

Classification of Time Series in Frequency Domain

Jianqiang Wang

Department of Statistics

Iowa State University

September 12, 2007

Abstract

Classification of time series in the frequency domain is discussed in this paper. We derived two kinds of classification methods, one is mean vector discrimination using the logarithm of periodogram, and the other one is based on asymptotic results on the real and imaginary part of DFT (Discrete Fourier Transformation). Simulation study has been conducted to compare the performance of different classification methods under different configurations.

1 Introduction

1.1 Classification

Classification is a popular supervised learning technique and it has applications in credit scoring, network intrusion detection, pattern recognition, DNA expression microarray and so forth. We come up with a classification rule based on training data and this classification rule will be applied to new data to predict its class label. Commonly used classification methods include logistic regression, classification and regression tree, neural networks, discriminant analysis, Support Vector Machines and a more recent method called boosting. A detailed description of these methods can be found in Hastie, Tibshirani, and Friedman (2001).

Classification method can be applied to classify different types of objects, including credit applicants in credit scoring, handwriting or facial expression in pattern recognition, or microarray expressions. Usually, a credit applicant can be represented by a

point living in a p -dim space, pixels from graphs like handwriting or facial expression can be sampled and entered into matrices as explanatory variables. Another kind of objects is time series, possibly unequal length and recorded on different time points. The classification of time series have been studied using different classification methods and under various model assumptions.

The classification of time series has applications in various scientific fields. Heart beat dynamics can be used to classify people in different physiological and pathological states (Peng et al. 1999), historical seismic data can be trained to discriminate between earthquakes, and different kinds of explosions (Kakizawa et al. 1998), microarray time series classification has also been explored by many authors.

A time series can be studied in time domain or frequency domain and the parameter we are concerned about in discriminating time series can be trend, seasonal terms or covariance structures. The primary focus of this paper is to classify second-order stationary time series with different covariance structures and our study lives mostly in the frequency domain. Model formulations and theoretical results are presented in Section 2. A simulation study is conducted to compare the performance of different classification methods as well as under different time series models, and the results are presented in Section 3. Section ?? concludes this paper and presents some empirical results I have found in simulation.

2 Method Description

2.1 Background and notations

Consider we have G classes of real second-order-stationary (SOS) time series models $X_1(t), X_2(t), \dots, X_G(t)$ which can be represented as linear processes, with corresponding White Noise series $Z_1(t), Z_2(t), \dots, Z_G(t)$,

$$\begin{aligned} X_1(t) &= \sum_j \psi_{1,j} Z_1(t), \\ &\dots \\ X_G(t) &= \sum_j \psi_{G,j} Z_G(t), \end{aligned}$$

where $Z_g(t) \sim WN(0, \sigma_g^2)$, and $g = 1, 2, \dots, G$.

The ACVF of $X_g(t)$ is,

$$\gamma_g(k) = Cov(X_g(1), X_g(1 + |k|)) = \sum_{j \in \mathcal{Z}} \psi_{g,j} \psi_{g,j+|k|} \sigma_g^2, \forall k \in \mathcal{Z}$$

and the spectral density of $X_g(t)$ is,

$$f_g(\lambda) = \frac{\sigma_g^2}{2\pi} \left| \psi_g(e^{-i\lambda}) \right|^2 \quad (1)$$

for $\lambda \in \Pi$, and $\bar{\Pi} = [-\pi, \pi]$.

Suppose we have m_g realized time series of length n under each time series model $X_g(t)$, denoted as $X_{g,1}(t), X_{g,2}(t), \dots, X_{g,m_g}(t)$.

Denote the Discrete Fourier Transformation of $X_{g,l}(t)$ at frequency λ as $d_{n,g,l}(\lambda)$,

$$d_{n,g,l}(\lambda) = \frac{1}{\sqrt{n}} \sum_{t=1}^n X_{g,l}(t) e^{i\lambda t},$$

where $\lambda \in \Pi$.

