Master in Applied Econometrics and Forecasting

Time Series Analysis and Forecasting

Class #6: Model Identification, Estimation, Testing and Selection

Jorge Caiado, PhD

CEMAPRE/REM/ISEG, University of Lisbon, jcaiado@iseg.ulisboa.pt

ARMA Models: Identification, Estimation, Testing and Selection

Model Identification

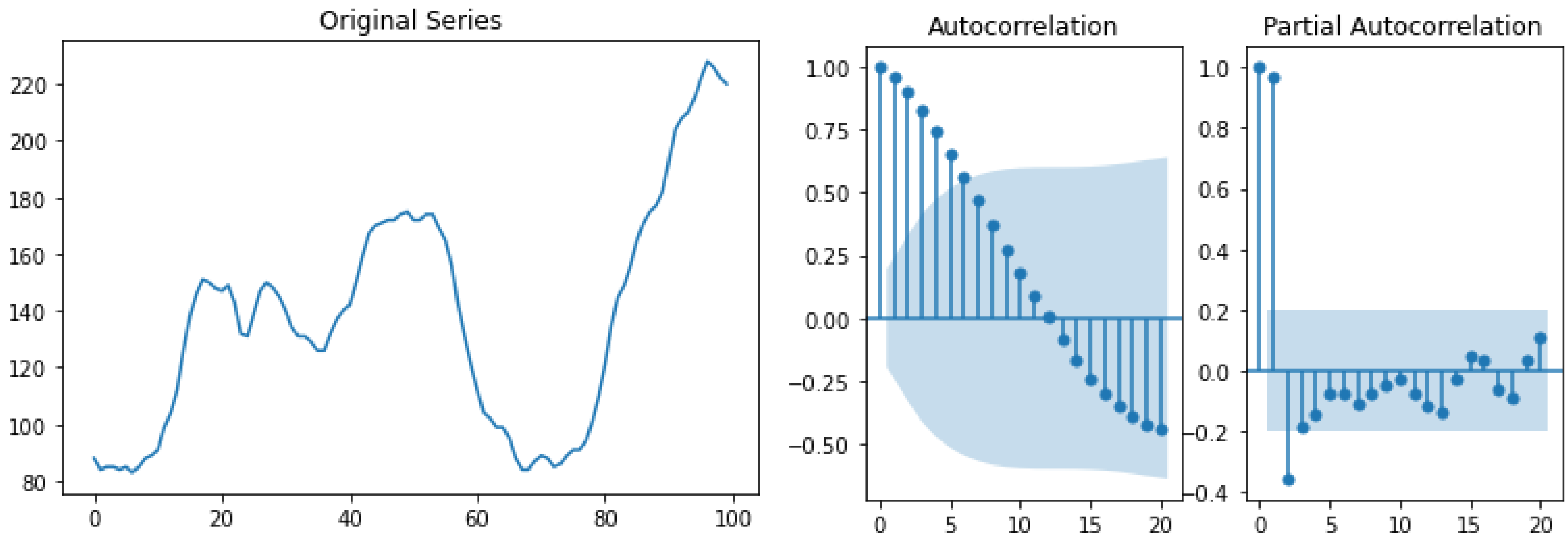
Plot the time series and examine whether the series contains a trend, seasonality, ouliers, nonconstant variances and other nonstationary phenomena. Choose proper variance-stabilizing (Box-Cox's power transformation) and differencing transformations.

Compute the sample ACF and the sample PACF of the original series and identify the degree of differencing *d* to achieve stationarity, $(1-B)^d X_t$. In practice, *d* is either 0, 1, or 2. Later, we will introduce a formal test to determine the order of integration (or degree of differentiation) of the time series (unit root test).

Compute the sample ACF and the sample PACF of the stationary series and identify the orders *p* and *q* for the autoregressive and moving average operators. Usually, the needed orders of integers *p* and *q* are less or equal to 3.

