Information-Theoretic Sample Size Selection for Linear Prediction

Demosthenes N. Tambakis[1]

Abstract: What is the appropriate number of past observations to use in forecasting univariate linear processes? A non-parametric statistic useful for sample size selection is proposed involving the data's average information content (AIC). It is shown that the asymptotic predictability of a process is increasing in its AIC. Monte Carlo simulations of stationary pdf's indicate that AIC increases with sample size, suggesting that "more is better", while for stock market returns over a large number of sample sizes the AIC and mean squared forecast error are signi¯cantly negatively correlated.

Key words: Predictability, average information content, sample size
Revised and accepted: 18 December 1999

1 Introduction

What is the optimal number of past observations to use in order to maximize the forecast accuracy of univariate linear models? Applied research often proceeds on the implicit assumption that more in-sample observations are better.[2] This is justi¯ed if the data's probability density function (pdf) is Gaussian, or more generally if OLS is consistent. The question of optimal sample size is related to that of the appropriate out-of-sample separation point for assessing the performance of competing forecasting models. In that respect, strict stationarity of a data vector implies independence from initial conditions. However, a non-Gaussian but linear | i.e. satisfying the Wold representation | data generating process (dgp) only satis¯es weak stationarity, which is necessary but not su±cient for strict stationarity (Hamilton (1994)). Therefore, sample size matters. In the case of extreme events such as stock market crashes, the decision to include the relevant observation in the sample can signi¯cant a®ect a forecasting model's out-of-sample performance.

[1]City University Business School, Frobisher Crescent, Barbican Centre, London EC2Y 8HB, U.K.

[2]See, for example, Granger and Newbold (1986).

This paper proposes a sample's average information content (AIC) as an information-theoretic predictability measure. The sample AIC is de¯ned as the sample entropy normalized by its alphabet length. In turn, non-parametric predictability is de¯ned as the mutual information between the random variable to be forecast and the ensemble of past observations, normalized by the alphabet length underlying the sample's empirical probability distribution. The main results are as follows. First, asymptotic predictability is shown to be increasing in the AIC and decreasing in the entropy rate of the dgp. Second, the general relation between forecast error probability and AIC is non-monotonic. Empirically, we examine the behavior of the AIC for simulated data from known pdf's and show that it is increasing in sample size, thus justifying using all available data. Then, using time series of daily returns for the Dow Jones and Nikkei stock market averages, we show that AIC is non-monotonic in the sample size. Estimating linear autoregressive models and comparing the evolution of the AIC against that of mean squared forecast error (MSE) for changing sample size suggests that AIC and MSE can be signi¯cantly negatively correlated.

In Section 2 the information-theoretic concepts of entropy and predictability are introduced and used to obtain the average information content of a dataset. In Section 3 the implications of the theoretical properties are explored for simulated and actual data. Section 4 concludes the paper.

2 Information theory

2.1 De¯nitions

An n{vector of observations $\{x_t; x_{t-1}; ...; x_{t-n+1}\}$ from discrete random variable X is denoted $\underline{x}^n$. If the data generating process (dgp) of X is strictly stationary and ergodic, the statistics of $\underline{x}^n$ (do not) depend on n (t). The sample entropy of X is $H_n^k(X) = -\sum_{i=1}^{k} p_i \log p_i$, where $\{p_i\}_{i=1}^{k}$ are the empirical probabilities of observations partitioned into equally-spaced percentiles $i = 1; ...; k$. The log is to base 2, so the entropy units are information bits. The percentile ensemble is de¯ned as the alphabet of the dgp. Its length k de¯nes the partition: a ¯ner (coarser) partition amounts to a bigger (smaller) alphabet. For discrete random variables, the value of maximum entropy occurs for the uniform probability density function (pdf) where $p_i = 1/k$ for

all i: $\max H_n^k(X) = \log k$.[3] In "normal" circumstances the alphabet length is invariant to the sample size. However, an "extreme" marginal observation added to the sample may necessitate a marginal increase in alphabet length from k to $k + 1$.