Write the periodogram of $X_{g,l}(t)$ at frequency λ as $I_{n,g,l}(\lambda)$,

$$I_{n,g,l}(\lambda) = |d_{n,g,l}(\lambda)|^2 = \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n X_{g,l}(t) e^{i\lambda t} \right|^2.$$

2.2 Discriminant analysis

We further assume $Z_g(t) \sim IID(0, \sigma_g^2)$, and $\sum_{j \in \mathcal{Z}} |\psi_{g,j}| < \infty$ for all $g = 1, 2, \dots, G$. By Theorem 10.3.2 of Brockwell and Davis (2002), if $f_g(\lambda) > 0$, $\lambda \in [-\pi, \pi]$, and $0 < \lambda_1 < \dots < \lambda_m < \pi$,

$$(I_{n,g,l}(\lambda_1), \dots, I_{n,g,l}(\lambda_m))^T \xrightarrow{d} (U_{g,1}, \dots, U_{g,m})^T, \quad (2)$$

where $U_{g,i}$ are independently distributed as $\exp(2\pi f_g(\lambda_i))$.

By Continuous Mapping Theorem,

$$\mathbf{Y}_{n,g,l} \hat{=} \begin{pmatrix} \log I_{n,g,l}(\lambda_1) \\ \log I_{n,g,l}(\lambda_2) \\ \dots \\ \log I_{n,g,l}(\lambda_m) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \log 2\pi f_g(\lambda_1) \\ \log 2\pi f_g(\lambda_2) \\ \dots \\ \log 2\pi f_g(\lambda_m) \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \dots \\ U_m \end{pmatrix} \quad (3)$$

where U_i are IID $\log \exp(1)$ random variables.

So $\mathbf{Y}_{n,g,l}$ have the same mean vector in the same class, and usual discriminant analysis methods can be applied to classify $\mathbf{Y}_{n,g,l}$, like linear discriminant analysis and quadratic discriminant analysis.

2.3 Likelihood-based classification

Theoretical results have been given by Brockwell and Davis (2002), Fuller (1996) concerning the asymptotic behavior of $d_{n,g,l}(\lambda)$. Under certain regularity conditions,

$$\mathbf{W}_{n,g,l} \hat{\stackrel{d}{\rightarrow}} N_{2m}(\mathbf{0}, \Sigma_{2m}) \quad (4)$$

$$\mathbf{W}_{n,g,l} \hat{\stackrel{d}{\rightarrow}} \begin{pmatrix} \text{Re}(d_{n,g,l}(\lambda_1)) \\ \text{Im}(d_{n,g,l}(\lambda_1)) \\ \dots \\ \text{Re}(d_{n,g,l}(\lambda_m)) \\ \text{Im}(d_{n,g,l}(\lambda_m)) \end{pmatrix}$$

where $\Sigma_{2m,g} = \text{Diag}\left(\frac{f_g(\lambda_1)}{2}, \frac{f_g(\lambda_1)}{2}, \dots, \frac{f_g(\lambda_m)}{2}, \frac{f_g(\lambda_m)}{2}\right)$.

So $\Sigma_{2m,g}^{-1/2} \mathbf{W}_{n,g,l}$ converges in distribution to a multivariate normal distribution with identity variance-covariance matrix. Let $n_{2m,g}(\cdot)$ be the density of $2m$ -variate MVN distribution, then we can use a likelihood-based approach to discriminate between different classes.

Let $X_k(t)$ denote a new time series which we want to classify, and $W_{n,k}$ defined as in Equation 4. We will calculate $\pi_g n_{2m,g}(\Sigma_{2m,g}^{-1/2} \mathbf{W}_{n,k})$ for each class $g = 1, 2, \dots, G$ and classify $X_k(t)$ into a class with the highest $\pi_g n_{2m,g}(\Sigma_{2m,g}^{-1/2} \mathbf{W}_{n,k})$. Usually, $\Sigma_{2m,g}$ is unknown and will be replaced by $\hat{\Sigma}_{2m,g}$ where we replace true spectral densities with their parametric estimates.

