ο αριθμός των παραμέτρων (συμπεριλαμβανομένων των μαθηματικών και στατιστικών μοντέλων).

ΑΙΚ για την περίπτωση μέγιστης πιθανοφάνειας

$$AIC = -2log(\mathcal{L}(\hat{\theta}|\boldsymbol{y}) + 2k.$$

Υπολογισμός της τιμής του ΑΙΚ για κάθε μοντέλο με το ίδιο σετ δεδομένων, και το "καλύτερο" μοντέλο θα είναι αυτό με την ελάχιστη τιμή.

ΑΙΚ για την περίπτωση ελάχιστων τετραγώνων: Υποθέτουμε ότι έχουμε σφάλματα ανεξάρτητα ομοίως κατανεμημένα με βάσει την κανονική κατανομή.

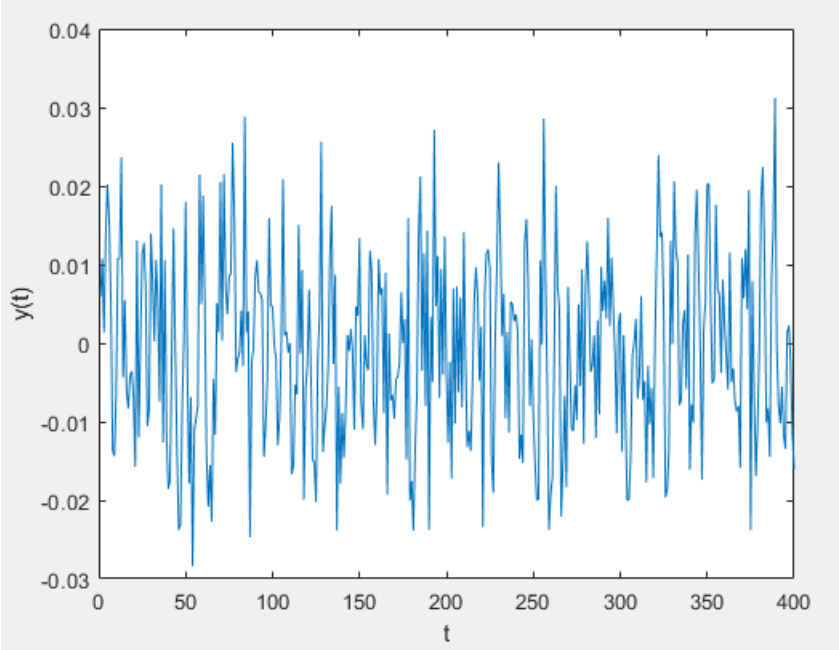
$$AIC = nlog(\frac{RSS}{n}) + 2k$$

όπου RSS είναι το άθροισμα των τετραγώνων των υπολοίπων του προσαρμοσμένου μοντέλου.

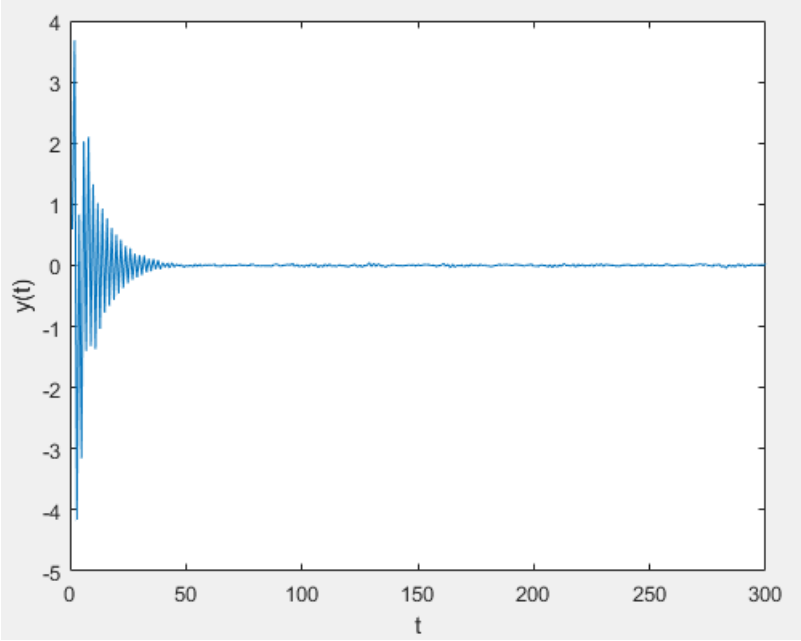
Αποτελέσματα

Για την κατανόηση και επαλήθευση των διαδικασιών δημιουργίας της χρονοσειράς AR και των ενδιάμεσων βημάτων. Εκτελέστηκαν αρχικά διάφορα σενάρια τα οποία επικυρώνουν τη μεθοδολογία και τα αποτελέσματα που προέκυψαν.