The joint and conditional entropies of random variables $\{X_1; X_2; :::; X_n\}$ with joint pdf $p(x_1; :::; x_n)$ are respectively:

$$H_n^k(X_1; :::; X_n) = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} p(x_1; :::; x_n) \log p(x_1; :::; x_n) \qquad (1)$$

$$H_n^k(X_n \mid \underline{X}^{n-1}) = \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} p(x_n \mid \underline{x}^{n-1}) \log p(x_n \mid \underline{x}^{n-1}) , \qquad (2)$$

where $p(x_n \mid \underline{x}^{n-1})$ is the conditional pdf of X_n given past observations $\underline{X}^{n-1} \{x_1; :::; x_{n-1}\}$. In general, the mutual information $I(X; Y)$ of random variables X and Y is the relevant information in one variable for predicting the other:

$$\begin{aligned} I(X; Y) &= H(X) - H(X \mid Y) \qquad (3) \\ &= H(X) + H(Y) - H(X; Y) , \end{aligned}$$

where X is the (unobservable) input forecast using a noisy observable channel output (signal) Y. Mutual information is symmetric and non-negative. If X and Y are independent then $H(X \mid Y) = H(X)$, so $I(X; Y) = 0$ and Y is useless in predicting X, while if X is a deterministic function of Y then $H(X \mid Y) = 0$ and mutual information is maximized.[4]

The univariate non-parametric predictability $P_n^k(X_n; \underline{X}^{n-1})$ of random variable X_n as a function of $\underline{X}^{n-1}$ is mutual information normalized by the maximum entropy of a discrete dgp with a k-alphabet:

$$P_n^k(X_n; \underline{X}^{n-1}) = \frac{H_n^k(X_n) - H_n^k(X_n \mid \underline{X}^{n-1})}{\log k} \qquad (4)$$

Normalization implies that $P_n^k(X_n)$ is bounded between 0 and 1.[5]

[3] See Applebaum (1996) and Golan, Judge and Miller (1996).

[4] If $X = x_{t+j}$ and $Y = x_t$ then $I(x_{t+j}; x_t)$ is the information about x_{t+j} contained in x_t. As $j \to 0$ then mutual information reduces to the entropy.

[5] Fraser (1989) and Palus (1993) define the numerator of the predictability statistic to be the (non-linear) redundancy measure.

2.2 Asymptotic predictability and forecast error

The entropy rate of a random variable sequence $\{X_i\}_{i=1}^{n}$ is defined as $H^k(n) = \lim_{n \to \infty} \frac{1}{n} H_n^k(X_1, \ldots, X_n)$. Thus $H^k(n)$ is the limit of the average joint entropy per observation. Clearly, if the $\{X_i\}$ sequence is iid then $H^k(n) = \lim_{n \to \infty} nH_n^k(X_n)/n = H_n^k(X_n)$. Khinchin (1957) shows the existence of $H^k(n)$ for strictly stationary processes. Moreover, for strictly stationary ergodic processes conditional entropy converges to the entropy rate:[6]

$$\lim_{n \to \infty} H_n^k(X_n \mid \underline{X}^{n-1}) = H^k(n) \tag{5}$$

Taking limits and substituting (5) in (4) yields:

$$\lim_{n \to \infty} P_n^k(X_n; \underline{X}^{n-1}) = \frac{1}{\log k}[\lim_{n \to \infty} (H_n^k(X_n)) - H^k(n)] \tag{6}$$

