9.073/HST.460 Statistics for Neuroscience Research
Emery N. Brown

Lecture 16: Introduction to Time-Series Analysis

I. Objectives

Understand the theoretical framework for stationary time-series analysis

Understand method-of-moments analysis of time-series

Understand parametric likelihood analyses of time-series

Understand the strategy for systematically selecting, analyzing and making an inference from a time-series model

II. Time-Series Analysis

A. Motivation

Example 3.2 MEG Background Noise Analysis (continued). Our analysis of these data in Lecture 3 showed that they are well described by a Gaussian distribution (**Figure 16.1**). These time-series plots suggest that there may be temporal structure in these data. Is there temporal structure in the individual sensor time-series? How can we quantify this temporal structure if it is present? Is there a similar type of temporal structure in the back sensor and the front sensor? Could this temporal structure reflect systematic inhomogeneities in the background magnetic field?

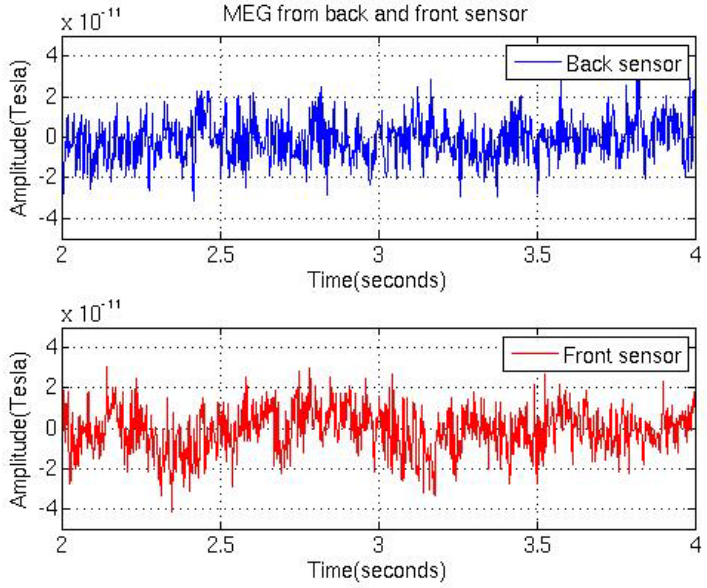
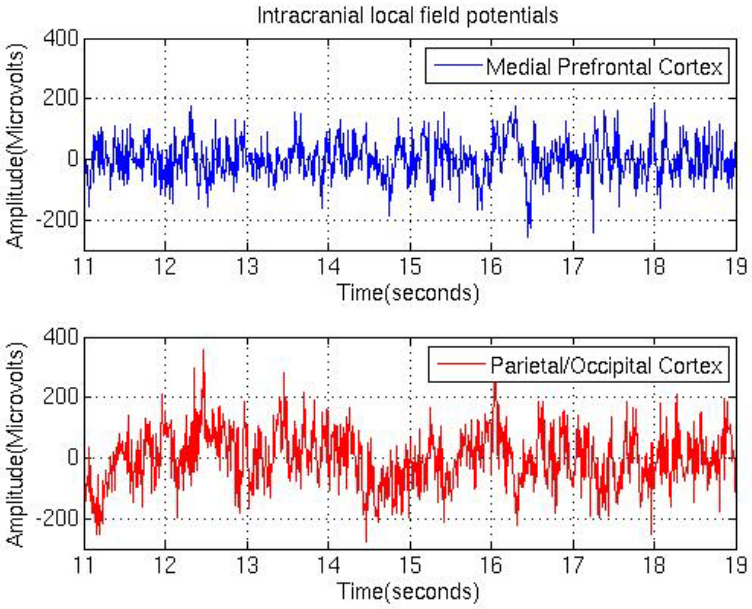


Figure 16.1 Four seconds of background noise recordings from a front and back sensor in the MEG scanner.

Example 16.1 EEG and LFP's from a Rat Under General Anesthesia. The objective of this study is to develop neural correlates of the states of general anesthesia. A rat has epidural screws as well as tetrode arrays implanted in four sites in a rostral caudal configuration. Local field potentials and epidural EEGs are measured as the animal is placed at increasing doses of anesthesia. The data in the current plots are epidural EEG recordings made on the animal prior

to giving the first dose of the anesthetic drug while the animal is sitting still and not executing any particular behavior (**Figure 16.2**). The intracranial local field potentials were recorded simultaneously from a prefrontal and an occiptal/parietal site. These time-series plot shows obvious temporal structure. How can we quantify this temporal structure? Is this temporal structure harmonic in nature? That is, are there oscillations in known frequencies such as theta, alpha and or gamma? Is there a similar type of temporal structure across different recording sites?



