TIME SERIES AND MONTE-CARLO SUMMARY

ANDREW TULLOCH

1. Basic Concepts

Definition (Mean, Covariance function)**.** *Define the mean function* $\mu_X(t) = \mathbb{E}(X_t)$.

Define the covariance function

$$\gamma_X(t,s) = Cov(X_t, X_s) = \mathbb{E}((X_t - \mu_X(t))(X_s - \mu_X(s))). \tag{1.1}$$

Definition (Weak Stationarity)**.** *A time series* X_t *is stationary if*

(i) $\mathbb{E}(|X_t|^2) < \infty$ *for all* $t \in \mathbb{Z}$
(ii) $\mathbb{E}(X_t) = c$ *for all* $t \in \mathbb{Z}$
(iii) $\gamma_X(t,s) = \gamma_X(t+h, s+h))$ *for all* $t, s, h \in \mathbb{Z}$

Definition (Strict Stationarity)**.** *A time series* X_t *is said to be strict stationary if the joint distributions of* $X_{t_1,\dots,X_{t_k}}$ *and* $X_{t_1+h},\dots,X_{t_k+h}$ *are identical for all* k *and for all* $t_1,\dots,t_k, h \in Z$.

Definition (Autocovariance function)**.** *For a stationary time series* X_t*, define the autocovariance function*

$$\gamma_X(t) = Cov(X_{t+h}, X_t). \tag{1.2}$$

and the autocorrelation function

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}. \tag{1.3}$$

Lemma (Properties of the autocovariance function)**.**

$$\gamma(0) \geq 0 \tag{1.4}$$
$$|\gamma(h)| \leq \gamma(0) \tag{1.5}$$
$$\gamma(h) = \gamma(-h) \tag{1.6}$$

for all h.

Note that these all hold for the autocorrelation function ρ*, with the additional condition that* $\rho(0) = 1$.

Theorem. A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and nonnegative definite.

Definition (Sample Autocovariance)**.** *The sample autocovariance function of* $\{x_1,\dots,x_n\}$ *is defined by*

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{j=1}^{n-h}(x_{j+h} - \bar{x})(x_j - \bar{x}), 0 \leq h < n \tag{1.7}$$

and $\hat{\gamma}(h) = \hat{\gamma}(-h)$, $-n < h \leq 0$.

Note that the divisor is n *rather than* $n-h$ *since this ensures that the sample autocovariance matrix*

$$\hat{\Gamma}_n = (\hat{\gamma}(i-j))_{i,j} \tag{1.8}$$

is positive semidefinite.

2. Stationary Processes

2.1. Linear Processes.

Definition (Wold Decomposition)**.** *If* X_t *is a nondeterministic stationary time series, then*

$$X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j} + V_t \tag{2.1}$$

where

(i) $\psi_0 = 1$ *and* $\sum_{j=0}^{\infty}\psi_j^2 < \infty$,
(ii) $Z_t \sim WN(0,\sigma^2)$,
(iii) $Cov(Z_s, V_t) = 0$ *for all* s, t,
(iv) $Z_t = \tilde{P}_t Z_t$ *for all* t,
(v) $V_t = \tilde{P}_s V_t$ *for all* s, t,
(vi) V_t *is deterministic.*

The sequences Z_t, ψ_j, V_t *are unique and can be written explicitly as*

$$Z_t = X_t - \tilde{P}_{t-1}X_t \tag{2.2}$$
$$\psi_j = \frac{\mathbb{E}(X_t Z_{t-j})}{\mathbb{E}(Z_t)^2} \tag{2.3}$$
$$V_t = X_t - \sum_{j=0}^{\infty}\psi_j Z_{t-j}. \tag{2.4}$$

Definition. *A times series* $\{X_t\}$ *is a* ***linear process*** *if it has the representation*

$$X_t = \sum_{j=-\infty}^{\infty}\psi_j Z_{t-j} \tag{2.5}$$

where $Z_t \sim WN(0,\sigma^2)$ *and* $\{\psi_j\}$ *is a sequence of constants with* $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$.

Theorem. The Taylor expansion of $\frac{1}{1-ax}$ is

$$\frac{1}{1-ax} = 1 + ax + a^2x^2 + \dots \tag{2.6}$$

converging for $|ax| \leq 1$

A linear process is called a **moving average** or MA(∞) if $\psi_j = 0$ for all $j < 0$, so

$$X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j}. \tag{2.7}$$

Proposition. *Let* Y_t *be a stationary time series with mean zero and coavariance function* γ_Y*. If* $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$*, then the time series*

$$X_t = \sum_{j=-\infty}^{\infty}\psi_j Y_{t-j} = \psi(B)Y_t \tag{2.8}$$

is stationary with mean zero and autocovariance function

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\psi_j\psi_k\gamma_Y(h+k-j). \tag{2.9}$$

In the special case where X_t *is a linear process,*

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty}\psi_j\psi_{j+h}\sigma^2. \tag{2.10}$$

Theorem. For the case where $X_t = \phi X_{t-1} + Z_t$, we have that if $|\phi| \leq 1$ then $X_t = \sum_{j\geq 0}\phi^j X_{t-j}$, with $\gamma_X(h) = \frac{\sigma^2\gamma^h}{1-\phi^2}$. This is unique in L^2 - take the difference of two solutions, show the tail converges to zero.

If $|\phi| \geq 1$ then the unique solution is $X_t = -\sum_{j\geq 1}\phi^{-j}Z_{t+j}$.

2.2. Forecasting Stationary Time Series.

Our goal is to find the linear combination of $1, X_n, X_{n-1},\dots,X_1$ that forecasts X_{n+h} with minimum mean squared error. The best linear predictor in terms of $1, X_n,\dots,X_1$ will be deonted by P_nX_{n+h} and clearly has the form

$$P_nX_{n+h} = a_0 + a_1X_n + \dots + a_nX_1. \tag{2.11}$$

This follows as

$$U_n = A_nX_n \tag{2.12}$$
$$\hat{X}_n = X_n - U_n \tag{2.13}$$
$$C_n = A_n^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \theta_1 & 1 & \dots & 0 \\ \dots & \dots & & \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \dots & 1 \end{pmatrix} \tag{2.14}$$
$$\hat{X}_n = (C_n - I_nU_n = \Theta_nU_n. \tag{2.15}$$
$$\hat{X}_{n+1} = \begin{cases} 0 & n = 0 \\ \sum_{j=1}^{n}\theta_{n-j}(X_{n+1-j} - \hat{X}_{n+1-j}) \end{cases} \tag{2.16}$$

To find these equations, we solve the convex problem by setting derivatives to zero, and obtain the result given below.

Theorem (Properties of h-step best linear predictor P_nX_{n+h})**.** (i)

$$P_nX_{n+h} = \mu + \sum_{i=1}^{n}a_i(X_{n+1-i} - \mu) \tag{2.17}$$

where $\mathbf{a}_n = (a_1, \dots, a_n)$ satisfies

$$\Gamma_n \mathbf{a}_n = \gamma_n(h) \tag{2.18}$$
$$\Gamma_n = [\gamma(i-j)]_{i,j=1}^n \tag{2.19}$$
$$\gamma_n(h) = (\gamma(h), \gamma(h+1), \dots, \gamma(h+n-1)) \tag{2.20}$$

(ii)
$$\mathbb{E}\big((X_{n+h} - P_n X_{n+h})^2\big) = \gamma(0) - \langle \mathbf{a}_n, \gamma_n(h)\rangle \tag{2.21}$$

(iii)
$$\mathbb{E}(X_{n+h} - P_n X_{n+h}) = 0 \tag{2.22}$$

(iv)
$$\mathbb{E}((X_{n+h} - P_n X_{n+h}) X_j) = 0 \tag{2.23}$$

for $j = 1, \dots, n$.