The second term in (6) converges to the entropy rate, implying that asymptotic predictability is greatest when $\lim_{n \to \infty} H_n^k(X_n)/\log k$ is maximized, while it is zero for an iid sequence. We define the resulting finite-sample statistic as the average information content (AIC) of a dataset of sample size n and alphabet length k:

$$AIC_n^k = \frac{H_n^k(X_n)}{\log k} \tag{7}$$

We turn to examine the relation between AIC and univariate forecast error. Let $g(\underline{X}^{n-1}) = \hat{X}_n$ be a (linear/non-linear) forecast of X_n. The error probability for sample size n can be written $E(n) = E(\hat{X}_n \neq X_n)$: There is no one-to-one relationship between predictability and forecast error probability, but a tight lower bound relating predictability and error probability is given by Fano's inequality.[7] This is given by:

$$E(n) \geq \frac{H_n^k(X_n \mid X_1, \ldots, X_{n-1}) - 1}{\log k} \tag{8}$$

The error probability is low only if the conditional entropy $H_n^k(X_n \mid \underline{X}^{n-1})$ is small. Taking limits and substituting the entropy rate from equation (5) yields $\lim_{n \to \infty} E(n) \geq (H^k(n) - 1)/\log k$. Thus, the lower bound of the error

[6] For the proof see Cover and Thomas (1991).

[7] If $P_e(n) = 0$ then $H_n^k(X_n \mid \underline{X}^{n-1}) = 0$. A tight upper bound for error probability has also been established by Feder and Merhav (1994).

probability increases in the entropy rate. We know that for iid processes, which are completely unpredictable, this expression becomes:

$$\lim_{n \to \infty} E(n) \approx \frac{H_n^k(X_n) - 1}{\log k} \tag{9}$$

Asymptotic error probability is increasing in the AIC. Therefore, for iid processes the relation between general loss function evaluation criteria | such as mean squared error | proxying for error probability and the AIC is monotonic. However, if a process is not iid then the relationship between asymptotic error probability, predictability and the AIC is likely non-linear.

3 Average information content and sample size

3.1 Monte Carlo simulations

We first illustrate the behavior of AIC_n^k as sample size changes from $n = 1$ to 500 using simulated data from the Gaussian $(0;1)$, uniform $[0;1]$ and gamma $(1;1)$ distributions. Each simulated n{vector is partitioned using a discrete alphabet of fixed length k: an n{vector thus yields k^n possible output signals. Given the alphabet length, the density function for a given sample size corresponds to the empirical frequency distribution. Figure 1 shows that given alphabet length $k = 100$, AIC_n^k increases with sample size.[8] This property is robust to alternative distributions and alphabet length. Also note that, overall, the greatest asymptotic AIC occurs for the uniform pdf. This follows from the maximum entropy principle because the simulated data, although drawn from continuous pdf's, have been discretized. Therefore, $n^* = \arg\max_n AIC_n^k$ coincides with the maximum sample size, implying that predictability increases in the number of observations. Intuitively, "more is better" because the underlying (true) dgp is strictly stationary. We now turn to analyze the behavior of average information content with sample size and its relation to forecast accuracy for actual financial data.

[8] The increase is not monotonic because there is only one sample: if many random samples of length n were generated then AIC_n^k would be smoothly increasing in n.

3.2 Illustration with stock market returns

The time series used are daily returns on the Dow Jones Industrial (DJIA) and the Nikkei (NIK) stock market averages over the period 1=1=1973{ 6=4=1998, a total 6; 591 observations. The data is (weakly) stationary over the sample period. The average information content and forecast accuracy statistics are both in°uenced by the choice of cut-o® observation and consequent size of the in-sample data vector. In addition, the average information content is a®ected by the alphabet length k. In turn, the forecast accuracy measure is also in°uenced by the speci¯cation of forecasting model, the information criterion for ARMA (parametric) lag order selection, the loss function used to evaluate forecast accuracy and the length of the forecast horizon.