ARMA Models: Identification, Estimation, Testing and Selection

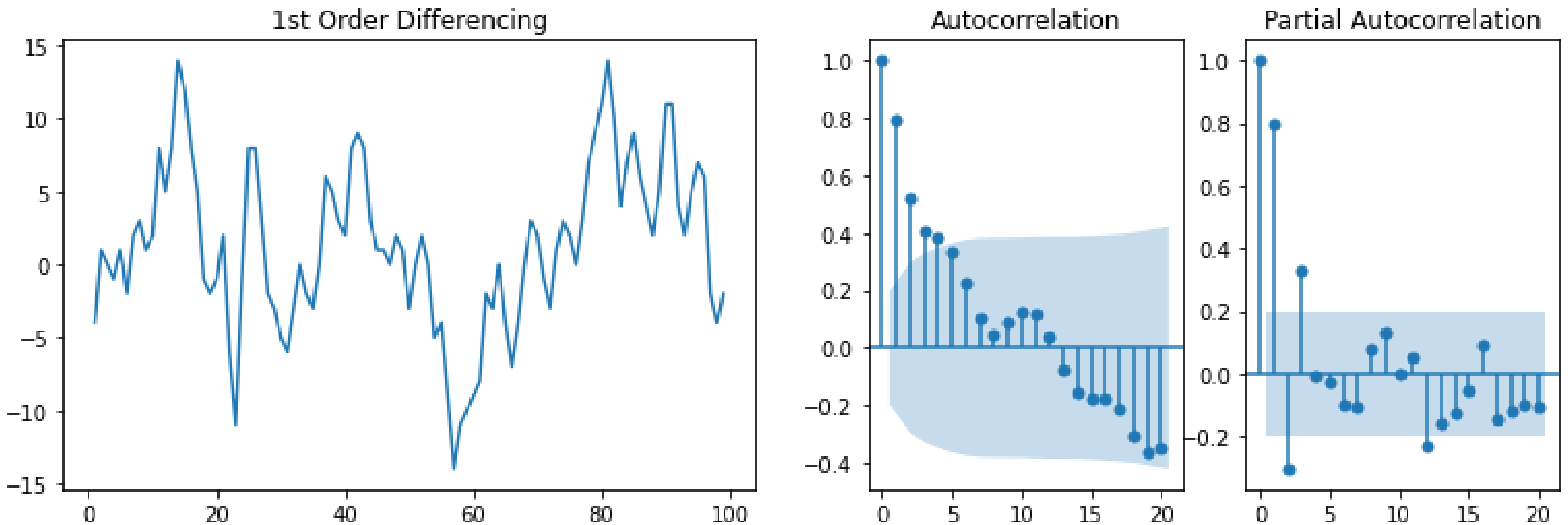
Example: Number of users connected to the Internet through a server every minute; ACF and PACF for the original series



Source: Durbin, J. and Koopman, S. J. (2001) and https://vincentarelbundock.github.io/Rdatasets/datasets.html

ARMA Models: Identification, Estimation, Testing and Selection

Example: First order differencing of the number of users connected to the Internet; ACF and PACF for the differenced series



ARMA Models: Identification, Estimation, Testing and Selection

Model Estimation

We discuss two widely used estimation procedures:

Maximum likelihood estimators (MLE) method. The parameter values of the ARMA model are obtained by minimizing the conditional log-likelihood function

$$ln\, L_*\,(\emptyset, \theta, \sigma_a) = -\frac{n}{2} ln\, 2\,\pi\sigma_a^2 - \frac{S_*(\emptyset, \theta)}{2\sigma_a^2}$$

where $S_*(\emptyset, \theta) = \sum_{t=p+1}^{n} \sigma_a^2(\emptyset, \theta|X)$ is the conditional sum of squares function

Ordinary Least Squares (OLS) method

OLS is the most commonly used regression method in data analysis. However, for ARMA(p,q) models, the OLS estimator will be inconsistent unless we have q=0. For more details, see Wei (2006).