Definition (Prediction Operator $P(\cdot|\mathbf{W})$). *Suppose that $\mathbb{E}(U^2) < \infty$, $\mathbb{E}(V^2) < \infty$, $\Gamma = Cov(\mathbf{W}, \mathbf{W})$, and $\beta, \alpha_1, \dots, \alpha_n$ are constants.*

(i)
$$P(U|\mathbf{W}) = \mathbb{E}(U) = \mathbf{a}'(\mathbf{W} - \mathbb{E}(\mathbf{W})) \tag{2.24}$$
where $\Gamma \mathbf{a} = Cov(U, \mathbf{W})$.

(ii)
$$\mathbb{E}((U - P(U|\mathbf{W}))\mathbf{W}) = 0 \tag{2.25}$$
and
$$\mathbb{E}(U - P(U|\mathbf{W})) = 0 \tag{2.26}$$

(iii)
$$\mathbb{E}\big((U - P(U|\mathbf{W}))^2\big) = \mathbb{V}(U) - \mathbf{a}' Cov(U, \mathbf{W}) \tag{2.27}$$

(iv)
$$P\alpha_1 + \alpha_2 V + \beta|\mathbf{W} = \alpha_1 P(U|\mathbf{W}) + \alpha_2 P(V|\mathbf{W}) + \beta \tag{2.28}$$

(v)
$$P(\sum_{i=1}^n \alpha_i W_i + \beta|\mathbf{W}) = \sum_{i=1}^n \alpha_i W_i + \beta \tag{2.29}$$

(vi)
$$P(U|\mathbf{W}) = EU \tag{2.30}$$
if $Cov(U, \mathbf{W}) = 0$.

2.3. Innovation Algorithm.

Theorem. Suppose X_t is a zero-mean series with $\mathbb{E}(|X_t|^2) < \infty$ for each t and $\mathbb{E}(X_i X_j) = \kappa(i,j)$. Let $\hat{X}_n = 0$ if $n = 1$, and $P_{n-1} X_n$ if $n = 2, 3, \dots$, and let $v_n = \mathbb{E}\big((X_{n+1} - P_n X_{n+1})^2\big)$.

Define the innovations, or one-step prediction errors, as $U_n = X_n - \hat{X}_n$. Then we can write

$$\hat{X}_{n+1} = \begin{cases} 0 & n = 0 \\ \sum_{j=1}^n \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) \end{cases} \tag{2.31}$$

where the coefficients $\theta_{n1}, \dots, \theta nn$ can be computed recursively from the equations

$$v_0 = \kappa(1,1) \tag{2.32}$$
$$\theta_{n,n-k} = \frac{1}{v_k}(\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j} v_j) \tag{2.33}$$

for $0 \le k < n$, and

$$v_n = \kappa(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j. \tag{2.34}$$

3. ARMA Processes

Definition. *X_t is an ARMA(p,q) process if X_t is stationary and if for every t,*
$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{3.1}$$
where $Z_t \sim WN(0, \sigma^2)$ and the polynomials $(1 - \phi_1 z - \cdots - \phi_p z^p)$ and $(1 + \theta_1 z + \cdots + \theta_q z^q)$ have no common factors.

It can be more convenient to write this in the form
$$\phi(B) X_t = \theta(B) Z_t \tag{3.2}$$
with B the back-shift operator.

ARMA$(0,q)$ is a moving average process of order q (MA(q)). ARMA$(p,0)$ is an autoregressive process of order p (AR(p)).

Theorem. A stationary solution of (3.1) exists (and is the unique stationary solution) if and only if
$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \tag{3.3}$$
for all $|z| = 1$

Definition. *An ARMA(p,q) process X_t is causal (or a causal function of Z_t) if there exists constants ψ_j such that $\sum_{j=0}^\infty |\psi_j| < \infty$ and*
$$X_t = \sum_{j=0}^\infty \psi_j Z_{t-j} \tag{3.4}$$
for all t.

Theorem. An ARMA(p,q) process is causal if and only if
$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \tag{3.5}$$
for all $|z| \le 1$.

Note that the coefficients ψ_j are determined by
$$\psi_j - \sum_{k=1}^p \theta_k \psi_{j-k} = \theta_j \tag{3.6}$$
for $j = 0, 1, \dots$. and $\theta_0 = 1$, $\theta_j = 0$ for $j > q$, and $\psi_j = 0$ for $j < 0$.

Definition. *An ARMA(p,q) is invertible if there exist constants π_j such that $\sum_{j=0}^\infty |\pi_j| < \infty$ and*
$$Z_t = \sum_{j=0}^\infty \pi_j X_{t-j} \tag{3.7}$$
for all t.

The coefficients π_j are determined by the equations
$$\pi_j + \sum_{k=1}^q \theta_k \pi_{j-k} = -\phi_j \tag{3.8}$$
where $\phi_0 = -1$, $\theta_j = 0$ for $j > p$, and $\pi_j = 0$ for $j < 0$.

Theorem. Invertibility is equivalent to the condition
$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0 \tag{3.9}$$
for all $|z| \le 1$.

3.1. ACF and PACF of an ARMA(p,q) Process.

Theorem. For a causal ARMA(p,q) process defined by
$$\phi(B) X_t = \theta(B) Z_t \tag{3.10}$$
we know we can write
$$X_t = \sum_{j=0}^\infty \psi_j Z_{t-j} \tag{3.11}$$
where $\sum_{j=0}^\infty \psi_j z^j = \theta(z)/\phi(z)$ for $|z| \le 1$.

Thus, the ACVF γ is given as
$$\gamma(h) = \mathbb{E}(X_{t+h} X_t) = \sigma^2 \sum_{j=0}^\infty \psi_j \psi_{j+|h|} \tag{3.12}$$

A second approach is to multiple each side by X_{t_k} and take expectations, and obtain a sequence of m homogenous linear difference equations with constant coefficients. These can be solved to obtain the $\gamma(h)$ values.

For example, for $X_t = \frac{1}{3} X_{t-1} + \frac{2}{9} X_{t-2} + Z_t$, we obtain $\rho_k - \frac{1}{3}\rho_{k-1} - \frac{2}{9}\rho_{k-2}$, $k \ge 2$. We attempt a solution of the form $p_k = A\lambda^k$, and thus must have $\lambda^2 - \frac{1}{3}\lambda - \frac{2}{9} = 0$, so $\lambda = \frac{2}{3}, \frac{-1}{3}$. Thus, must have
$$p_k = A(\frac{2}{3})^k + B(\frac{-1}{3})^k \tag{3.13}$$
with $p_0 = A + B = 1$, $p_1 = \frac{1}{3} + \frac{2}{9} p_1$, so $p_1 = \frac{3}{7}$.

Definition (PACF). *The partial autocorrelation function (PACF) of an AMRA process X is the function $\alpha(\cdot)$ defined by*
$$\alpha(0) = 1 \tag{3.14}$$
$$\alpha(h) = \phi_{hh}, h \ge 1 \tag{3.15}$$
where ϕ_{hh} is the last component of $\phi_h = \Gamma_h^{-1}\gamma_h$, where $\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$, and $\gamma_h = [\gamma(1), \gamma(2), \dots, \gamma(h)]$.