We discuss each factor in turn. First, the empirical distribution is used to compute the time series average information content AIC_n^k, as in Abarbanel (1996). As discussed in Section 2, the subscript n denotes the changing sample size while the superscript k denotes the ¯xed alphabet length. For both time series, the cut-o® obervation of the in-sample data is ¯xed at 6; 500, so the length of the out-of-sample period is ¯xed at 91. The in-sample size is increased incrementally from n_{MIN} to n_{MAX} observations by moving the ¯rst in-sample observation backward one day at a time. For illustration purposes, for the Dow Jones $n_{MIN} = 400$ and $n_{MAX} = 700$, or 300 rolling sample sizes, while for the Nikkei $n_{MIN} = 400$ and $n_{MAX} = 4; 400$, or 4; 000 rolling sample sizes. The data is partitioned in $k = 100$ equally-spaced percentiles.[9] Second, for the same sample size, parametric forecast error is evaluated according to the mean-squared error (MSE) criterion. The forecast horizon is ¯xed at $j = 10$ days ahead and the forecasts are dynamic. The MSE_n^j statistic is computed using a linear AR speci¯cation, where the lag order is determined is determined using the modi¯ed Schwartz information criterion of Neumaier and Schneider (1997). The AR parameters, including the order speci¯cation, are reestimated at each sample size increment.

For each stock market average, the left panels in Figure 2 plot the evolution of the AIC_n^k and MSE_n^j statistics on separate scales with changing sample size n. The plotted values are conditional upon the alphabet length, forecasting model, AR order information criterion and forecast horizon. Unlike the simulated datasets for the known pdf's, AIC is clearly non-monotonic

[9] Following Golan, Judge and Miller (1996), the alphabet length k must be less than the sample size n in order for the recovery of the probability vector $\{p_i\}_{i=1}^k$ to be well-de¯ned. This constraint is unlikely to be binding in practice.

in sample size: average information content is very sensitive to the arrival of "extreme" datapoints (outliers). This suggests that maximizing AIC may contribute to lower mean squared error. The correlation coefficients of the two statistics are $\rho_{DJIA}(AIC_{300}^{100}; MSE_{300}^{10}) = -0.495$ for the Dow Jones Industrial and $\rho_{NIK}(AIC_{4000}^{100}; MSE_{4000}^{10}) = -0.648$ for the Nikkei average. The significance of these values is examined by bootstrapping the underlying statistics. The right panels in Figure 2 show the empirical distribution of each bootstrap correlation statistic for 1,000 bootstrap replications of the AIC_n^k and MSE_n^j vectors. In each case, the bootstrap histograms strongly suggest that the true correlation coefficient are significantly negative. Robustness of the results to alternative datasets and to the factors mentioned above is the subject of current research.

4 Conclusion

This paper analyzed the problem of sample size selection in an information-theoretic framework. Theoretically, it was shown that a higher average information content improves predictability, while the link between the AIC and error probability was less clear. Simulations of known stationary pdf's suggested that AIC increases with sample size. An empirical application to financial data indicated that AIC and mean square forecast error are negatively correlated, suggesting that AIC can be used to improve the forecast accuracy of linear parametric models by appropriate selection of in-sample size. Extensions to this framework include relating parametric and parametric predictability using Fisher information, and analyzing multivariate predictability using the mutual information between different datasets.

References

[1] Abarbanel, H.D.I. 1996. Analysis of Observed Chaotic Data. New York: Springer Verlag.

[2] Applebaum, D. 1996, Probability and Information Theory: An Integrated Approach, Cambridge: Cambridge University Press.

[3] Cover, T. and Thomas, J. 1991, Elements of Information Theory. New York: Wiley Interscience.

[4] Feder, M. and Merhav N. 1994. Relations Between Entropy and Error Probability. IEEE Transactions on Information Theory 40(1): 259-266.

[5] Fraser, A.M. 1989. Reconstructing Attractors from Scalar Time Series: A Comparison of Singular System and Redundancy Criteria. Physica D 34: 391-404.