Example 16.1. Nine-seconds of local field potential recordings from a prefrontal and a parietal/occipital intracranial electrode in a rat at baseline prior to receiving an anesthetic drug.

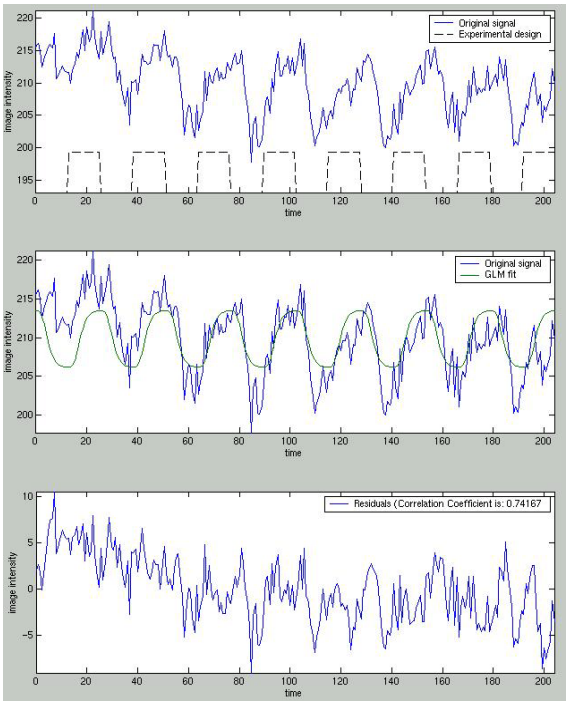


Figure 16.3. Plots of fMRI Time-Series, motor stimulus, and fit of hemodynamic response plus white model and residuals.

Example 15.3 fMRI Visual-Motor Stimulation Experiment (continued). Our analysis of these data in **Lecture 15** (Class 7-8) showed that they are well described by a hemodynamic plus drift model with independent, identically distributed (white) Gaussian noise. The time-series analyzed in the Class 7-8 had much more structure in the residuals (**Figure 16.3**). Can we develop a correlated noise model that would be more appropriate for the non-hemondynamic and non-drift components in these fMRI time-series?

B. Definitions

1. A **stochastic process** is a family, or collection, series of random variables, typically defined on the same outcome space and indexed by time or space. If we take the index to be time, then there are 4 types of stochastic processes we can define. To do so, we call the value of the random variable the value or state. For our purposes, the value or state of the random variable can be either discrete or continuous. Similarly, time can be either discrete or continuous.

State		Discrete	Continuous
	Continuous	**Digital Signals** **LFP's** **EEG** x_t **MEG** **Body Temperature** **Plasma Hormone Level**	**Analog Signals** **LFP's** **EEG** $x(t)$ **MEG** **Body Temperature** **Plasma Hormone Level**
	Discrete	**Spike Counts** **Binary Behavior Response** n_t	**Point Processes** **Spike Trains** **Heartbeats** $N(t)$

In this lecture, we will consider time-series (stochastic processes) that have continuous-valued states and are observed in discrete time.

2. In **Lecture 3**, we defined for a random X with probability density $f(x)$ the j^{th} moment of X as

$$\mu_j = E(X^j) = \int x^j f(x) dx$$

and the j^{th} sample moment as

$$m_j = n^{-1} \sum_{i=1}^{n} x_i{}^j.$$

The sample mean is $\bar{X} = m_1$ and the sample variance is $\hat{\sigma}^2 = m_2 - m_1^2$.

3. For two random variables X and Y with a joint probability density $f(X,Y)$ we have the marginal moments

$$\mu_{x,j} = E(X^j)$$

$$\mu_{y,j} = E(Y^j)$$

the cross-moment (covariance)

$$\text{cov}(X,Y) = E(X-\mu_x)(Y-\mu_y) = \int\int (x-\mu_x)(y-\mu_y) f(x,y)dxdy = \sigma_{xy}$$

the correlation coefficient is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

When we studied the simple linear regression model in **Lecture 14** we saw the sample versions of these quantities namely,

the sample covariance is

$$s_{xy} = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

the sample correlation coefficient is

$$r_{xy} = \frac{s_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}.$$

We need to define similar quantities for a time-series where there is a different probability density at each time point.

