Time Series: A First Course with Bootstrap Starter

Contents

Lesson 8-1: Introduction to Entropy

- We introduce **entropy** as a measure of randomness and unpredictability.
- Modeling time series involves increasing the entropy, to where we cannot predict anything.

Definition 8.1.8.

- The **entropy** of a discrete random variable X is denoted $H(X)$:

$$H(X) = -\sum_{k} \mathbb{P}[X = k] \log(\mathbb{P}[X = k]).$$

- This is non-negative, with higher values corresponding to greater uncertainty.
- A value of zero corresponds to X being deterministic almost surely.

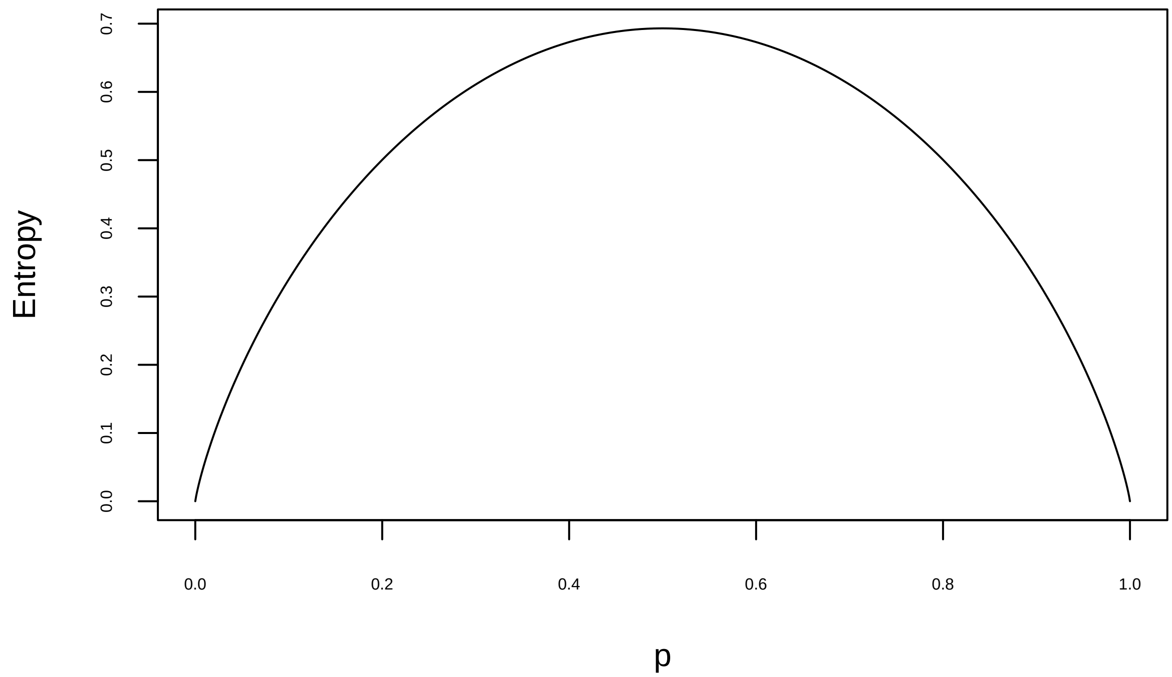
Example 8.1.10. Bernoulli Entropy

- Let X be a Bernoulli random variable with success probability p. The entropy is

$$-(p \log p + (1-p) \log(1-p)).$$

- Entropy is highest for $p = 1/2$, and lowest for $p = 0, 1$.

```
pvals <- seq(0,1000)/1000
ber.ent <- -pvals*log(pvals) - (1-pvals)*log(1-pvals)
ber.ent[1] <- 0
ber.ent[1001] <- 0
plot(ts(ber.ent,start=0,frequency=1000),xlab="p",ylab="Entropy",
    yaxt="n",xaxt="n")
axis(1,cex.axis=.5)
axis(2,cex.axis=.5)
```



Definition 8.1.12.

- The **differential entropy** of a continuous random variable X is

$$H(X) = -\int p(x)\ \log(p(x))\,dx,$$

  where p is the probability density function.
- We integrate over the support (where p is positive).
- Note that $H(X) = -\mathbb{E}[\log p(X)]$.
- Can take negative values, but interpretation is the same as discrete case.
- Concepts extends to random vectors by taking the joint pdf.

Example 8.1.13. Gaussian Entropy

- Suppose $\underline{X}$ is normal with mean zero and n-dimensional covariance matrix Σ, which is assumed to be non-singular. Then

$$\log p(X) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}\underline{X}'\Sigma^{-1}\underline{X}.$$

- The entropy is the expectation of this times -1:

$$H(X) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det\Sigma + \frac{n}{2},$$

  since $\mathbb{E}[\underline{X}'\Sigma^{-1}\underline{X}] = n$ (see Lesson 2-1).