Different software give different estimates...

ARMA Models: Identification, Estimation, Testing and Selection

Example: Model fitted to the number of users connected to the Internet

```
    .... plt.show()
                             ARIMA Model Results
==============================================================================
Dep. Variable:               D.value   No. Observations:                   99
Model:                ARIMA(3, 1, 0)   Log Likelihood                -251.832
Method:                      css-mle   S.D. of innovations              3.056
Date:               Thu, 10 Dec 2020   AIC                            513.665
Time:                       14:47:15   BIC                            526.641
Sample:                            1   HQIC                           518.915

=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.9799      1.650      0.594      0.553      -2.254       4.214
ar.L1.D.value     1.1460      0.095     12.017      0.000       0.959       1.333
ar.L2.D.value    -0.6593      0.135     -4.880      0.000      -0.924      -0.394
ar.L3.D.value     0.3346      0.095      3.532      0.000       0.149       0.520
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1960           -0.0000j            1.1960           -0.0000
AR.2            0.3872           -1.5326j            1.5808           -0.2106
AR.3            0.3872           +1.5326j            1.5808            0.2106
-----------------------------------------------------------------------------
```

ARMA Models: Identification, Estimation, Testing and Selection

Model Testing

Analysis of the quality of parameter estimates. To test the null hypothesis $\mathrm{H_0}: \mathrm{\beta_i} = 0$, we use the test statistic:

$$|t| = \left|\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}\right| > t_{(n-m)} \Rightarrow \text{Reject } H_0: \beta_i = 0$$

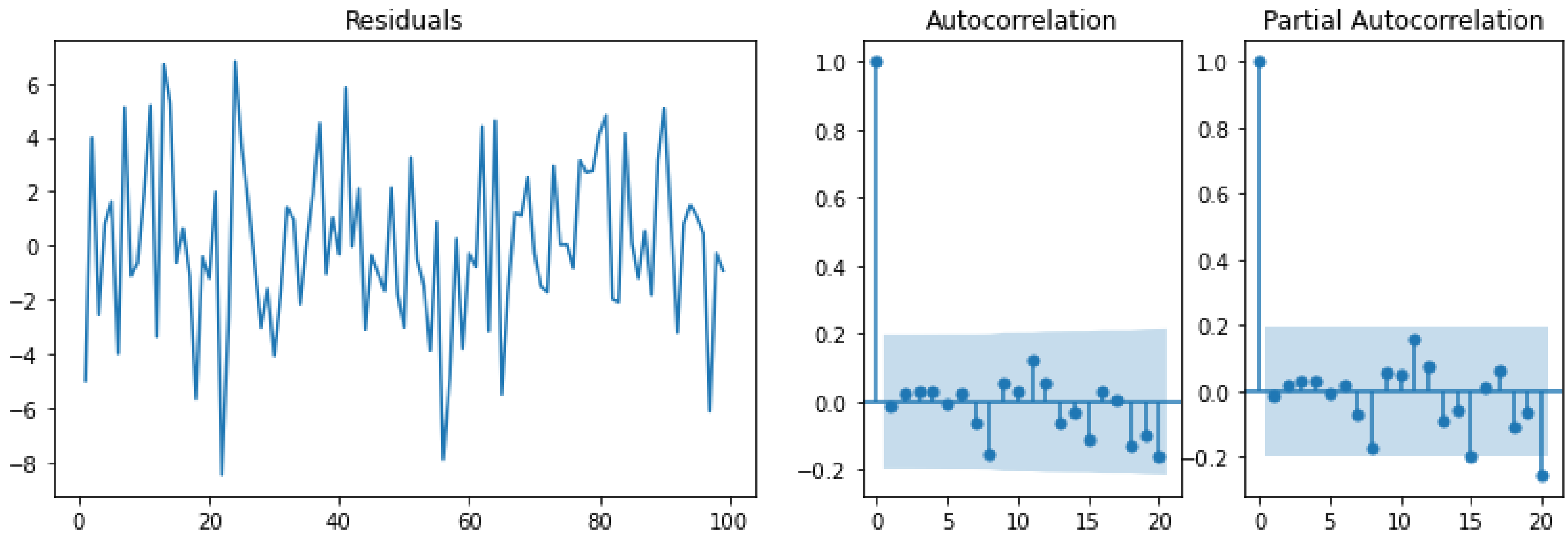
Check whether the residuals are approximately white noise. Compute the sample ACF and sample PACF of the residuals. Tests for residual autocorrelation: H_0: $\rho_k = 0$. If $|\hat{\rho}_k| > \mp\frac{2}{\sqrt{N}} = 0$, Reject H_0. Box and Pierce (1970) introduced a 'portmanteu' test to check the null hypothesis $H_0: \rho_1 = \rho_2 = \cdots = \rho_k = 0$, with the test statistic $Q_k = T\sum_{i=1}^{k} \hat{\rho}_i^2$, which is asymptotically distributed as χ^2 with $k - m$ degrees of freedom, with m the number of estimated parameters. Ljung e Box (1978) proposed a modified version of the statistic Q_k,