Model	ACF	PACF
AR(p)	decaying	0 for $h > p$
MA(q)	0 for $h > q$	decaying
ARMA(p, q)	decaying	decaying

Theorem. For an AR(p) process, the sample PACF values at lags greater than p are approximately independent $N(0, \frac{1}{n})$ random variables. Thus, if we have a sample PACF satisfying

$$|\hat{\alpha}(h)| > \frac{1.96}{\sqrt{n}} \tag{3.16}$$

for $0 \leq h \leq p$ and

$$|\hat{\alpha}(h)| < \frac{1.96}{\sqrt{n}} \tag{3.17}$$

for $h > p$, this suggests an AR(p) model for the data.

Theorem (PACF summary)**.** For an AR(p) process X_t, the PACF $\alpha(\cdot)$ has the properties that $\alpha(p) = \phi_p$, and $\alpha(h) = 0$ for $h > p$. For $h < p$ we can compute numerically from the expression that $\phi_h = \Gamma_h^{-1}\gamma_h$.

3.2. **Forecasting ARMA Processes.** For the causal ARMA(p, q) process

$$\phi(B)X_t = \theta(B)Z_t, Z_t \sim WN(0, \sigma^2) \tag{3.18}$$

we can avoid using the full innovations algorithm.

If we apply the algorithm to the transformed process W_t given by

$$W_t = \begin{cases} \frac{1}{\sigma}X_t & t = 1, \ldots, m \\ \frac{1}{\sigma}\phi(B)X_t & t > m \end{cases} \tag{3.19}$$

where $m = \max(p, q)$.

For notational convenience, take $\theta_0 = 1$, $\theta_j = 0$ for $j > q$.

Lemma. *The autocovariances $\kappa(i, j) = \mathbb{E}(W_i W_j)$ are found from*

$$\kappa(i,j) = \begin{cases} \sigma^2\gamma_X(i-j) & 1 \leq i, j \leq m \\ \sigma^2(\gamma_X(i-j) - \sum_{r=1}^{p}\phi_r\gamma_X(r - |i-j|)) & \min(i,j) \leq m < \max(i,j) \\ \sum_{r=0}^{q}\theta_r\theta_{r+|i-j|} & \min(i,j) > m \\ 0 & \textit{otherwise} \end{cases} \tag{3.20}$$

Applying the innovations algorithm to the process W_t, we obtain

$$\hat{W}_{n+1} = \begin{cases} \sum_{j=1}^{n}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}) & 1 \leq n < m \\ \sum_{j=1}^{q}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}) & n \geq m \end{cases} \tag{3.21}$$

where the coefficients θ_{nj} and MSE $r_n = \mathbb{E}\Big((W_{n+1} - \hat{W}_{n+1})^2\Big)$ are found recursively using the innovations algorithm.

Since the equations (3.19) allow us to write X_n as a linear combination of $W_j, 1 \leq j \leq n$, and conversely, each $W_n, n \geq 1$ to be written as a linear combination of $X_j, 1 \leq j \leq n$. Thus the best linear predictor of the random variable Y in terms of $\{1, X_1, \ldots, X_n\}$ is the same as the best linear predictor of Y in terms of $\{1, W_1, \ldots, W_n$. Thus, by linearity of $\hat{P}_n$, we have

$$\hat{W}_t = \begin{cases} \frac{1}{\sigma}\hat{X}_t & t = 1, \ldots, m \\ \frac{1}{\sigma}(\hat{X}_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}) & t > m \end{cases} \tag{3.22}$$

which shows that

$$X_t - \hat{X}_t = \sigma(W_t - \hat{W}_t) \tag{3.23}$$

Substituting into (3.20) and (3.21), we obtain

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & 1 \leq n < m \\ \phi_1 X_n + \cdots + \phi_p X_{n+1-p} + & \\ \quad \sum_{j=1}^{q}\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & n \geq m \end{cases} \tag{3.24}$$

and

$$\mathbb{E}\Big((X_{n+1} - \hat{X}_{n+1})^2\Big) = \sigma^2\mathbb{E}\Big((W_{n+1} - \hat{W}_{n+1})^2\Big) = \sigma^2 r_n \tag{3.25}$$

where θ_{nj} and r_n are found using the innovation algorithm.

4. Estimation of ARMA Processes

4.1. **Yule-Walker Equations.** Consider estimating a causal AR(p) process. We can write

$$X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j} \tag{4.1}$$

where $\sum_{j=0}^{\infty}\psi_j z^j = \frac{1}{\phi(z)}$ for $z \leq 1$.

Multiplying each side by Z_{t-j}, and taking expectations, we obtain the Yule-Walker equations

$$\Gamma_p\phi = \gamma_p \tag{4.2}$$

and $\sigma^2 = \gamma(0) - \langle\phi, \gamma_p\rangle$ where $\Gamma_p = [\gamma(i-j)]_{i,j=1}^{p}$ and $\gamma_p = (\gamma(1), \gamma(2), \ldots, \gamma(p))$.

If we replace the covariances by the sample covariances $\hat{\gamma}(j)$, we obtain a set of equations for the so-called Yule-Walker estimators $\hat{\phi}$ and $\hat{\sigma}^2$, given by

$$\hat{\Gamma}_p\hat{\phi} = \hat{\gamma}_p \tag{4.3}$$

and $\hat{\sigma}^2 = \hat{\gamma}(0) - \left\langle\hat{\phi}, \hat{\gamma}_p\right\rangle$

Theorem. If X_t is the causal AR(p) process and $\hat{\phi}$ is the Yule-Walker estimator of ϕ, then

$$n^{\frac{1}{2}}(\hat{\phi} - \phi) \xrightarrow{d} N(0, \sigma^2\Gamma_p^{-1}) \tag{4.4}$$

Moreover, $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

Theorem. If X_t is a causal AR(p) process and $\hat{\phi}_m$ is the Yule-Walker estimate of order $m > p$, then

$$n^{\frac{1}{2}}(\hat{\phi}_m - \phi_m) \xrightarrow{d} N(0, \sigma^2\Gamma_m^{-1}) \tag{4.5}$$

where $\hat{\phi}_m$ is the coefficient vector of the best linear predictor $\langle\phi_m, \mathbf{X}_m\rangle$ of X_{m+1} based on $X_m, \ldots, X_1$. So $\phi_m = R_m^{-1}\rho_m$. In particular, for $m > p$,

$$n^{\frac{1}{2}}\hat{\phi}_{mm} \xrightarrow{d} N(0, 1) \tag{4.6}$$

Theorem (Durbin-Levinson Algorithm for AR models)**.** Consider fitting an AR(m) process

$$X_t - \hat{\theta}_{m1}X_{t-1} - \cdots - \hat{\theta}_{mm}X_{t-m} = Z_t \tag{4.7}$$

with $Z_t \sim WN(0, \hat{v}_m)$.