4. For a time-series $X_1,...,X_n,$ we assume $E(X_t) = \mu$ for all $t.$ We define the

a. **Autocovariance function** as

$$\gamma(h) = Cov(X_t, X_{t+h}) = E[X_t - \mu)(X_{t+h} - \mu)$$

for $h = 0,1,2,...,$

b. **Autocorrelation function** is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

for $h = 1, 2, ...,$

Remark 1. The autocovariance function has the following properties

i) $\gamma(0) \geq 0$ (variance of X_t)

ii) $|\gamma(h)| \leq \gamma(0)$ for all h

iii) $\gamma(h) = \gamma(-h)$ for all h

For these definitions to be useful in a practical sense, we need some additional definitions.

A time-series is **second-order (weakly) stationary** if for all t and τ $Cov(X_{t+h}, X_t) = Cov(X_{t+h+\tau}, X_{t+\tau})$. That is, the covariance only depends on the separation h between the values and not the time t.

A time-series is **strictly stationary** if $\{X_1, ..., X_n\}$ and $\{X_{1+h}, ..., X_{n+h}\}$ have the same distribution for all h.

Remark 2. Strict stationarity implies weak stationarity but not vice-versa, necessarily.

Remark 3. A sequence of independent, identically distributed observations is strictly stationary.

Remark 4. A weakly stationary Gaussian process is strictly stationary because a Gaussian process is completely characterized by its first two moments and cross moments.

Remark 5. In practice, we can show that stationarity does not hold. It is harder (impossible) to show that it holds.

Example 16.2. A first-order autoregressive process AR(1) is defined as

$$x_t = \rho x_{t-1} + \varepsilon_t$$

where we typically assume $|\rho| < 1,$ to insure stationarity and the ε_t are independently, identically distributed Gaussian random variables with zero mean and variance σ^2. We have $E(X_t) = 0$ for all t. It can be shown that

$$\gamma(h) = \gamma^{|h|}\gamma(0)$$

$$\gamma(0) = (1 - \rho^2)^{-1}\sigma^2.$$

This process is stationary and a plot of $\gamma(h)$ as a function of h shows geometric (exponential) decay. What happens if $|\rho|>1$?

5. Just as we did for observations from a sample probability density we can define for a time-series sample moment estimates of the mean, autocovariance function and the autocorrelation function. These are defined as follows. Let $x_1,...,x_n$ be observations from a time-series. The **sample mean** is

$$\overline{x} = n^{-1}\sum_{t=1}^{n} x_t.$$

The **sample autocovariance** is

$$\hat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-|h|}(x_{t+|h|}-\overline{x})(x_t-\overline{x})$$

for $-n<h<n.$

The **sample autocorrelation function** is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

for $-n<h<n.$

Remark 6. These quantities are the **method-of-moments estimates** of the mean autocovariance and autocorrelation functions because we are estimating the theoretical moment by the sample moment. Dividing by n, rather than $n-h$ is important to insure that the sample covariance function is nonnegative definite.

C. Properties of the Method-of-Moments Estimates

1. Sample Mean

It can be shown that

$$E(\overline{x}) = \mu$$

$$Var(\overline{x}) = n^{-1}\sum_{h=-n}^{n}(1-\frac{|h|}{n})\gamma(h)$$

and that

$$n^{\frac{1}{2}}(\overline{x}-\mu) \sim N(0, \sum_{h=-n}^{n}\left[(1-\frac{|h|}{n})\gamma(h)\right]).$$

It follows that an approximate 95% confidence interval for μ is

$$\overline{x} \pm \frac{1.96\hat{v}}{n^{\frac{1}{2}}}$$

where $\hat{v} = [\sum_{n=-h}^{h}(1-\frac{|h|}{n}\hat{\gamma}(h)]^{\frac{1}{2}}$.

Example 16.2 (continued). Suppose the AR(1) process is now

$$x_t - \mu = \rho(x_{t-1}-\mu) + \varepsilon_t$$

then

$$\gamma(h) = (1-\rho^2)^{-1}\sigma^2\rho^{|h|}$$

$$v = \frac{\sigma^2}{(1-\rho)^2}$$

and the 95% confidence interval for μ is

$$\overline{x} \pm \frac{1.96\sigma}{(1-\rho)n^{\frac{1}{2}}}.$$

Remark 7. The quantity $(1-\rho)^2 n$ is sometimes referred to as the **number of equivalent independent observations**. Note that when $\rho = 0$ we have our standard formula. In most cases, we will remove the mean and work with the zero mean data series. An important exception is when the mean process is not a simple function as in the case of the fMRI time-series analysis in **Example 15.3**. Therefore, μ is time-varying and defined by

$$\mu = \mu_t = h(t) = \int_0^t c(t-u)g(u)du.$$

where $h(t)$ is the physiological response, $c(t)$ is the input stimulus and $g(t)$ is the hemodynamic response.