$$Q_k^* = n(n+2)\sum_{i=1}^{k} \frac{\hat{\rho}_i^2}{n-i}.$$

This modified form of the 'portmanteu' test statistic is much closer to the $\chi^2(k - m)$ distribution for typical sample sizes n. Thus, if the calculated Q_k^* statistic exceeds the value $\chi^2(k - m)$ then the adequacy of the fitted ARMA model would be questioned.

ARMA Models: Identification, Estimation, Testing and Selection

Example: Residuals from the model fitted to the number of users connected to the Internet; Residual ACF and PACF



ARMA Models: Identification, Estimation, Testing and Selection

Model Selection

Selection criteria are based on summary statistics from residuals, computed from a fitted model (or on forecast errors calculated from out-of-sample forecasts).

Akaike Information Criteria (AIC): Assume that a statistical model of *m* parameters is fitted to a given time series. Akaike (1974) introduced an information criterion defined as

$$\text{AIC} = -2\log L + 2m,$$

where *L* is the maximum likelihood and *n* is the effective number of observations (or number of computed residuals from the series). Some software packages compute the AIC value as

$$\text{AIC} = n\,\widehat{\log \sigma_{\hat{a}}^2} + 2m,$$

where $\hat{\sigma}_{\hat{a}}^2$ is the residual variance for the fitted model.

Schwartz Bayesian criterion (SBC): Schwartz (1978) introduced the following Bayesian criterion of model selection:

$$SBC = n\,\widehat{\log \sigma_{\hat{a}}^2} + m\log n,$$

where $\hat{\sigma}_{\hat{a}}^2$ is the residual variance for the fitted model, *m* is the number of parameters and *n* is the effective number of observations.

ARMA Models: Identification, Estimation, Testing and Selection

Example: Model 1 and Model 2 fitted to the number of users connected to the Internet. Select the “best” model using AIC, BIC and HQIC.

```
     ....: plt.show()
                             ARIMA Model Results
==============================================================================
Dep. Variable:                D.value   No. Observations:                   99
Model:                 ARIMA(3, 1, 0)   Log Likelihood                -251.832
Method:                       css-mle   S.D. of innovations              3.056
Date:                Thu, 10 Dec 2020   AIC                            513.665
Time:                        14:47:15   BIC                            526.641
Sample:                             1   HQIC                           518.915

=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.9799      1.650      0.594      0.553      -2.254       4.214
ar.L1.D.value     1.1460      0.095     12.017      0.000       0.959       1.333
ar.L2.D.value    -0.6593      0.135     -4.880      0.000      -0.924      -0.394
ar.L3.D.value     0.3346      0.095      3.532      0.000       0.149       0.520
```