If $\hat{\gamma}(0) > 0$, then the fitted autoregressive models for $m = 1, 2, \ldots, n-1$ can be determined recursively from the relations

$$\hat{\phi}_{11} = \hat{\rho}(1) \tag{4.8}$$

$$\hat{v}_1 = \hat{\gamma}(0)(1 - \hat{\rho}^2)(1) \tag{4.9}$$

$$\hat{\phi}_{mm} = \frac{\hat{\gamma}(m) - \sum_{j=1}^{m-1}\hat{\phi}_{m-1,j}\hat{\gamma}(m-j)}{\hat{v}_{m-1}} \tag{4.10}$$

$$\begin{Bmatrix} \hat{\phi}_{m1} \\ \vdots \\ \hat{\phi}_{m,m-1} \end{Bmatrix} = \hat{\phi}_{m-1} - \hat{\phi}_{mm}\begin{Bmatrix} \hat{\phi}_{m-1,m-1} \\ \vdots \\ \hat{\phi}_{m-1,1} \end{Bmatrix} \tag{4.11}$$

$$\hat{v}_m = \hat{v}_{m-1}(1 - \hat{\phi}_{mm}^2) \tag{4.12}$$

Theorem (Confidence intervals for AR(p) estimation)**.** Under the assumption that the order p of the fitted model is the correct value, for large sample-size n, the region

$$\{\phi \in \mathbb{R}^p | (\phi - \hat{\phi}_p)'\hat{\Gamma}_p(\phi - \hat{\phi}_p) \leq \frac{1}{n}\hat{v}_p\chi_{1-\alpha}^2(p)\} \tag{4.13}$$

contains ϕ_p with probability close to $1 - \alpha$ where $\chi_{1-\alpha}^2(p)$ is the $(1-\alpha)$ quantile of the chi-squared distribution with p degrees of freedom.

Similarly, if $\Phi_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution and $\hat{v}_{jj}$ is the j-th diagonal element of $\hat{v}_p\hat{\Gamma}_p^{-1}$, then for large n

$$\{\hat{\phi}_{pj} \pm \Phi_{1-\frac{\alpha}{2}}\frac{1}{n^{\frac{1}{2}}}\hat{v}_{jj}^{\frac{1}{2}}\} \tag{4.14}$$

contains ϕ_{pj} with probability close to $(1-\alpha)$.

4.2. **Estimation for Moving Average Processes Using the Innovations Algorithm.** Consider estimating

$$X_t = Z_t + \hat{\theta}_{m1}Z_{t-1} + \cdots + \hat{\theta}_{mm}Z_{t-m} \tag{4.15}$$

with $Z_t \sim WN(0, \hat{v}_m)$.

Theorem. We can apply the innovation estimates by applying the recursive relations

$$\hat{v}_0 = \hat{\gamma}(0) \tag{4.16}$$

$$\hat{\theta}_{m,m-k} = \frac{1}{\hat{v}_k}(\hat{\gamma}(m-k) - \sum_{j=0}^{k-1} \hat{\theta}_{m,m-j}\hat{\theta}_{k,k-j}\hat{v}_j) \tag{4.17}$$

for $k = 0, \dots, m-1$, and

$$\hat{v}_m = \hat{\gamma}(0) - \sum_{j=0}^{m-1} \hat{\theta}^2_{m,m-j}\hat{v}_j. \tag{4.18}$$

Theorem. Let X_t be the causal invertible ARMA process $\phi(B)X_t = \theta(B)Z_t$ with $Z_t \sim WN(0, \sigma^2)$, $\mathbb{E}(Z_t^4) < \infty$, and let $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$ for $|z| \leq 1$, and $\psi_0 = 1$ and $\psi_j = 0$ for $j < 0$.

Then for any sequence of positive integers m_n, such that $m < n$, $m \to \infty$, and $m = o(n^{\frac{1}{3}})$ as $n \to \infty$, we have for each k,

$$n^{\frac{1}{2}}(\hat{\theta}_{m1} - \psi_1, \dots, \hat{\theta}_{mk} - \psi_k) \xrightarrow{d} N(0, A) \tag{4.19}$$

where $A = [a_{ij}]_{i,j=1}^k$ and

$$a_{ij} = \sum_{r=1}^{\min(i,j)} \psi_{i-r}\psi_{j-r} \tag{4.20}$$

and

$$\hat{v}_m \xrightarrow{p} \sigma^2. \tag{4.21}$$

Remark. *Note that for the AR(p) process, the Yule-Walker estimator is a consistent estimator of ϕ_p. However, for an MA(q) process, the estimator $\hat{\theta}_q$ is not consistent for the true parameter vector as $n \to \infty$. For consistency, it is necessary to use the estimators with m satisfying the conditions given in Theorem .*

Theorem (Asymptotic confidence regions for the θ_q)**.**

$$\{\theta \in R || \theta - \hat{\theta}_{mj}| \leq \Phi_{1-\frac{\alpha}{2}} \frac{1}{n^{\frac{1}{2}}} (\sum_{k=0}^{j-1} \hat{\theta}^2_{mk})^{\frac{1}{2}}\} \tag{4.22}$$

is an $(1-\alpha)$ confidence interval for θ_{mj}.

4.3. **Maximum Likelihood Estimation.** Consider X_t a gaussian time series with zero mean and autocovariance function $\kappa(i,j) = \mathbb{E}(X_i X_j)$. Let $\hat{X}_j = P_{j-1}X_j$. Let Γ_n be the covariance matrix and assume Γ_n is nonsingular. The likelihood of X_n is

$$L(\Gamma_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{(\det \Gamma_n)^{\frac{1}{2}}} \exp(-\frac{1}{2}\mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n) \tag{4.23}$$

Theorem. The likelihood of the vector $\mathbf{X}_n$ reduces to

$$L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n \prod_{i=0}^{n-1} r_i}} \exp(-\frac{1}{2}\sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}}) \tag{4.24}$$

Remark. *Even if X_t is not Gaussian, the large sample estimates are the same for $Z_t \sim IID(0, \sigma^2)$, regardless of whether or not Z_t is Gaussian.*

Theorem (Maximum Likelihood Estimators for ARMA processes)**.**

$$\hat{\sigma}^2 = \frac{1}{n} S(\hat{\phi}, \hat{\theta}) \tag{4.25}$$

where $\hat{\phi}, \hat{\theta}$ are the values of ϕ, θ that minimize

$$\ell(\phi, \theta) = \ln(\frac{1}{n} S(\theta, \theta)) + \frac{1}{n}\sum_{j=0}^{n-1} \ln r_j \tag{4.26}$$

and

$$S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}} \tag{4.27}$$

Theorem (Asyptotic Distribution of Maximum Likelihood Estimators)**.** For a large sample from an ARMA(p,q) process,

$$\hat{\beta} = N(\beta, \frac{1}{n} V\beta) \tag{4.28}$$

where

$$V(\beta) = \sigma^2 \begin{bmatrix} \mathbb{E}(U_t U_t') & \mathbb{E}(U_t V_t') \\ \mathbb{E}(V_t U_t') & \mathbb{E}(V_t V_t') \end{bmatrix}^{-1} \tag{4.29}$$

and U_t are the autoregressive process $\phi(B)U_t = Z_t$ and $\theta(B)V_t = Z_t$.

Note that for $p = 0$, $V(\beta) = \sigma^2[\mathbb{E}(V_t V_t')]^{-1}$, and for $q = 0$, $V(\beta) = \sigma^2[\mathbb{E}(U_t U_t')]^{-1}$.