[6] Golan, A., Judge, G. and Miller, D. 1996, Maximum Entropy Econometrics: Robust Estimation With Limited Data. New York: John Wiley & Sons.

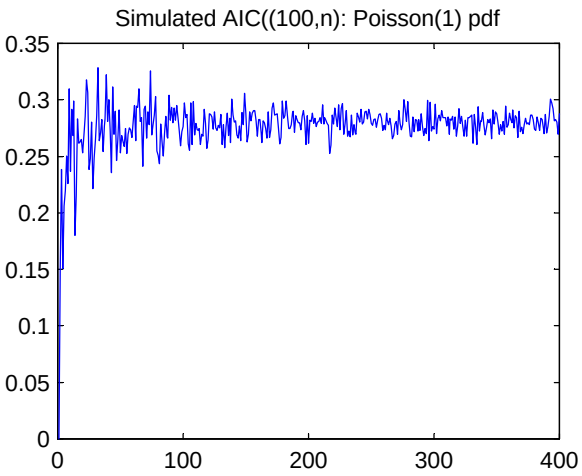
[7] Granger, C.W.J. and Newbold D. 1986: Forecasting Economic Time Series, Cambridge: Cambridge University Press.

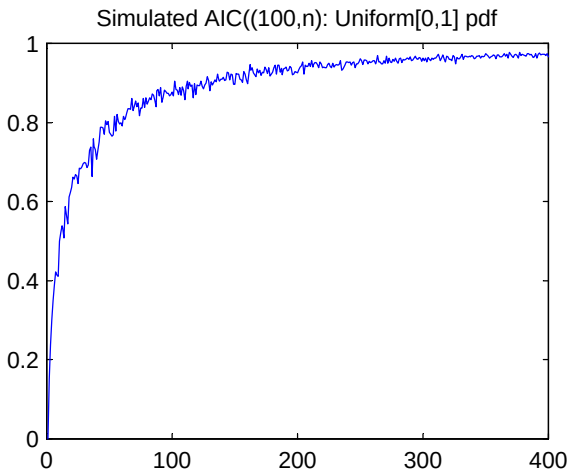
[8] Hamilton, J. (1994): Time Series Analysis. Princeton: Princeton University Press

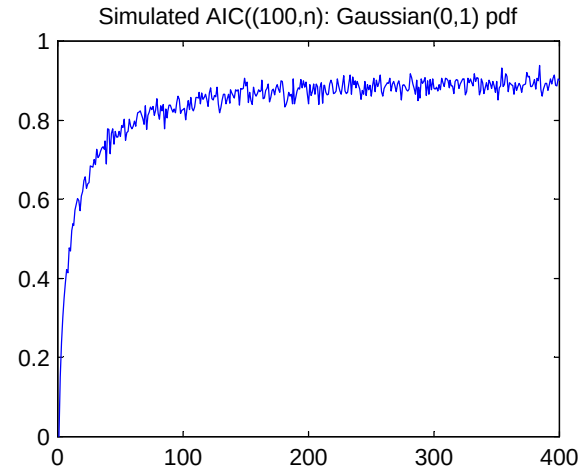
[9] Khinchin, A.I. (1957): Mathematical Foundations of Information Theory. New York: Dover.

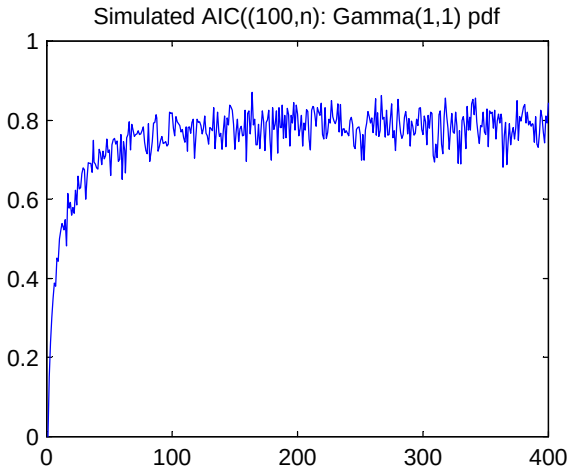
[10] Neumaier, A. and Schneider, T. 1997, Multivariate Autoregressive and Ohrstein-Uhlenbeck processes: Evidence for order, parameters, spectral information and con¯dence regions, submitted to ACM Trans. Math. Soft.