Model 1

```
     ....: print(model_fit2.summary())
                             ARIMA Model Results
==============================================================================
Dep. Variable:                D.value   No. Observations:                   99
Model:                 ARIMA(0, 1, 3)   Log Likelihood                -255.325
Method:                       css-mle   S.D. of innovations              3.166
Date:                Thu, 10 Dec 2020   AIC                            520.651
Time:                        14:57:33   BIC                            533.626
Sample:                             1   HQIC                           525.901

=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             1.2077      0.937      1.288      0.198      -0.630       3.045
ma.L1.D.value     1.2068      0.101     11.920      0.000       1.008       1.405
ma.L2.D.value     0.6473      0.123      5.281      0.000       0.407       0.888
ma.L3.D.value     0.1201      0.104      1.154      0.248      -0.084       0.324
                                    Roots
```

Model 2

ARMA Models: Identification, Estimation, Testing and Selection

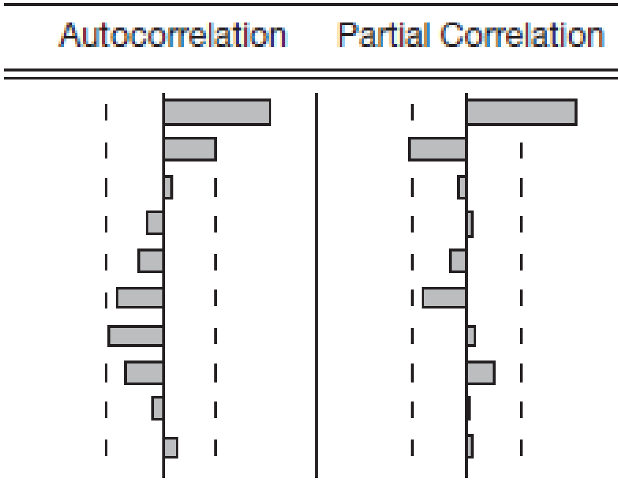
Problems:

1. The sample ACF and PACF were calculated for a series of 30 annual working hours per employee in the US as shown below (Source: Gloria Gonzalez-Rivera, 2013):

a) Test the significance of the sample ACF in the lags 1 and 3

b) Identify potential models for the series

Sample: 1977 2006
Included observations: 30



	AC	PAC	Q-Stat	Prob
1	0.737	0.737	17.987	0.000
2	0.364	-0.392	22.537	0.000
3	0.062	-0.058	22.676	0.000
4	-0.086	0.039	22.951	0.000
5	-0.162	-0.126	23.957	0.000
6	-0.288	-0.295	27.270	0.000
7	-0.352	0.052	32.432	0.000
8	-0.253	0.185	35.229	0.000
9	-0.064	0.001	35.416	0.000
10	0.114	0.034	36.035	0.000

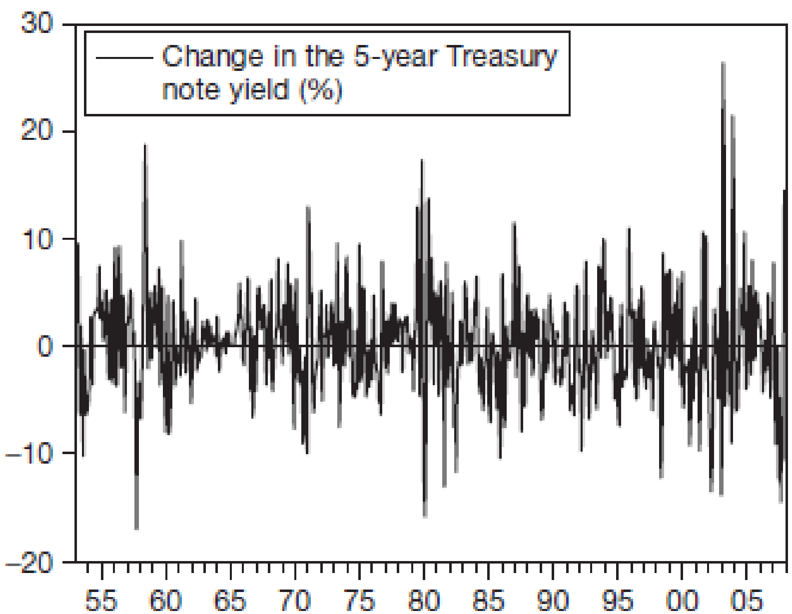
ARMA Models: Identification, Estimation, Testing and Selection