4.4. **Order Selection.**

Definition (Kullback-Leibler divergence)**.** *The Kullback-Leibler (KL) divergence between $f(\cdot;\psi)$ and $f(\cdot;\theta)$ is defined as*

$$d(\psi|\theta) = \Delta(\psi|\theta) - \Delta(\theta|\theta) \tag{4.30}$$

where

$$\Delta(\psi|\theta) = \mathbb{E}_\theta(-2\ln f(X;\psi)) \tag{4.31}$$

is the Kullback-Leibler index of $f(\cdot;\psi)$ relative to $f(\cdot;\theta)$.

Theorem (AICC of ARMA(p,q) process)**.**

$$AICC(\beta) = -2\ln L_X(\beta, \frac{S_X(\beta)}{n}) + \frac{2(p+q+1)n}{n-p-q-2} \tag{4.32}$$

Theorem (AIC of ARMA(p,q) process)**.**

$$AIC(\beta) = -2\ln L_X(\beta, \frac{S_X(\beta)}{n}) + 2(p+q+1) \tag{4.33}$$

Theorem (BIC of ARMA(p,q) process)**.**

$$BIC(\beta) = (n-p-q)\ln\frac{n\hat{\sigma}^2}{n-p-q} + n(1+\ln\sqrt{2\pi}) + \tag{4.34}$$

$$(p+q)\ln\frac{\sum_{t=1}^n X_t^2 - n\hat{\sigma}^2}{p+q} \tag{4.35}$$

where $\hat{\sigma}^2$ is the MLE estimate of the white noise variance.

5. Spectral Analysis

Let X_t be a zero-mean stationary time series with autocovariance function $\gamma(\cdot)$ satisfying $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$.

Definition. *The spectral density of X_t is the function $f(\cdot)$ defined by*

$$f(\lambda) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-ih\lambda} y(h) \tag{5.1}$$

The summability implies that the series converges absolutely.

Theorem. (i) f is even
(ii) $f(\lambda) \geq 0$ for all $\lambda \in (-\pi, \pi]$.
(iii) $\gamma(k) = \int_{-\pi}^{\pi} e^{-k\lambda} f(\lambda) d\lambda = \int_{-\pi}^{\pi} \cos(k\lambda) f(\lambda) d\lambda$.

Definition. *A function f is the **spectral density** of a stationary time series X_t with autocovariance function $\gamma(\cdot)$ if*
(i) $f(\lambda) \geq 0$ for all $\lambda \in (0, \pi]$,
(ii) $\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda$ for all integers h.

Lemma. *If f and g are two spectral density corresponding to the autocovariance function γ, then f and g have the same Fourier coefficients and hence are equal.*

Theorem. A real-valued function f defined on $(-\pi, \pi]$ is the spectral density of a stationary process if and only if
(i) $f(\lambda) = f(-\lambda)$,
(ii) $f(\lambda) \geq 0$
(iii) $\int_{-\pi}^{\pi} f(\lambda) d\lambda < \infty$.

Theorem. An absolutely summable function $\gamma(\cdot)$ is the autocovariance function of a stationary time series if and only if it is even and

$$f(\lambda) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h) \geq 0 \tag{5.2}$$

for all $\lambda \in (-\pi, \pi]$, in which case $f(\cdot)$ is the spectral density of $\gamma(\cdot)$.

Theorem (Spectral Representation of the ACVF)**.** A function $\gamma(\cdot)$ defined on the integers is the ACVF of a stationary time series if and only if there exists a right-continuous, nondecreasing, bounded function F on $[-\pi, \pi]$ with $F(-\pi) = 0$ such that

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} dF(\lambda) \tag{5.3}$$

for all integers h.

Remark. *The function F is a **generalized distribution function** on $[-\pi, \pi]$ in the sense that $G(\lambda) = \frac{F(\lambda)}{F(\pi)}$ is a probability distribution function on $[-\pi, \pi]$. Note that since $F(\pi) = \gamma(0) = \mathbb{V}(X_1)$, the ACF of X_t has the spectral representation function*

$$\rho(h) = \int_{-\pi}^{\pi} e^{ih\lambda} dG(\lambda) \tag{5.4}$$

The function F is called the spectral distribution function of $\gamma(\cdot)$. If $F(\lambda)$ can be expressed as $F(\lambda) = \int_{-\pi}^{\lambda} f(y)dy$ for all $\lambda \in [-\pi, \pi]$, then f is the spectral density function and the time series is said to have a continuous spectrum. If F is a discrete function, then the time series is said to have a discrete spectrum.

Theorem. A complex valued function $\gamma(\cdot)$ is the autocovariance function of a stationary process X_t if and only if either

(i) $\gamma(h) = \int_{-\pi}^{\pi} e^{-ihv} dF(v)$ for all $h = 0, \pm 1, \dots$ where F is a right-continuous, non-decreasing, bounded function on $[-\pi, \pi]$ with $F(-\pi) = 0$, or

(ii) $\sum_{i,j=1}^{n} a_i \gamma(i-j)\overline{a}_j \geq 0$ for all positive integers n and all $a = (a_1, \dots, a_n \in \mathbb{C}^n)$.

5.1. The Spectral Density of an ARMA Process.

Theorem. If Y_t is any zero-mean, possibly complex-valued stationary process with spectral distribution function $F_Y(\cdot)$ and X_t is the process $X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j}$ where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then X_t is stationary with spectral distribution function $F_X(\lambda) = \int_{-\pi,\lambda} |\sum_{j=-\infty}^{\infty} \psi_j e^{-ijv}|^2 dF_Y(v)$ for $-\pi \leq \lambda \leq \pi$.

If Y_t has a spectral density $f_Y(\cdot)$, then X_t has a spectral density $f_X(\cdot)$ given by $f_X(\lambda) = |\Psi(e^{-i\lambda})|^2 f_Y(\lambda)$ where $\Psi(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}$.

Theorem. Let X_t be an ARMA(p, q) process, not necessarily causal or invertible satisfying $\phi(B)X_t = \theta(B)Z_t$, $Z_t \sim WN(0, \sigma^2)$ where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ have no common zeroes and $\phi(z)$ has no zeroes on the unit circle. Then X_t has spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{-i\lambda})|^2}{||\phi(e^{-i\lambda})|^2} \tag{5.5}$$

for $-\pi \leq \lambda \leq \pi$.

Theorem. The spectral density of the white noise process is constant, $f(\lambda) = \frac{\sigma^2}{2\pi}$.

Theorem. If we have a mean-zero stationary porcess with $\phi(B)X_t = \theta(B)Z_t$,with $\phi(z)\theta(z) \neq 0$ for $|z| = 1$, then factorizingyields

$$\phi(B) = \prod_{j=1}^{p} (1 - a_j^{-1} B) \tag{5.6}$$

$$\theta(B) = \prod_{j=1}^{q} (1 - b_j^{-1} B) \tag{5.7}$$

where $|a_j| > 1, 1 \leq j \leq r$, $|a_j| \leq 1$, $r \leq j \leq p, |b_j| > 1, 1 \leq j \leq s, |b_j| \leq 1, s \leq j \leq q$.