2. The Sample Autocovariance and Autocorrelation Functions

Both $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ are biased. That is,

$$E[\hat{\gamma}(h)] = \gamma(h) + \frac{\text{terms}}{n}$$

$$E[\hat{\rho}(h)] = \rho(h) + \frac{\text{terms}}{n}$$

We can define the sample covariance matrix up to lag k as

$$\hat{\Gamma}_k = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(\text{k-1}) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(\text{k-2}) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\gamma}(\text{k-1}) & \hat{\gamma}(\text{k-2}) & \cdots & \hat{\gamma}(0) \end{bmatrix}$$

and the sample autocorrelation function up to lag k as

$$\hat{R}_k = \frac{\hat{\Gamma}_k}{\hat{\gamma}(0)}.$$

The approximate distribution of the $\hat{\gamma}(h)$ has been well worked out. For practical purposes we are more interested in the distribution of the $\hat{\rho}(h)$. If we define $\hat{\rho}_k = (\rho(0),...,\hat{\rho}(k))$ then it can be shown that

$$\hat{\rho}_k \sim N(\rho_k, n^{-1}W)$$

where W is a matrix whose ij^{th} element is

$$W_{ij} = \sum_{k=1}^{\infty} [\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)][\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)].$$

Example 12.3. If the X_t independent, identically distributed Gaussian with mean 0 and variance σ^2, $\rho(h) = 0$ for all $|h| > 0$, and hence $W_{ij} = 1$ if $i = j$ and 0 otherwise. For large n $\hat{\rho}(1),...,\hat{\rho}(h)$ are independent, identically distributed with mean 0 and variance n^{-1}. This is a key result we will use to estimate the approximate confidence intervals for the autocorrelation function (acf) in order to understand the nature of the time dependence.

Example 12.2 (continued). For the AR(1) process we have

$$W_{ii} = (1-\rho^{2i})(1+\rho^2)(1-\rho^2)^{-1} - 2i\rho^{2i}$$

and the approximate 95% confidence intervals of

$$\hat{\rho}(i) \pm \frac{1.96 W_{ii}^{1/2}}{n^{\frac{1}{2}}}.$$

D. Autoregressive Models of Order P $(AR(p))$.

1. Model Formulation

As an alternative to simply working with the sample acf to analyze a time-series, we will consider a model based approach using autoregressive models of order p defined as

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + \varepsilon_t$$

where we assume

i) ε_t are i.i.d. Gaussian with mean 0 and variance σ^2.

ii) the polynomial defined as

$$\phi(z) = 1 - \sum_{j=1}^{p} z^j \alpha_j$$

has all its roots outside the unit circle.

This later condition insures stationarity. For the AR(1) model this requirement simplifies to $|\alpha_1| = |\rho| < 1$.

2. Conditional Likelihood Analysis

We want to derive the likelihood in order to estimate $\alpha = (\alpha_1, ..., \alpha_p)$ so we require the joint probability density of $X_1, ..., X_n$. To do this note that we can express the joint probability density as

$$f(X_1, ..., X_n) = f(X_{p+1}, ..., X_n \mid X_1, ..., X_p) f(X_1, ..., X_p)$$

$$= \prod_{t=p+1}^{n} f(X_t \mid X_{t-p}, ..., X_{t-1}) f(X_1, ..., X_p)$$

$$= \prod_{t=p+1}^{n} f(X_t \mid X_{t-p}, ..., X_{t-1}) f(X_1, ..., X_p)$$

where the second step above is possible because the dependence of X_t on history goes back not to the beginning but only p lags by the definition of X_t. Let us further assume that $n >> p$ so that conditioning on the first p observations, i.e. assuming they are known is not a substantial loss. We have then that

$$f(X_1, ..., X_n) \approx \prod_{t=1}^{n} f(X_t \mid X_{t-p}, ..., X_{t-1}) = f(X_{p+1}, ..., X_n \mid X_1, ..., X_p).$$