Problems:

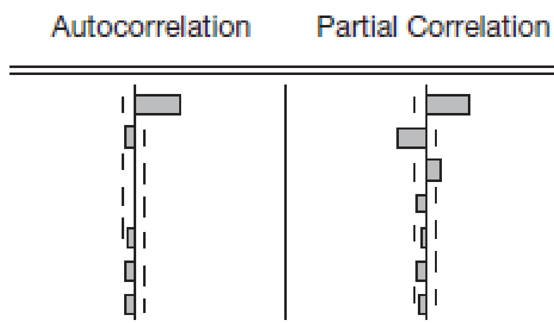
2. The sample ACF and PACF were calculated for a series of "Change in the 5-year Treasury note yield (%) the US" (660 data samples) as shown below (Source: Gloria Gonzalez-Rivera, 2013):

a) Are there any significant ACF and PACF lags?

b) Identify potential models for the series



Sample: 1953M04 2008M04
Included observations: 660



	AC	PAC
1	0.339	0.339
2	-0.073	-0.213
3	0.007	0.129
4	0.014	-0.063
5	-0.043	-0.017
6	-0.073	-0.060
7	-0.069	-0.035

ARMA Models: Identification, Estimation, Testing and Selection

Problems:

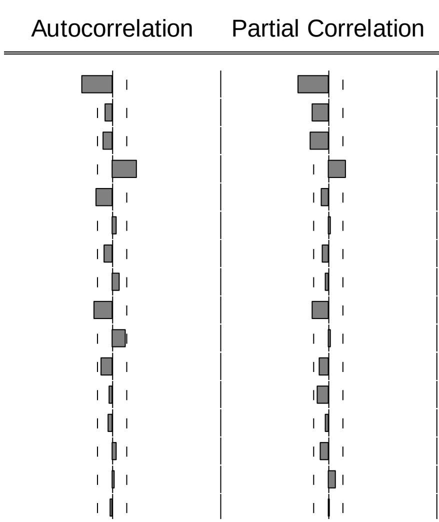
3. Below are the ACF and PACF for the US inflation rate (log differences of CPI).

a) Write down the MA model you feel is appropriate to describe the inflation rate series? Explain your choice

b) Test the null hypothesis that ACF is zero at order 3. Write down the null and alternative hypotheses and explain how you reach a conclusion.

c) Test the null joint hypothesis that the ACF is zero up to order 12. Write down the null and alternative hypotheses and explain in which numbers you base your conclusion.

Sample: 1953Q1 2004Q4
Included observations: 206



	AC	PAC	Q-Stat	Prob
1	-0.286	-0.286	17.055	0.000
2	-0.060	-0.154	17.818	0.000
3	-0.086	-0.168	19.387	0.000
4	0.240	0.171	31.584	0.000
5	-0.154	-0.057	36.659	0.000
6	0.043	0.024	37.059	0.000
7	-0.072	-0.050	38.175	0.000
8	0.066	-0.023	39.128	0.000
9	-0.166	-0.149	45.151	0.000
10	0.135	0.029	49.130	0.000
11	-0.102	-0.074	51.395	0.000
12	-0.025	-0.106	51.534	0.000
13	-0.034	-0.032	51.789	0.000
14	0.047	-0.068	52.274	0.000
15	0.032	0.068	52.509	0.000
16	-0.013	0.006	52.545	0.000