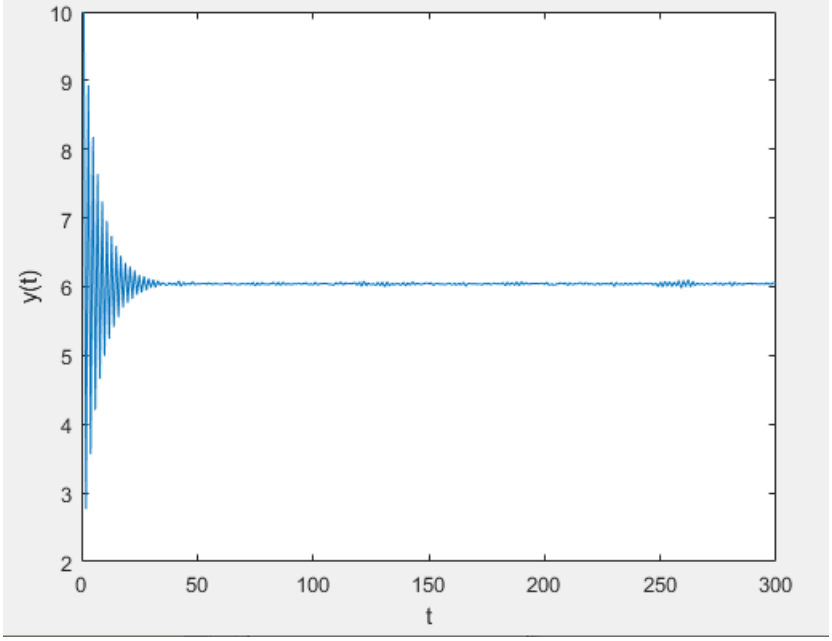
Αρχικά επιβεβαιώθηκε η παραγωγή ψευδοτυχαίας AR χρονοσειράς. [Να σημειωθεί ότι οι τιμές για το $t = 0$ παραλείπονται στα παρακάτω γραφήματα.] Στην παρακάτω εικόνα φαίνεται μία τυπική χρονοσειρά με μέγεθος δείγματος 300 και $y_0 = 0, a_0 = 0, \sigma = 0.01$:



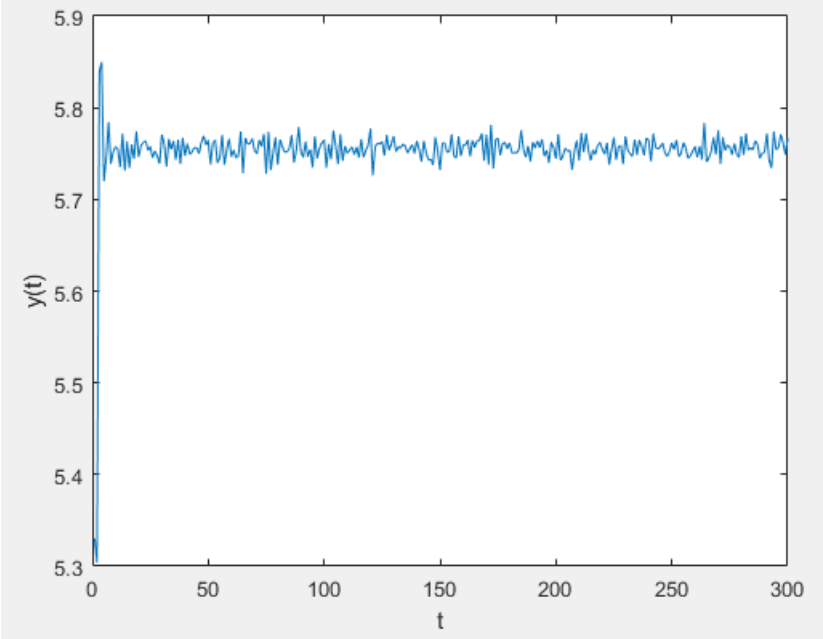
Στην δεύτερη περίπτωση φαίνεται μία τυπική χρονοσειρά με μέγεθος δείγματος 300 και $y_0 = 10, a_0 = 0, \sigma = 0.01$. Αυτό που είναι προφανές από την δεύτερη εικόνα είναι η χρησιμότητα του discard sample size, όταν το y_0 είναι μη μηδενικό. Το "ενδιαφέρον" κομμάτι της διαδικασίας είναι η μόνιμη κατάσταση. Φαίνεται ότι η μέση τιμή στη μόνιμη κατάσταση τείνει στο 0:



Η μέση τιμή στη μόνιμη κατάσταση καθορίζεται από το a_0. Στο παρακάτω σενάριο $y_0 = 0, a_0 = 10, \sigma = 0.01$