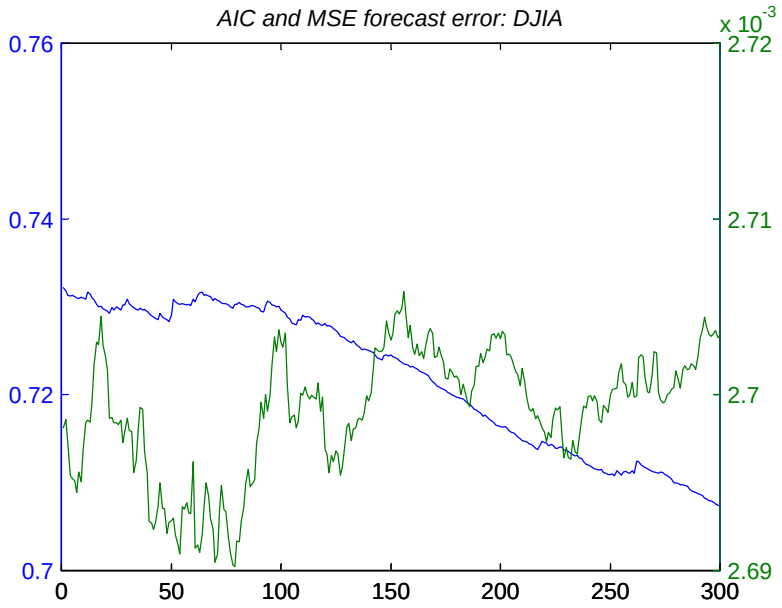
[11] Palus, M. 1993. Identifying and Quantifying Chaos by Using Information-Theoretic Functionals, in Time Series Prediction: Forecasting the Future and Understanding the Past, eds. A.S. Weigend and N.A. Gershenfeld, SFI Studies in the Sciences of Complexity, Proc. Vol. XV, Addison-Wesley.

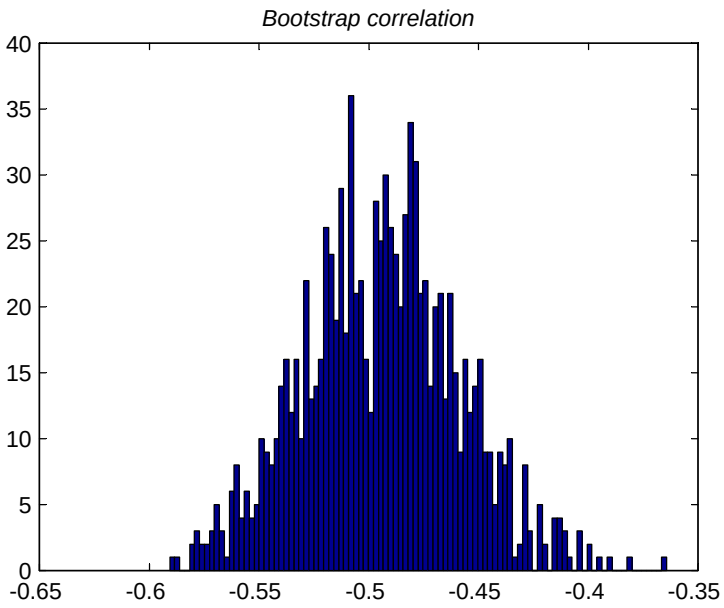

Simulated AIC((100,n): Poisson(1) pdf
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
100
200
300
400