3 Simulation Study

In my simulation study, we have 2 classes of real SOS time series model, both of which are ARMA models, written as

$$\begin{aligned} \phi_1(B)X_1(t) &= \theta_1(B)Z_1(t), \\ \phi_2(B)X_2(t) &= \theta_2(B)Z_2(t), \end{aligned}$$

where $\phi_1(B) = 1 - 0.5B$, $\theta_1(B) = 1 + 0.75B$,
 $\phi_2(B) = 1 - 1.65B + B^2 - 0.2625B^3 + 0.025B^4$, $\theta_2(B) = 1 - 0.2B - 0.24B^2$, $Z_1(t) \sim WN(0, 0.4)$ and $Z_2(t) \sim WN(0, 0.4)$.

We randomly generate 175 time series from class I and 125 series of length 200, and plot one from each classes in Figure 1. The Autocovariance functions are plotted in Figure 2, from which we can see that time series from the second class have a longer memory than those from class II.

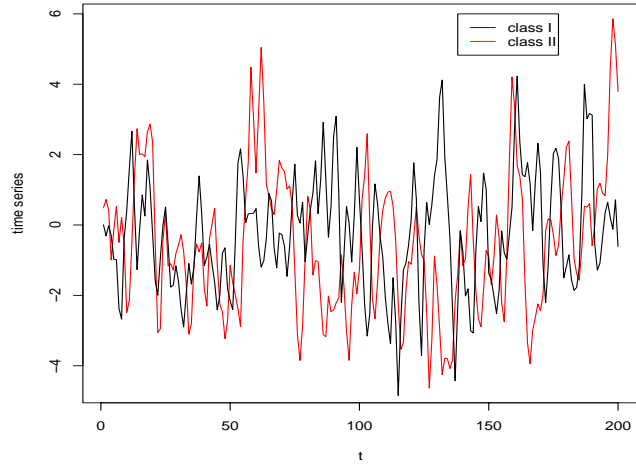


Figure 1: Time series plot using one time series from each class

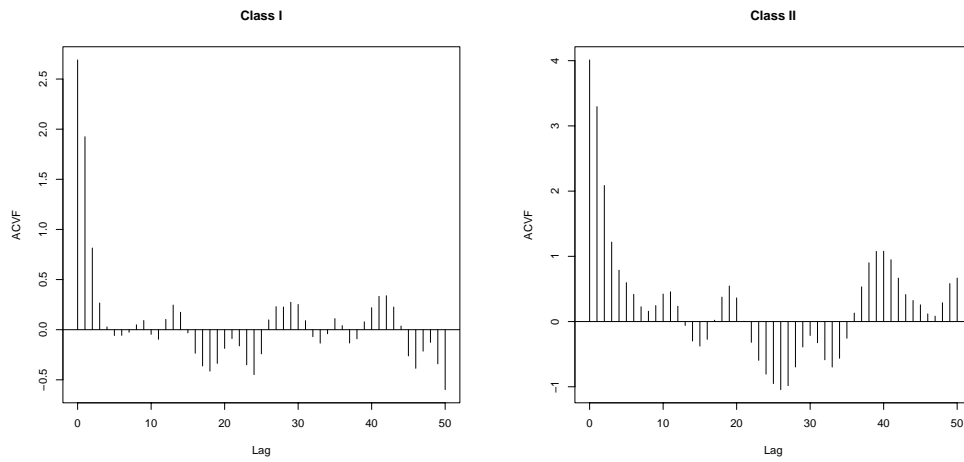


Figure 2: Autocovariance function based on one time series from each class.

The true spectral densities calculated using model parameters are plotted in Figure 3. We can see that the spectral density of class I is unimodal and that of class II is bimodal.