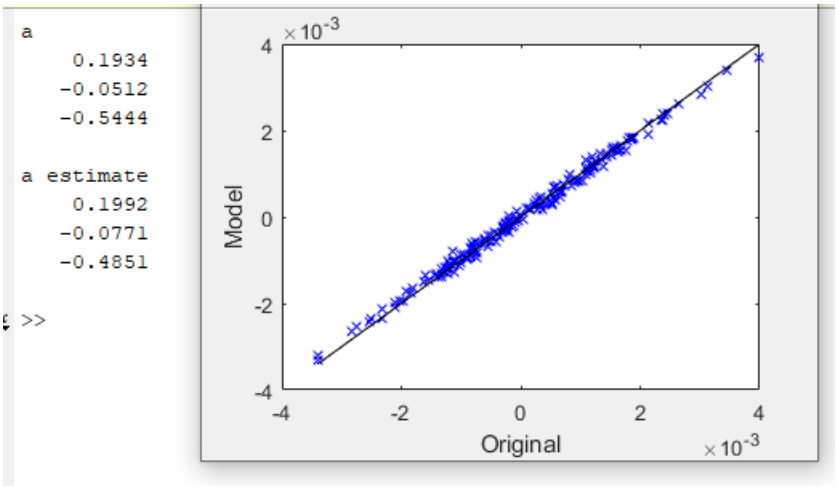


Τέλος, παρουσιάζεται μία συνδυαστική περίπτωση με $y_0 = 10, a_0 = 10, \sigma = 0.01$.

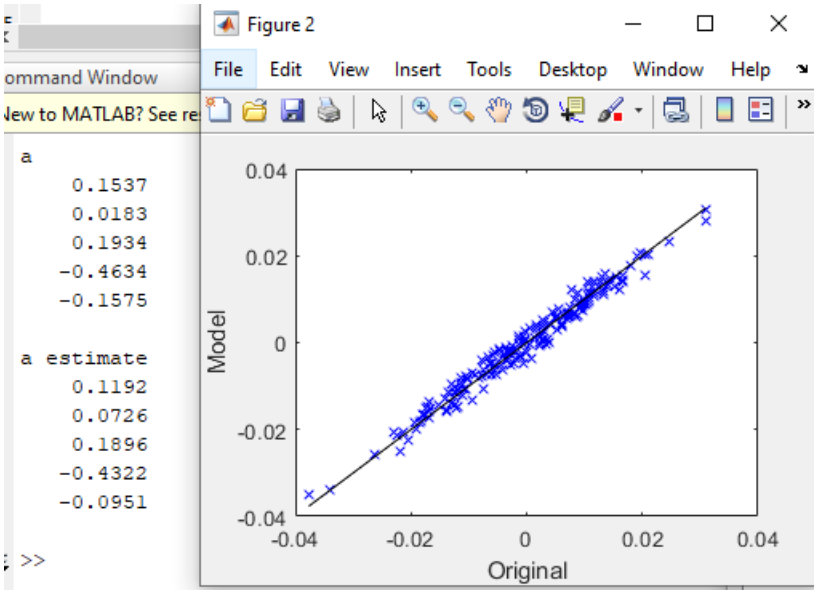


Σημαντικό κομμάτι της ανάλυσης είναι η σωστή εκτίμηση των παραμέτρων α. Για τον σκοπό

αυτό χρησιμοποιήθηκε η μέθοδος των Ελαχίστων Τετραγώνων. Λόγω της ύπαρξης του θορύβου (σφαλμάτων) η εκτίμηση των α είναι κοντά αλλά δεν ταυτίζεται με τις original τιμές. Πραγματοποιήθηκε εκτίμηση για διάφορα p. Παρακάτω φαίνεται η εκτίμηση για $p = 3$ και οι τιμές που έχουν αναπαραχθεί με την χρήση των "νέων εκτιμούμενων" a.