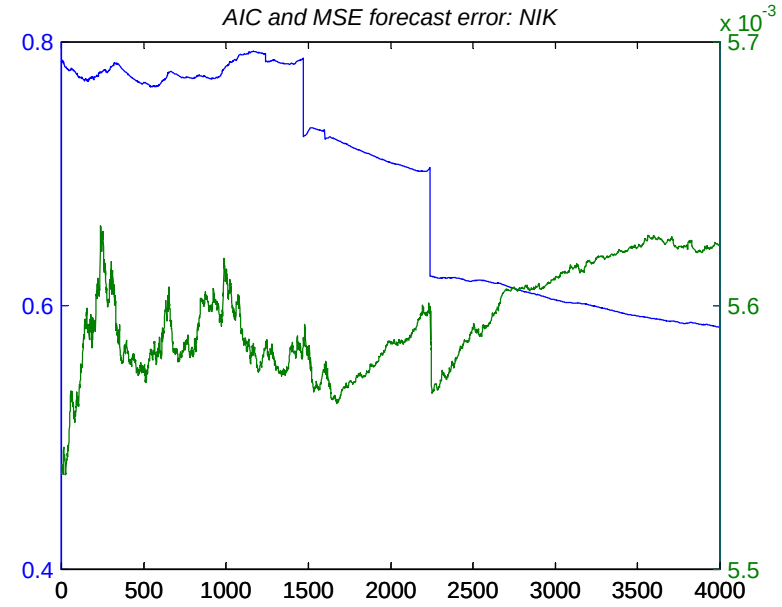

Simulated AIC((100,n): Uniform[0,1] pdf
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400

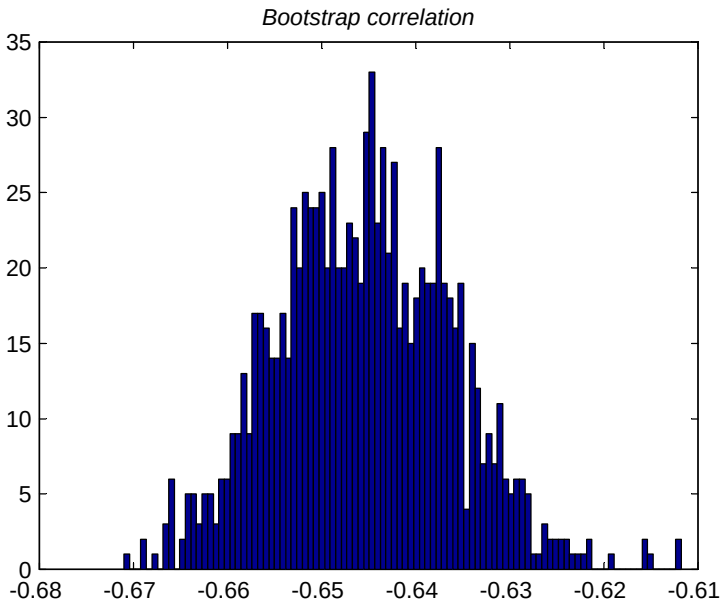

Simulated AIC((100,n): Gaussian(0,1) pdf
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400


Simulated AIC((100,n): Gamma(1,1) pdf
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400


AIC and MSE forecast error: DJIA
x 10^-3
0.76
0.74
0.72
0.7
2.72
2.71
2.7
2.69
0
50
100
150
200
250
300


Bootstrap correlation
40
35
30
25
20
15
10
5
0
-0.65
-0.6
-0.55
-0.5
-0.45
-0.4
-0.35


AIC and MSE forecast error: NIK
x 10^-3
0.8
0.6
0.4
5.7
5.6
5.5
0
500
1000
1500
2000
2500
3000
3500
4000


Bootstrap correlation
35
30
25
20
15
10
5
0
-0.68
-0.67
-0.66
-0.65
-0.64
-0.63
-0.62
-0.61