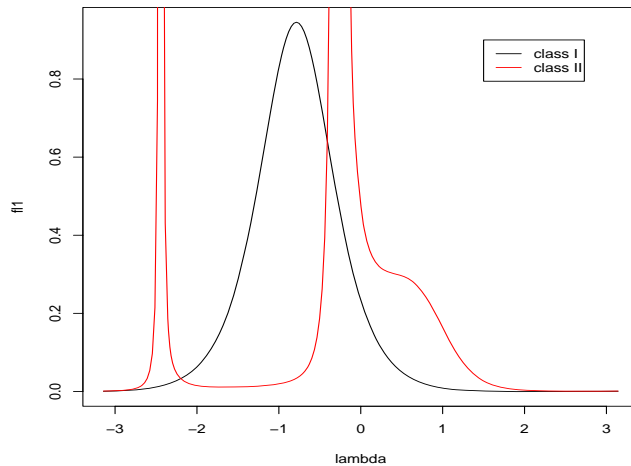


Figure 3: Spectral densities calculated using model parameters

We have 400 new time series generated from model I or model II, then we use each of the three classification methods to classify those 400 new series. We choose $m = 10$ frequencies satisfying $f_g(\lambda_i) > 0.02$, so as to avoid frequencies on which the spectral density is 0. The cross-tabulated misclassifications are listed in Table 1 using priors proportional to fractions in training data, and Table 2 shows the results with equal priors. In our prediction data, the proportion of time series from class I and those from class II are close to 1/2, and we can see by comparing Table 1 and Table 2 that all three classification methods do better using equal priors than prior probabilities proportional to training set. LDA is doing better than both QDA and likelihood classification, and there is a serious problem classifying time series into class I using both QDA and likelihood method.

In order to quantify the overall performance of each classification method and their stability, we calculated the mean, standard deviation of each misclassification rate using 20 simulations under each classification method and listed the results in Table 3. We can see from Table 3 that the misclassification rate is smallest using LDA, followed by QDA and likelihood classification has the worse accuracy. In terms of method stability, likelihood method is the most stable, followed by LDA and QDA is the most unstable.

	LDA		QDA		Likelihood	
True	class I	class II	class I	class II	class I	class II
I	204	14	183	35	214	4
II	90	92	160	22	172	10

Table 1: Misclassification of different methods on test data, prior proportional to the fraction in training data

	LDA		QDA		Likelihood	
True	class I	class II	class I	class II	class I	class II
I	123	84	124	83	156	51
II	43	150	90	103	135	58

Table 2: Misclassification of different methods on test data, equal prior

	LDA	QDA	Likelihood
mean	0.2776	0.3593	0.4656
standard deviation	0.0375	0.0663	0.0271

Table 3: Mean, variance and C.V of misclassification rate under each method

In the previous simulation, σ_g^2 is held constant at 0.4, and we are interested in the effects of σ_g^2 on misclassification rates. So we increase σ_g^2 from 0.1 to 2, and want to examine the trend of misclassification rate when increasing σ_g^2 . Figure 4 plots misclassification rates against σ_g^2 . From the plot, we can see misclassification rate increases when σ_g^2 gets larger in LDA and QDA, but the change is not as obvious using likelihood classification. As a result, when σ_g^2 increases, the performance of three classification methods will be similar regarding misclassification rates.

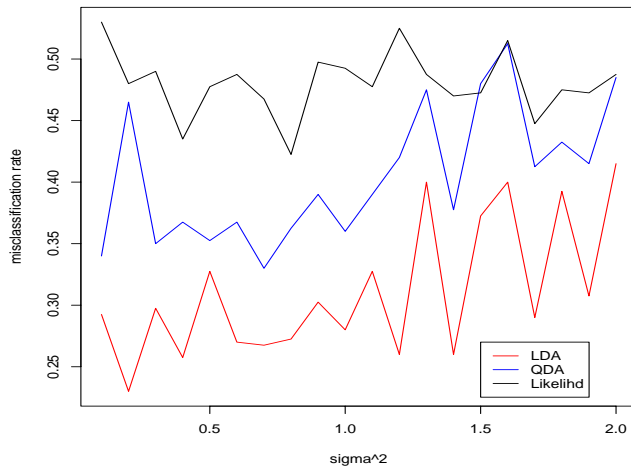


Figure 4: The changes of misclassification rates as we increase σ_g^2

We are also interested in the number of dimensions m and its effects on classification. From Figure 5, we can see that LDA will do even better if you increase the number of frequencies, and the misclassification rate is around 0.1 when we use 40 frequencies to define $\mathbf{Y}_{n,g,l}$. The misclassification rate does not change much in QDA or likelihood classification method, and if we increase the number of dimensions of $\mathbf{Y}_{n,g,l}$ to 40, likelihood method will give an even higher misclassification rate.