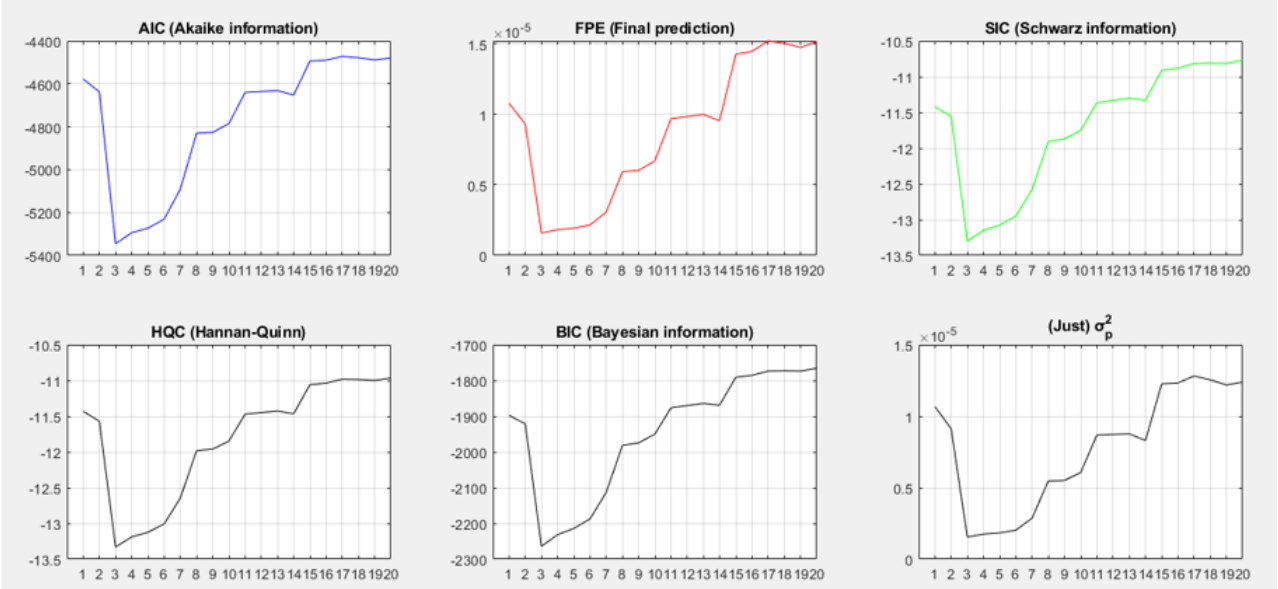


Σε μία επόμενη δοκιμή για $p = 5$ φαίνεται επίσης οτι οι τιμές των y_t που έχουν αναπαραχθεί είναι επίσης κοντά στις πρωτότυπες.

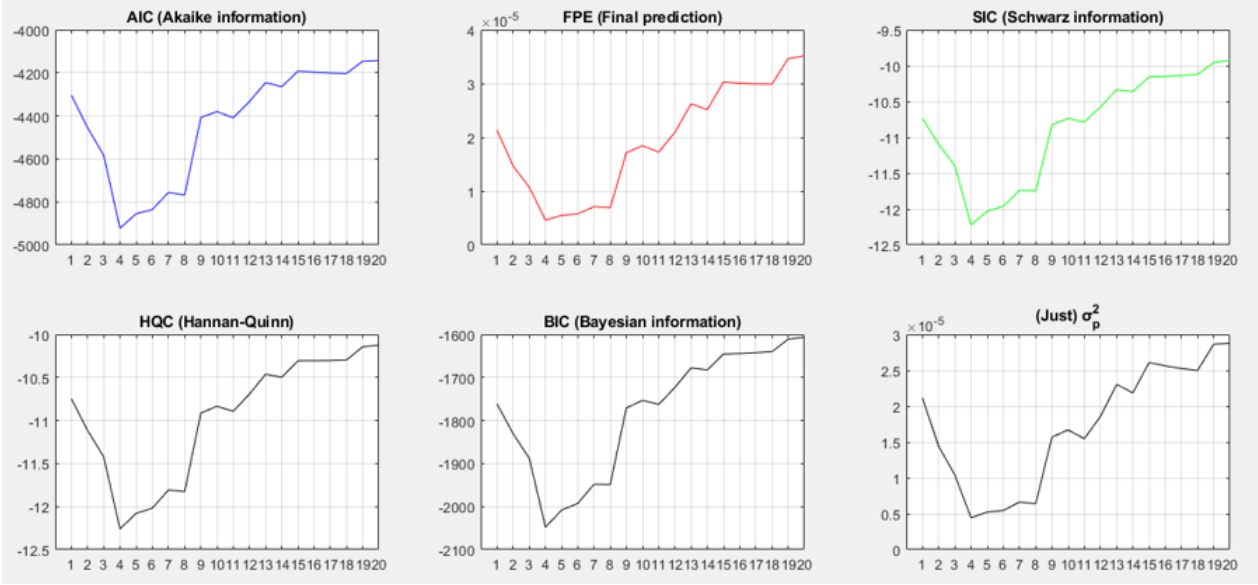


Μετά την επιβεβαίωση της μεθόδου ελαχίστων τετραγώνων που χρησιμοποιήθηκε για την εκτίμηση των παραμέτρων αυτοπαλινδρόμησης, εκτελέστηκαν μεμονωμένα σενάρια εκτίμησης βέλτιστης