Noting that

$$|1 - \overline{b}_j e^{-i\lambda}| = |1 - b_j e^{-\lambda}| \tag{5.8}$$

$$= |b_j e^{i\lambda}| |1 - b_j^{-1} e^{-i\lambda}| \tag{5.9}$$

$$= |b_j| |1 - b_j^{-1} e^{-i\lambda}| \tag{5.10}$$

5.2. The Periodogram.

Definition. *The periodogram of $(x_1, \dots, x_n)$ is the function*

$$I_n(\lambda) = \frac{1}{n} |\sum_{t=1}^{n} x_t e^{-it\lambda}|^2 \tag{5.11}$$

Theorem. If $x_1, \dots, x_n$ are any real numbers and ω_k is any of the nonzero Fourier Frequencies $\frac{2\pi k}{n}$ in $(-\pi, \pi]$, then $I_n(\omega_k) = \sum_{|h|<n} \hat{\gamma}(h) e^{-ih\omega_k}$ where $\hat{\gamma}(h)$ is the sample ACVF of $x_1, \dots, x_n$.

Theorem. Let X_t be the linear process $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, $Z_t \sim IID(0, \sigma^2)$, with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Let $I_n(\lambda)$ be the periodogram of $X_1, \dots, X_n$, and let $f(\lambda)$ be the spectral density of X_t.

(i) If $f(\lambda) > 0$for all $\lambda \in [-\pi, \pi]$ and if $0 < \lambda_1 < \cdots < \lambda_m < \pi$, then the random vector $(I_n(\lambda_1), \dots, I_n(\lambda_m))$ converges in distribution to a vector of independent and exponentially distributed random variables, the i-th component which has mean $2\pi f(\lambda_i)$, $i = 1 \dots, m$.

(ii) If $\sum_{j=-\infty}^{\infty} |\psi_j||j|^{\frac{1}{2}} < \infty$, $\mathbb{E}(Z_1^4) = \nu\sigma^4 < \infty$, $\omega_j = \frac{2\pi j}{n} \geq 0$, and $\omega_k = \frac{2\pi k}{n} \geq 0$, then

$$\mathrm{Cov}(I_n(\omega_j), I_n(\omega_k)) = \begin{cases} 2(2\pi)^2 f^2(\omega_j) + O(n^{-\frac{1}{2}}) & \omega_j = \omega_k = \{0, \pi\} \\ (2\pi)^2 f^2(\omega_j) + O(n^{-\frac{1}{2}}) & 0 < \omega_j = \omega_k < \pi \\ O(n^{-1}) & \omega_j \neq \omega_k \end{cases} \tag{5.12}$$

Definition. *The estimator $\hat{f}(\omega) = \hat{f}(g(n, \omega))$ with $\hat{f}(\omega_j)$ defined by*

$$\frac{1}{2\pi} \sum_{|k| \leq m_n} W_n(k) I_n(w_{j+k}) \tag{5.13}$$

*with $m \to \infty$, $\frac{m}{n} \to 0$, $W_n(k) = W_n(-k)$, $W_n(k) \geq 0$ for all k, and $\sum_{|k| \leq m} W_n(k) = 1$, and $\sum_{|k|} W_n^2(k) \to 0$ as $n \to \infty$ is called a **discrete spectral average estimator** of $f(w)$.*

Theorem. Let X_t be the linear process $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, $Z_t \sim IID(0, \sigma^2)$, with $\sum_{j=-\infty}^{\infty} |\psi_j||j|^{\frac{1}{2}} < \infty$ and $\mathbb{E}(Z_1^4) < \infty$. If $\hat{f}$is a discrete spectral average estimator of the spectral density f, then for $\lambda, \omega \in [0, \pi]$,

(i) $\lim_{n\to\infty} \mathbb{E}\Big(\hat{f}(\omega)\Big) = f(\omega)$

(ii)

$$\lim_{n\to\infty} \frac{1}{\sum_{|j| \leq m} W_n^2(j)} \mathrm{Cov}\Big(\hat{f}(\omega), \hat{f}(\lambda)\Big) = \begin{cases} 2f^2(\omega) & w = \lambda = \{0, \pi\} \\ f^2(\omega) & 0 < \omega = \lambda < \pi \\ \\ 0 & \omega \neq \lambda. \end{cases} \tag{5.14}$$

6. ARIMA Processes

Definition. *If d is a non-negative integer, X_t is said to be an ARIMA(p, d, q) process if $Y_t = (1 - B)^d X_t$ is a causal* ARMA(p, q) *process*

Theorem. ARIMA models should be used when there is a slowly decaying positive sample autocorrelation function.

If there is a slowly decaying oscillatory sample ACF, applying the operator $(1 - B + B^2)$ can applied to produce a series with a more rapidly autocorrelation function.

6.1. The Box-Cox Transformation.

Definition. *The Box-Cox transformation is a**variance-stabilizing transformation** - and should be used whenever the standard deviation increases **linearly** with the mean. The equation is*

$$f_\lambda(U_t) = \begin{cases} \frac{(U_t - 1)}{\lambda} & U_t \geq 0, \lambda > 0 \\ \ln U_t & U_t > 0, \lambda = 0. \end{cases} \tag{6.1}$$

6.2. Unit Roots Test.

Theorem. Let $X_1, \dots, X_n$ be observations from the AR(1) model $X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t$ for $Z_t \sim WN(0, \sigma^2)$ where $|\phi_1 < 1$ and $\mathbb{E}(X_t) = \mu$. For large n, the maximum likelihood estimator of $\hat{\phi}_1$ is approximately $N(\phi_1, \frac{1-\phi_1^2}{n})$. In the unit root case, the normal approximation is no longer applicable, which precludes its use for testing the unit root hypothesis $H_0 : \phi_1 = 1$ vs $H_1 : \phi_1 < 1$.

To construct a test of H_0, write the model as $\Delta X_t = X_t - X_{t-1} = \phi_0^\star + \phi_1^\star X_{t-1} + Z_t$, $Z_t \sim WN(0, \sigma^2)$ where $\phi_0^\star = \mu(1 - \phi_1)$ and $\phi_1^\star = \phi_1 - 1$. Now, let $\hat{\phi}_1^\star$ be the OLS estimator of $\phi_1^\star$ found by regression δX_t on 1 and X_{t-1}. The estimated standard error of $\phi_1^\star$ is

$$\hat{SE}(\hat{\phi}_1^\star) = S(\sum_{t=1}^{n} (X_{t-1} - \overline{X})^2)^{-\frac{1}{2}} \tag{6.2}$$

where $S^2 = \sum_{t=2}^{n} (\delta X_t - \hat{\phi}_0^\star - \hat{\phi}_1^\star X_{t-1})^2/(n-3)$ and $\overline{X}$ is the sample mean of $X_1, \dots, X_{n-1}$.

Then the limit distribution of the t-ration $\hat{\tau}_\mu = \frac{\hat{\phi}_1^\star}{\hat{SE}(\hat{\phi}_1^\star)}$ under the unit root assumption can be derived.

Theorem. Let X_t be a causal and invertible ARMA(p, q) process satisfying $\phi(B)X_t = \theta(B)Z_t$, $Z_t \sim WN(0, \sigma^2)$. Then the differenced series $Y_t = \Delta X_t$ is a non-invertible ARMA$(p, q+1)$ process with moving average polynomial $\theta(z)(1 - z)$. Thus, testing for a unit root in a MA polynomial is equivalent to testing that the time series has not been over-differenced.

Let $X_1, \dots, X_n$ be observations from the MA(1) model $X_t = Z_t + \theta Z_{t-1}$, $Z_t \text{IID}(0, \sigma^2)$. Then under the assumption $\theta = -1$, $n(\hat{\theta} + 1)$ where $\hat{\theta}$ is the MLE converges in distribution. A test of $H_0 : \theta = -1$ vs $H_1 : \theta > -1$ can be fashioned on this limiting result by rejecting H_0 when $\hat{\theta} > -1 + \frac{c_\alpha}{n}$ where c_α is the $(1 - \alpha)$ quantile of the limit distribution of $n(\hat{\theta} + 1)$.