To derive this joint conditional density we have by definition and the change-of-variable formula (**Lecture 4)**

$$f(X_{p+1},...,X_n \mid X_1,...,X_p) = f(\varepsilon_{p+1},...,\varepsilon_n) \mid J \mid$$

where J is the Jacobian of the transformation. We have by the definition of the ε_t's

$$f(\varepsilon_{p+1},...,\varepsilon_n) = \prod_{t=p+1}^{n} (\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}\varepsilon_t^2\}$$

$$= (2\pi\sigma^2)^{-\frac{n-(p+1)}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\varepsilon_t^2\}.$$

Now

$$\varepsilon_t = X_t - \sum_{j=1}^{p}\alpha_j X_{t-j}$$

and J is lower triangular so its determinant $\mid J \mid = \prod_{t=p+1}^{n} \mid J_{tt} \mid$. We have $J_{tt} = \frac{\partial \varepsilon_t}{\partial x_t} = 1$ for all t. Hence, $\mid J \mid = 1$. Substituting for ε_t in terms of $X_t, X_{t-1}, \ldots, X_{t-p}$ in $f(\varepsilon_{p+1},...,\varepsilon_n)$ gives

$$f(X_{p+1},...,X_n) \mid X_1,...,X_p) = (2\pi\sigma^2)^{-\frac{[n-(p+1)]}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{t=p+1}^{n}(X_t - \sum_{j=1}^{p}\alpha_j X_{t-j})^2\}.$$

Therefore,

$$L(\alpha,\sigma^2 \mid x_1,...x_n) \approx f(x_{p+1},...,x_n,\alpha,\sigma^2)$$

is the **conditional likelihood** of α and σ^2. Based on the results in **Lecture 15** we have a linear model in matrix notation of the form

$$Y = X\alpha + \varepsilon$$

where

$$\begin{bmatrix} x_{p+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_p & \cdots & x_1 \\ \vdots & & \vdots \\ x_{n-1} & & x_{n-p} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} + \begin{bmatrix} \varepsilon_{p+1} \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

It follows from the results in **Lecture 15** that the approximate (conditional) maximum likelihood estimates of α and σ^2 are

$$\hat{\alpha} = (X^T X)^{-1} X^T Y$$

$$\hat{\sigma}^2 = [n-(p+1)]^{-1} \sum_{t=p+1}^{n} (x_t - \sum_{j=1}^{p} \hat{\alpha}_j x_{t-j})^2.$$

Remark 8. The $AR(p)$ model is a p^{th} order Markov process. That is, the probability density at time t depends not on the full history put only on the previous p observations.

Remark 9. Nothing we have done insure that the α_j's satisfy the stationary conditions. This is usually not an issue when n is large and $n >> p$. It is possible to carry out the estimation so that the α_j's are stationary.

Remark 10. When n is large relative to p, we can show that the (conditional) maximum likelihood solution can be computed as

$$\hat{\alpha} = \hat{R}_p^{-1} \hat{\rho}_p$$

where $\hat{R}_p$ and defined above. These equations are the well-known **Yule-Walker equations** and give the method-of-moments estimate of $\hat{\alpha}$. It is no surprise that the method-of-moments and maximum likelihood agree for Gaussian processes. This result follows because for large n

$$(X^T X) \approx n\hat{R}_p.$$

The Yule-Walker equations are one of the two options for estimating AR models in Matlab. The other is the Burg algorithm mentioned below.

Remark 11. Because $\hat{\alpha}$ is an approximate ML estimate we have from **Lecture 9** that

$$\hat{\alpha} \sim N(\alpha, -I(\alpha)^{-1})$$

where $I(\alpha)$ is the Fisher information for α. It is straightforward to show that the approximate observed Fisher information is

$$-\hat{I}_n(\alpha)^{-1} = \sigma^2 (X^T X)^{-1} \approx (n\hat{R})^{-1}$$

for n large. Hence, we can use likelihood theory from **Lectures 9** and **15** to compute approximate 95% confidence intervals for α_j as

$$\hat{\alpha}_j \pm 1.96\hat{\sigma}[(X^T X)^{-1}_{jj}]^{\frac{1}{2}}$$

where $(X^T X)^{-1}_{jj}$ is the j^{th} diagonal element of $(X^T X)^{-1}$ for $j = 1,..., p.$

Remark 12. Additional information about an $AR(p)$ model can be had by examining the **partial autocorrelation coefficients** or **partial autocorrelation function (pacf)**. Let $\pi_k = -\alpha_k,$ where α_k is the last coefficient in an $AR(k)$ model. Then π_k is partial autocorrelation coefficient between x_t and x_{t+k} holding $x_{t+1},...,x_{t+k-1}$ fixed. If the true model is $AR(m),$ the $\pi_m = 0$ for $k > m$ and the partial autocorrelation function vanishes after a finite number of terms. Hence, we can estimate $\hat{\pi}_k$ as $-\hat{\alpha}_k$ where $\hat{\alpha}_k$ is the estimate of the last coefficient for an $AR(k)$ process. Now a plot $\hat{\pi}_k$ versus k should indicate the true order of the model.