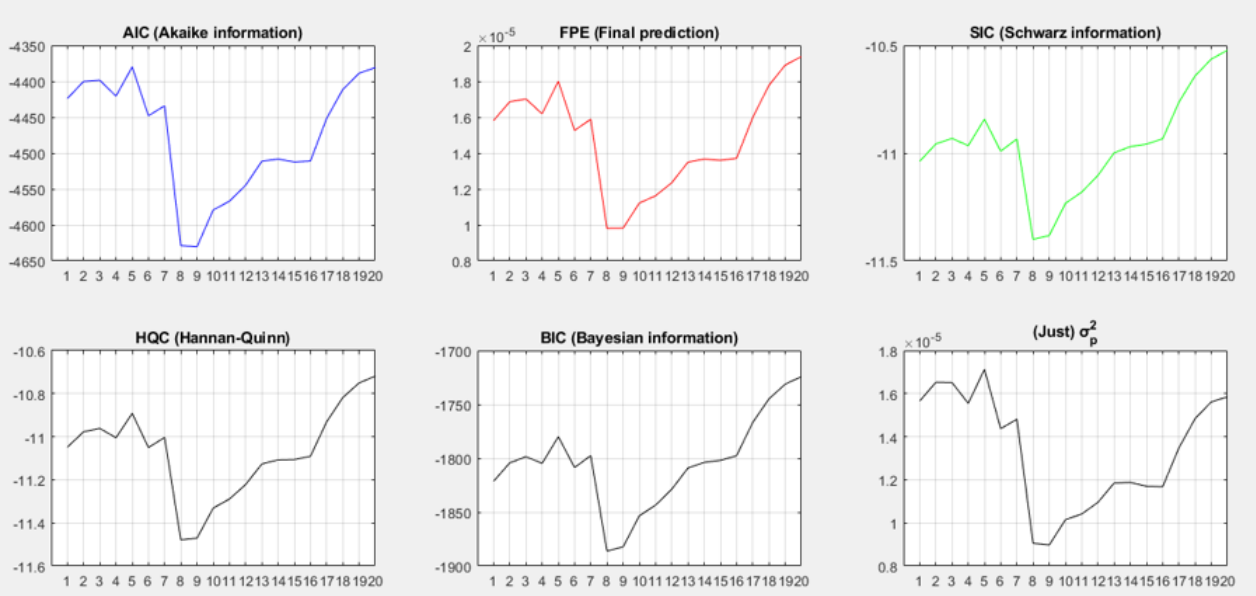
παραμέτρων για διάφορα p. Παρακάτω φαίνονται οι διάφορες τιμές των κριτήριων για μία διαδικασία $AR(3)$. Όπως φαίνεται, όλα τα κριτήρια έχουν ελάχιστο στην τιμή $p = 3$ (οριζόντιος άξονας).



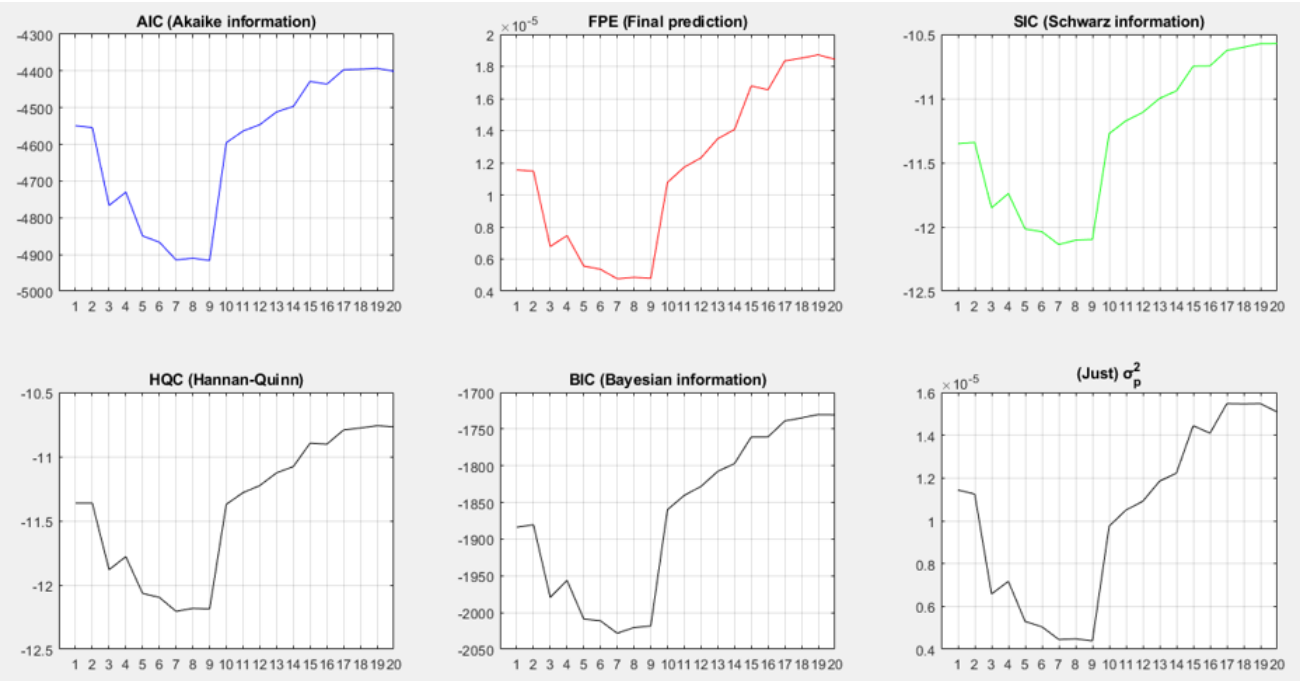
Για $p = 4$ τα αποτελέσματα είναι εξίσου καλά:



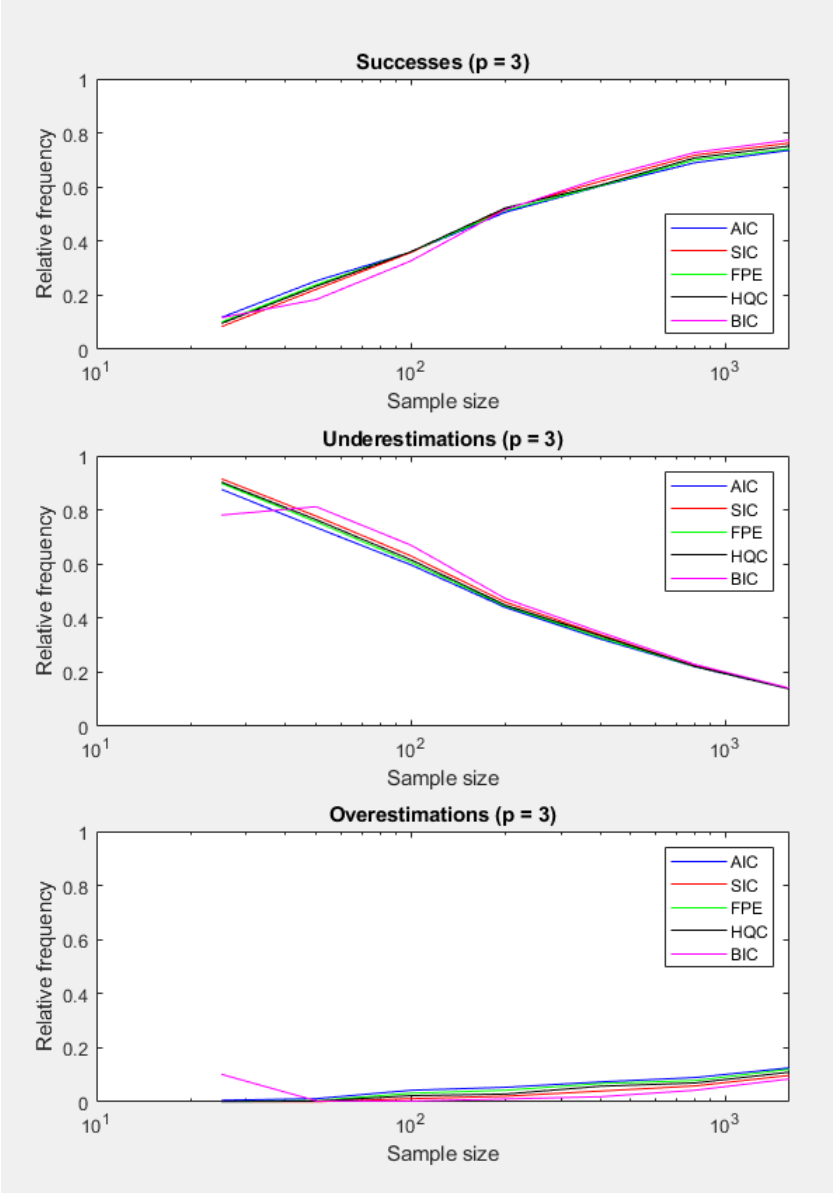
Για $p = 8$ φαίνεται ότι αρχίζουν να εμφανίζονται τα "προβλήματα" των κριτήριων. Η σωστή πρόβλεψη είναι οριακή:



Για $p = 10$ παρακάτω, παρατηρείται το αναμενόμενο: τα κριτήρια δεν αποδίδουν καλά γιατί ακριβώς έχουν φτιαχτεί για να αποφεύγεται το overestimate (όχι το underestimate). Εδώ εμφανίζεται ως βέλτιστη η τιμή για $\hat{p} = 9$:



Έχοντας επαληθεύσει όλους τους ενδιάμεσους αλγορίθμους, παρήχθεισαν τα ζητούμενα διαγράμματα για p=3. Οι δοκιμές φαίνεται ότι επαληθεύουν μερικώς τους ισχυρισμούς του Liew. Για μεγάλα δείγματα πράγματι τα κριτήρια βρίσκουν σωστά την τάξη κοντά στο 80%. Όχι όμως για μικρά δείγματα.