Figure 6 shows the trend of misclassification rates as we increase the length of each time series. We can not see a clear pattern in misclassification rates with the length of time series, but I would expect the variation in misclassification rates to decrease as we increase the length of time series because of convergence.

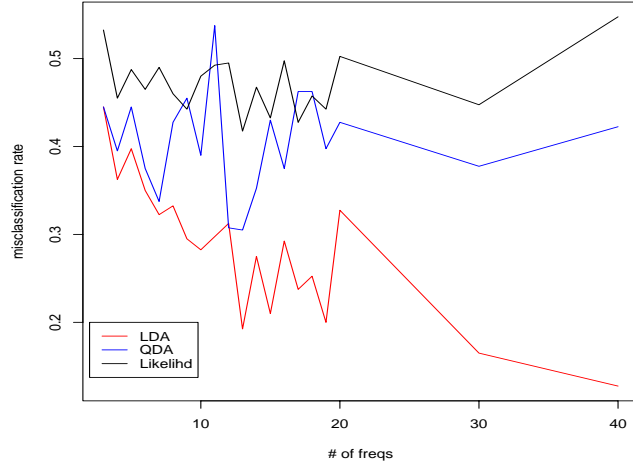


Figure 5: The changes of misclassification rates as we increase the number of frequencies

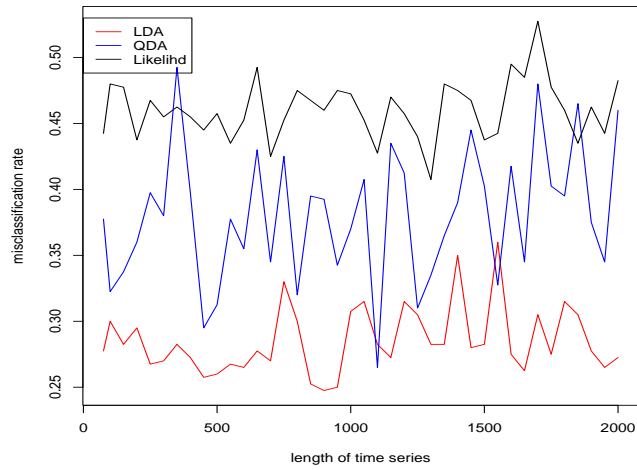


Figure 6: The changes of misclassification rates as we increase the length of each time series

4 Conclusion

1. Misclassification rate: LDA $\hat{}$ QDA $\hat{}$ Likelihood-based
2. Variation of Misclassification rate: Likelihood-based $\hat{}$ LDA $\hat{}$ QDA
3. The increase of σ_g^2 will decrease misclassification rates, but the effect varies between classification methods
4. The effect of length of each time series on misclassification isn't significant

5 Acknowledgement

I want to express my gratitude to Dr S.N. Lahiri for guiding me through this project and suggesting various ideas. I highly appreciate his patience and help during the process of working on this project.

References

- Brockwell, P. and R. Davis (2002). *Introduction to Time Series and Forecasting*. Springer-Verlag, New York.
- Fuller, W. A. (1996). *Introduction to Statistical Time Series*. John Wiley & Sons.
- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer.
- Kakizawa, Y., R. H. Shumway, and M. Taniguchi (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association* 93, 328–340.
- Peng, C.-K., J. Mietus, Y. Liu, G. Khalsa, P. Douglas, H. Benson, and A. Goldberger (1999). Exaggerated heart rate oscillations during two meditation techniques. *International Journal of Cardiology* 70, 101–107.