We have that for an $AR(k)$ process the π_m is $\pm 1.96 n^{-\frac{1}{2}}$. If we let α_j, k be the j^{th} coefficient in an $AR(k)$ model for $j = 1,..., k,$ then the **Levinson-Durbin formula** for computing the partial autocorrelation coefficients is

$$\hat{\alpha}_{k+1,j} = \hat{\alpha}_{k,j} - \hat{\alpha}_{k+1,k+1}\hat{\alpha}_{k,k-j+1}$$

$$\hat{\alpha}_{k+1,k+1} = \frac{\hat{\rho}^{(k+1)} - \sum_{j=1}^{k} \alpha_{k,j}\hat{\rho}^{(k+1-j)}}{1 - \sum_{j=1}^{k} \hat{\alpha}_{m,j}\hat{\rho}^{(j)}}.$$

N.B. $\hat{\alpha}_k$ is often called the **reflection coefficient**. This is the term used in Matlab.

Remark 13. We can use the **AIC** to help identify the model order. It is

$$\text{AIC}(p) = n \log(\hat{\sigma}^2_{\text{ML},p}) + 2p$$

as we discussed for the multiple linear regression analyses in **Lecture 15**. Here $\hat{\sigma}^2_{\text{ML},p}$ is the ML estimate of σ^2 from the $AR(p)$ model.

Remark 14. $AR(p)$ processes are a special case of **autoregressive moving average processes** $(ARMA(p,q))$. They are defined as

$$X_t - \sum_{j=1}^{p} \alpha_j X_{t-j} = \sum_{k=1}^{q} \beta_k \varepsilon_{t-k} + \varepsilon_t.$$

These processes are discussed in detail in nearly every book on time-series analysis. An important feature of an $ARMA$ process is that it has a representation as an infinite order AR process. It has been shown empirically that on $ARMA(1)$ model provides an accurate description of physiological and scanner noise in an fMRI experiment (Purdon et al., 2001).

Remark 15. The Burg algorithm provides another approach for computing partial autocorrelations and AR parameters. The algorithm proceeds by recursively estimating the partial autocorrelation coefficient by minimizing the sum of the forward and backward prediction error variance. The AR coefficients for the model of order p are computed from the p partial autocorrelation coefficient using the Durbin algorithm. It has been shown that the distribution of the AR parameter estimates is the same as that of the Yule-Walker or conditional maximum likelihood estimates. All the likelihood theory we discussed above applies to these parameter estimates as well. Another compelling feature of the Burg algorithm is that it is fast and guarantees parameter estimates that are stationary. The Burg algorithm is the other algorithm used in Matlab to fit AR models.

Remark 16. The uncertainty in the AR parameter estimates is computed based on large sample method-of-moments and maximum likelihood theory. It is easy to adapt the bootstrap prescription we developed in **Lecture 11** to estimate the uncertainty in the AR parameter estimate.

E. Data Analysis: Application of the Time-Series Analysis Paradigm

Example 3.2 MEG Background Noise Analysis (continued). To illustrate the time-series analysis paradigm, we now analyze the temporal structure in the MEG background noise measurements.

III. Summary

Time-series analysis for continuous-valued data recorded in discrete time uses method-of-moments, maximum likelihood and prediction error analysis methods to conduct model fitting.

All of these approaches are standard parts of statistical packages including Matlab. Spectral analysis is another approach to time-series analysis we investigate in **Lecture 17**.

Acknowledgments

I am grateful to Supratim Saha for performing the data analysis and making the figures and to Julie Scott for technical assistance.

Text References

Brockwell PJ, Davis RA. Introduction to Time-Series and Forecasting, 2nd edition. New York: Springer, 2002.

Box GEP, Jenkins GM, Reinsel GC. Time-Series Analysis: Forecasting and Control, 3rd edition. Englewood Cliffs: Prentice-Hall, 1994.

Literature Reference

Purdon PL, Solo V, Weisskoff RM, Brown EN. Locally regularized spatiotemporal modeling and model comparison for functional MRI. *NeuroImage*, 2001, 14:912-923.