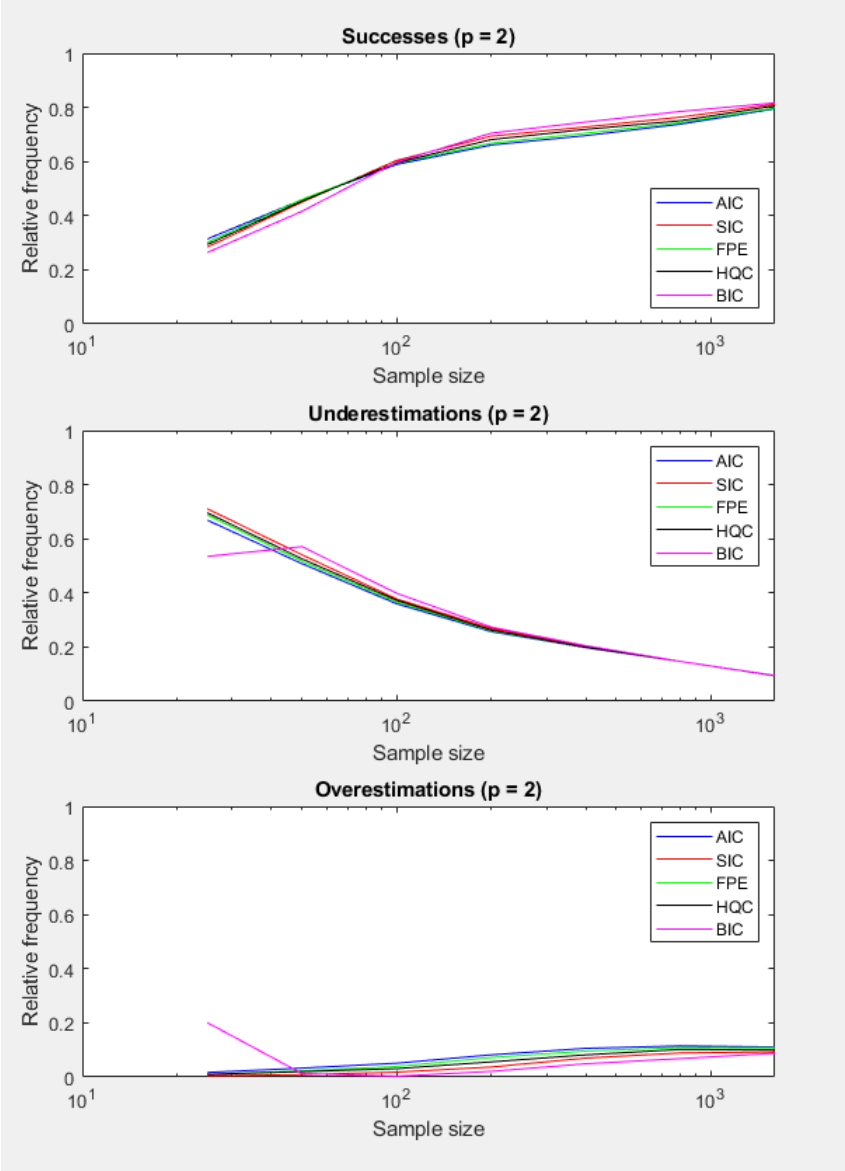
Τιμές για $p = 3$:

```
Successes
    0.1700    0.1400    0.1500    0.1400    0.1500
    0.2900    0.2300    0.3000    0.2900    0.1900
    0.3500    0.3300    0.3300    0.3200    0.2900
    0.5200    0.5200    0.5300    0.5200    0.5000
    0.5700    0.6300    0.5800    0.5800    0.6100
    0.6600    0.6700    0.6600    0.6600    0.6800
    0.7400    0.7600    0.7300    0.7500    0.7700

Underestimations
    0.8300    0.8600    0.8500    0.8600    0.8000
    0.6900    0.7600    0.6900    0.7000    0.8000
    0.6300    0.6700    0.6500    0.6600    0.7100
    0.4200    0.4500    0.4200    0.4500    0.4700
    0.3400    0.3400    0.3400    0.3400    0.3800
    0.2300    0.2300    0.2300    0.2300    0.2300
    0.1600    0.1700    0.1700    0.1700    0.1700

Overestimations
         0         0         0         0    0.0500
    0.0200    0.0100    0.0100    0.0100    0.0100
    0.0200         0    0.0200    0.0200         0
    0.0600    0.0300    0.0500    0.0300    0.0300
    0.0900    0.0300    0.0800    0.0800    0.0100
    0.1100    0.1000    0.1100    0.1100    0.0900
    0.1000    0.0700    0.1000    0.0800    0.0600
```

Τα ποσοστά βελτιώνονται για p=2 για μικρά δείγματα.