7. Random Variate Generation

Theorem. Consider a discrete random variable X with M mass points $\{m_1, \dots, m_M\}$ with probability $p_1, \dots, p_M$, such that $p \geq 0$, $\sum p_i = 1$. Then the CDF of such a random variable is $F(x) = \mathbb{P}(X \leq x) = \sum_{i=1}^{\lfloor x \rfloor_x} p_i$. Write $F(m_i) = F_i$, and note that $p_i = F_i - F_{i-1}$. Then we have that to draw from F, we simulate $U \sim \mathcal{U}_{[0,1]}$, and set $X = m_i$ if $F_{i-1} \leq U \leq F_i$.

Theorem. Let $X \sim F$ be a scalar random variable on $X \subset \mathbb{R}$. Assume F is strictly increasing and continuous. Then $U = F(X) \sim \mathcal{U}_{[0,1]}$.

Proof. $\mathbb{P}(U \leq u) = \mathbb{P}(F(x) \leq u) = \mathbb{P}\big(X \leq F^{-1}(u)\big) = F(F^{-1}(u)) = u.$

Thus $X = F^{-1}(U) \sim F$, as $\mathbb{P}(X \leq x) = \mathbb{P}\big(F^{-1}(U) \leq x\big) = \mathbb{P}(U \leq F(x)) = F(x)$. □

Definition. *Assume that we can generate a uniform random variable U on some domain $\mathcal{U}$. Then let μ the density of this uniform random variable. Assume we now wish to simulate a random variable X uniformly on some domain $\mathcal{X} \subset \mathcal{U}$ ($\mu(\cdot|\mathcal{X})$). Then do to so, simply simulate U uniformly on $\mathcal{U}$, and if $U \in \mathcal{X}$, set $X = U$, otherwise reject U and repeat. Then if $\mu(X) = p > 0$, X is distributed uniformly on $\mathcal{X}$.*

Theorem. Suppose we have two density's f, g and g defined on $X \subset \mathbb{R}$. Suppose that we can samply according to g and there exists $M \in [1, \infty)$ such that $f(x) \leq Mg(x)$ for all $x \in \mathcal{X}$. Then g is an envelope density that dominates f. If we then simulate $Y \sim g$, and some $U \sim \mathcal{U}_{[0,1]}$, and if $U \leq \frac{f(Y)}{Mg(Y)}$, set $X = Y$, otherwise re-sample.

Then the output X has density f.

Theorem. To generate a Gaussian random variable, we can use the **Box-Muller** method. This proceeds by generating $U \sim \mathcal{U}_{[0,1]}$, and setting $\theta = 2\pi U$, and then simulating $V \sim \mathcal{U}_{[0,1]}$, and setting $R^2 - 2\log V$. We can then set $X_1, X_2 = R\sin\theta, R\cos\theta$.

Alternatively, we can generate $U, V \in \mathcal{U}_{[0,1]}$ uniformly on the unit disk, and set $W = U^2 + V^2$ (by rejecting sampling), and set $X_1 = U\sqrt{-2\frac{\log W}{W}}$, $X_2 = V\sqrt{-2\frac{\log W}{W}}$.

8. Monte-Carlo Method and Non-Parametric Inference

Definition. *The plug-in estimator is defined as*

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \tag{8.1}$$

Theorem. Consider estimating $\theta(F) = \mathbb{E}_F \phi(X) = \int_{\mathcal{X}} \phi(x) f(x) dx$, with plug-in estimator $\theta(\hat{F}_n) = \mathbb{E}_{\hat{F}_n} \phi(X) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

Assume that $\int_{\mathcal{X}} \phi(x)^2 f(x) dx < \infty$. Then $\theta(\hat{F}_n)$ is such that

$$\mathbb{E}_{F^n}(\theta(\hat{F}_n)) = \theta(F), \tag{8.2}$$

$$\mathop{\mathrm{V}}_{F^n}(\theta(\hat{F}_n)) = \frac{1}{n}\Big(\int_{\mathcal{X}} (\phi(x) - \theta(F))^2 f(x) dx\Big) \tag{8.3}$$

Theorem. Assume that $\int_{\mathcal{X}} \phi(x)^2 f(x) dx < \infty$. Then the following statements hold:

(i) $\theta(\hat{F}_n)$ converges almost surely to $\theta(F)$

(ii) $\sqrt{n}(\theta(\hat{F}_n) - \theta(F))$ converges in distribution to $Normal(0, \int_{\mathcal{X}} (\phi(x) - \theta(F))^2 f(x) dx)$

Theorem. Let g be a density that is strictly positive whenever $f\phi$ is non-zero. Let $X_1, \dots, X_n$ be IID samples from g, and define the following estimate of $\theta(F)$ as $\hat{\theta}_g = \frac{1}{n}\sum_{i=1}^{n} w_i \phi(X_i)$ where $w_i = \frac{f(X_i)}{g(X_i)}$.

Then $\hat{\theta}_g$ is such that

$$\mathbb{E}_{G^n}(\hat{\theta}_g) = \theta(F), \tag{8.4}$$

$$\mathop{\mathrm{V}}_{G^n}(\hat{\theta}_g) = \frac{1}{n}\Big(\int_{\mathcal{X}} \frac{\phi(x)^2 f(x)^2}{g(x)} dx - \theta(F)^2\Big) \tag{8.5}$$

Theorem. Note that the variance of this estimate depends crucially on the function $\psi(x) = \frac{\phi(x)^2 f(x)^2}{g(x)}$, and the density that minimizes the variance of $\hat{\theta}_g$ is $g^\star = \frac{f|\phi|}{\int_{\mathcal{X}} f|\phi|}$.

Definition. *If we sample randomly on our space $\mathcal{X}$ according to f, we create a source of randomness and thus of error. By using **stratified sampling** we reduce the error by reducing the amount of randomness in the picking of the points.*

(i) Divide the domain into K strata Ω_i that are measurable according-to f and form a partition of the domain, and we know exactly $w_i = \mathbb{P}_f(\Omega_i)$.

(ii) Sample exactly T_i in each stratum Ω_i according to $f|\Omega_i$. The numbers T_i are deterministic with $\sum T_i$. Writing the conditional empirical mean in stratum Ω_i as $\mu_i = \frac{1}{T_i}\sum_{j=1}^{n} X_j \mathbb{I}(X_j \in \Omega_i)$, and return the weighted estimate of the integral $\hat{\mu} = \sum_{i=1}^{K} w_i \hat{\mu}_i$.

Theorem. $\mathbb{E}(\hat{\mu}) = \mu$, and $\mathrm{V}\,\hat{\mu} = \sum_{i=1}^{K} \frac{w_i^2 \sigma_i^2}{T_i}$ where $\mu_i = \frac{1}{w_i}\int_{\Omega_i} \phi(x) f(x) dx$ is the conditional mean, and $\sigma_i^2 = \frac{1}{w_i}\int_{\Omega_i} (\phi(x) - \mu_i)^2 f(x) dx$ is the conditional variance in stratum Ω_i.