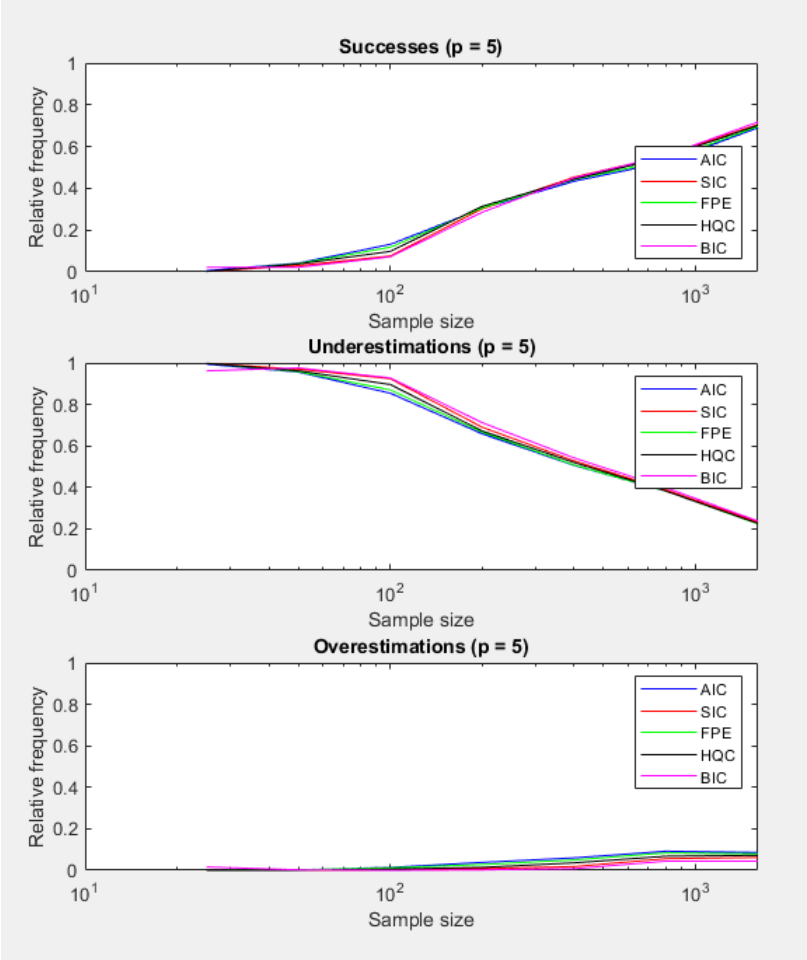
Τα δεδομένα για $p = 2$:

```
Successes
    0.3500    0.3300    0.3300    0.3300    0.2500
    0.4300    0.4200    0.4300    0.4400    0.4100
    0.6000    0.5800    0.6100    0.5900    0.5400
    0.7500    0.7400    0.7400    0.7400    0.7400
    0.7300    0.7800    0.7400    0.7700    0.7800
    0.7100    0.7300    0.7100    0.7100    0.7600
    0.7800    0.7800    0.7800    0.7800    0.8100

Underestimations
    0.6200    0.6600    0.6400    0.6500    0.5000
    0.5300    0.5600    0.5300    0.5300    0.5800
    0.3700    0.4200    0.3900    0.4100    0.4600
    0.2200    0.2300    0.2300    0.2300    0.2400
    0.1900    0.1900    0.1900    0.1900    0.1900
    0.1400    0.1400    0.1400    0.1400    0.1400
    0.1200    0.1200    0.1200    0.1200    0.1200

Overestimations
    0.0300    0.0100    0.0300    0.0200    0.2500
    0.0400    0.0200    0.0400    0.0300    0.0100
    0.0300         0         0         0         0
    0.0300    0.0300    0.0300    0.0300    0.0200
    0.0800    0.0300    0.0700    0.0400    0.0300
    0.1500    0.1300    0.1500    0.1500    0.1000
    0.1000    0.1000    0.1000    0.1000    0.0700
```

Τα ποσοστά όπως αναμένεται χειροτερεύουν για τα μικρά δείγματα για μεγαλύτερα p.



Τα δεδομένα για $p = 5$:

Successes

0.0200	0.0100	0.0200	0.0200	0.0200
0.0100	0.0100	0.0100	0.0100	0
0.1000	0.0700	0.0900	0.0800	0.0700
0.2800	0.2700	0.2800	0.2700	0.2600
0.4300	0.4800	0.4400	0.4700	0.4700
0.5200	0.5500	0.5200	0.5300	0.5400
0.5900	0.6300	0.6000	0.6200	0.6400

Underestimations

0.9800	0.9900	0.9800	0.9800	0.9700
0.9900	0.9900	0.9900	0.9900	1.0000
0.9000	0.9300	0.9100	0.9200	0.9300
0.6900	0.7300	0.6900	0.7100	0.7400
0.5200	0.5200	0.5200	0.5200	0.5300
0.3700	0.3900	0.3900	0.3900	0.4000
0.3300	0.3300	0.3300	0.3300	0.3300

Overestimations

0	0	0	0	0.0100
0	0	0	0	0
0	0	0	0	0
0.0300	0	0.0300	0.0200	0
0.0500	0	0.0400	0.0100	0
0.1100	0.0600	0.0900	0.0800	0.0600
0.0800	0.0400	0.0700	0.0500	0.0300

Συμπερασματικά, ενώ επαληθεύονται τα αποτελέσματα σε υψηλούς αριθμούς δειγμάτων, τα αποτελέσματα σε χαμηλούς αριθμούς δειγμάτων δεν επαληθεύονται.

Βιβλιογραφία

1. Akaike, H. 1969. Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics, 21, 243 – 247.

2. Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, B. N. Petrov and F. Csaki (eds.), Akademiai Kiado, Budapest, 267 – 281.

3. Akaike, H. 1979. A Bayesian extension of the minimum AIC procedure of Autoregressive

model fitting. Biometrika, 66, 237 – 242.

4. Hannan, E. J. and Quinn, B. G. 1978. The determination of the order of an autoregression. Journal of Royal Statistical Society, 41, 190 – 195.

5. Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics, 6, 461 – 464.