Definition. *Uniform stratified sampling takes $T_i^u = w_i n$. This **consolidates** the random sampling while preserving the shape of the density f. With this choice of T_i, we have $\mathbb{E}(\hat{\mu}_u) = \mu$, and $\mathrm{V}\,\hat{\mu}_u = \sum_{i=1}^{K} \frac{w_i \sigma_i^2}{n}$.*

Theorem. If we solve the minimization problem

$$\min_{(T_i)_i} \mathrm{V}\,\hat{\mu} = \sum_{i=1}^{K} \frac{w_i^2 \sigma_i^2}{T_i} \tag{8.6}$$

such that $T_i \geq 0$, $\sum_i T_i = n$ we obtain the unique solution

$$T_i^\star = \frac{w_i \sigma_i}{\sum_j w_j \sigma_j} n, \tag{8.7}$$

known as **oracle stratified sampling**. In this case, $\mathbb{E}(\hat{\mu}^\star) = \mu$, and $\mathrm{V}\,\hat{\mu}_n^\star = \frac{1}{n}(\sum_{i=1}^{K} w_i \sigma_i)^2$.

Theorem. If we have Lipschitz function ϕ that we are integrating uniformly over $[0, 1]$, we have there exists $C > 0$ such that for any u, v, $|\phi(u) - \phi(v)| \leq C|u - v|$. If we uniformly divide $[0, 1]$ into K sections, we have $\Omega_i = [\frac{i}{K}, \frac{i+1}{K}]$, with $w_i = \frac{1}{K}$. Then we have $\mathrm{V}\,\hat{\mu}_u \leq \sum_{i=1}^{n} \frac{1}{n}\int_{\Omega_i} \frac{C^2}{n^2} dx = \frac{C^2}{n^3}$, which is a significant gain over the Monte-Carlo estimate of $\frac{1}{n}$.

Definition. *To estimate the c_α quantile with tail probability α of a distribution F, we can do this with Monte-Carlo by*

(i) Choose $B \in \mathbb{N}$.

(ii) Restrict $\alpha \in k = \{1, \dots, B\}$, with $\alpha = \frac{k}{B+1}$.

(iii) Simulate $T_1, \dots, T_B$ according to F.

(iv) Let $\hat{c}_\alpha = T_{(k)}$ where $T_{(k)}$ is the k-th order statistic of $T_1, \dots, T_B$.

Theorem. If F corresponds to a density f, then $\mathbb{E}(F(\hat{c}_\alpha)) = \alpha$.

Definition. *Suppose we are interested in the distribution $K_n(F)$ of a root or pivot $R_n(X, F)$ where $X = (X_1, \dots, X_n)$ (e.g. the distribution of the statistic $T(X_1, \dots, X_n)$ in hypothesis tests). The boostrapboostrap estimator of $K_n(F)$is $K_n(\hat{F})$.*

Definition. *To approximate the Boostrap estimator by Monte-Carlo, we compute the following:*

(i) Draw B independent boostrap samples $X_b^\star = (X_{b,1}^\star, \dots, X_{b,n}^\star)$ from $\hat{F}^n$.

(ii) Approximate $K_n(\hat{F})$ by the empirical distribution function of $(R_n(X_b^\star, \hat{F})$

9. Bayesian Inference and Associated Methods

Definition. *Our objective is to generate samples according to the posterior $\pi(\theta|\overline{X}) = L(X, \theta)p(\theta)$.*

The solution is the generate a Markov chain $(\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(t)}, \dots)$ such that π is the stationary distribution.

Definition. *The Gibbs sampler proceeds as follows. Let $\overline{\theta} = (\theta_1, \dots, \theta_p)$ be the parameter of interest and $\pi(\theta_i|\overline{\theta}_{(-i)}) = \pi_i(\overline{\theta}_{(-i)})$ be the conditional posterior distributions. The Gibbs sampler works as follows:*

(i) Set the initial vector $\theta^{(0)}$,

(ii) At time $t + 1$:

(i) Set $\theta_1^{(t+1)} \sim \pi_1(\theta_2^{(t)}, \theta_3^{(t)}, \dots) = \pi_1(\overline{\theta}_{(-1)}^{(t)})$

(ii) Set $\theta_2(t+1) = \pi_2(\theta_1^{t+1}, \theta_3^{(t)}, \dots, \theta_p^{(t)}) = \pi_2(\overline{\theta}_{(-2)}^{(t)})$

(iii) ...

(iv) Set $\theta_p^{(t+1)} \sim \pi_p(\theta_{(-p)}^{(t+1)})$

(iii) Collect T samples using this iteration.

(iv) Throw away the first b samples and consider only the last samples.

This method generates a Markov chain whose stationary distribution is the posterior under not-too-strong assumptions.

*Note that the samples $\overline{\theta}^{(t)}$ are **not independent**.*

Proof. The probability of transition from a to b, $K(b,a)$ is given as $K(b,a) = \min\{\frac{\pi(b)}{\pi(a)},1\} + (1 - \sum_{b'=1}^{M} \min\{\frac{\pi(b)}{\pi(a)},1\})\mathbb{I}(a=b)$. Then

$$(K\pi)(b) = \sum_{a=1}^{M} K(b,a)\pi(a) \tag{9.1}$$

$$= \sum_{a=1}^{M} (\min\{\frac{\pi(b)}{\pi(a)},1\} + (1 - \sum_{b'=1}^{M} \min\{\frac{\pi(b')}{\pi(a)},1\})\mathbb{I}(a=b))\pi(a) \tag{9.2}$$

$$= \pi(b) \tag{9.3}$$

so $K\pi = \pi$, and so π is the invariant point of K and thus the stationary measure. □

Definition. *The Metropolis Hastings algorithm is a sequential form of rejection sampling. It is an Markov Chain Monte-Carlo method to construct a Markov chain whose stationary distribution is π.*

Consider a distribution π defined on a domain $\mathcal{X}$. Assume that π is such that for any atom $x \in \mathcal{X}$, $\{x\}$ is measurable according to π. For any $x \in \mathcal{X}$, define a transition measure $\mu(\cdot|x)$ on $\mathcal{X}$. The method proceeds as follows:

(i) Set the initial vector $\theta^{(0)}$.
(ii) At time $t+1$,
  (i) Simulate $X \sim \mu(\cdot|\theta^{(t)})$ and $U \sim \mathcal{U}_{[0,1]}$
  (ii) If $U \leq \frac{\pi(X)\mu(\theta^{(t)}|X)}{\pi(\theta^{(t)})\mu(X|\theta^{(t)})}$, then $\theta^{(t+1)} = X$, otherwise $\theta^{(t+1)} = \theta^{(t)}$.
(iii) Collect T samples like that.
(iv) Throw away the first b samples, consider the last samples (and possibly sub-sample to reduce correlations.)

Note that the initial state is important (particular for unbounded distributions), the transition probability is important (for fast convergence), and the number of samples discarded is problem-dependent.

Definition. *Given MCMC methods produce a correlated chain of samples - so t samples do not provide the sample information as t IID samples. A common notion for measuring this is the **effective sample size**. For any $l \geq 0$, define $\gamma(l)$ as the correlation between two samples of lag l. Defining $\rho(l) = \frac{\gamma(l)}{\gamma(1)}$, the effective sample size $\tilde{T}$of a chain of length t is*

$$\tilde{T} = \frac{t}{1 + 2\sum_{l=1}^{T-1} \rho(l)}. \tag{9.4}$$

References