Exercise 8.4. Poisson Entropy Computation

- We compute the entropy of a Poisson random variable of parameter λ. The probability mass function is

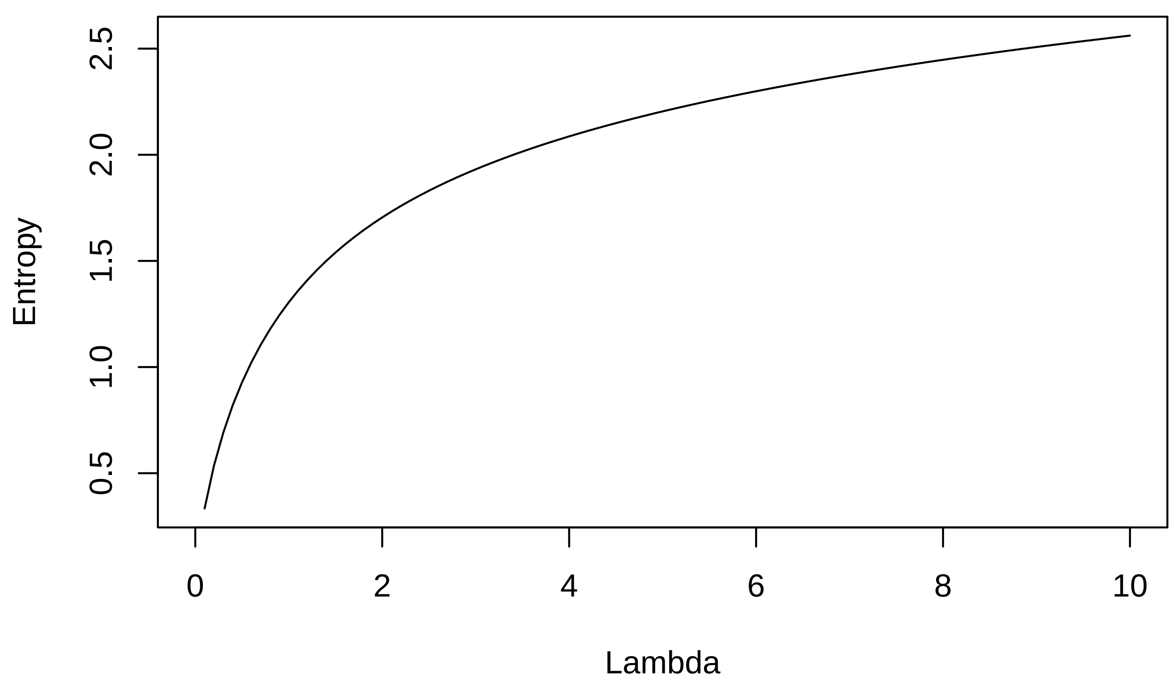
$$\mathbb{P}[X = k] = \lambda^k e^{-\lambda}/k!.$$

- The entropy is computed through the following function. This uses a truncation of the infinite summation.

```
pois.ent <- function(lambda,trunc)
{
    ent <- exp(-lambda)*sum(lambda^(seq(0,trunc))*log(factorial(seq(0,trunc)))/factorial(seq(0,trunc)))
    ent <- ent + lambda*(1 - log(lambda))
    return(ent)
}
```

- We plot the entropy for various λ, truncating at 50.

```
lambda <- seq(0,100)/10
my.ents <- NULL
for(i in 1:length(lambda))
{
    my.ents <- c(my.ents,pois.ent(lambda[i],50))
}
plot(ts(my.ents,start=0,frequency=10),xlab="Lambda",ylab="Entropy")
```



Lesson 8-2: Entropy Mixing

- We can measure how far apart two random variables are through entropy.
- "Mixing" refers to a property of some time series, whereby random variables that are temporally far apart have less dependence.

Fact 8.3.9. Entropy of a Random Sample

- If the components of $\underline{X}$ are independent then entropy is additive: $p(\underline{x}) = \prod_{k=1}^n p_k(x_k)$ implies

$$H(\underline{X}) = -\mathbb{E}[\log \prod_{k=1}^n p_k(X_k)] = -\sum_{k=1}^n \mathbb{E}[\log p_k(X_k)] = \sum_{k=1}^n H(X_k).$$

- So in general, we can measure dependence by comparing $H(\underline{X})$ to $\sum_{k=1}^n H(X_k)$.
- When the X_k are identically distributed, $H(\underline{X}) = nH(X_1)$.

Paradigm 8.2.11. Entropy Mixing

- For any two random variables X and Y, the *entropy mixing coefficient* is

$$\beta(X, Y) = H(X) + H(Y) - H(X, Y).$$

- This is always non-negative.
- If $\{X_t\}$ is strictly stationary, $\beta(X_t, X_{t+h})$ does not depend on t, and we write $\beta_X(h)$.

Example 8.2.12. Entropy Mixing for Gaussian Time Series

- Suppose $\{X_t\}$ is mean zero stationary Gaussian with ACVF $\gamma(h)$. Then

$$\beta_X(h) = 1 + \log(2\pi) + \log\gamma(0) - (1 + \log(2\pi)) - \frac{1}{2}\log\det\Sigma,$$

  using the $n = 1, 2$ cases of Example 8.1.13. Here

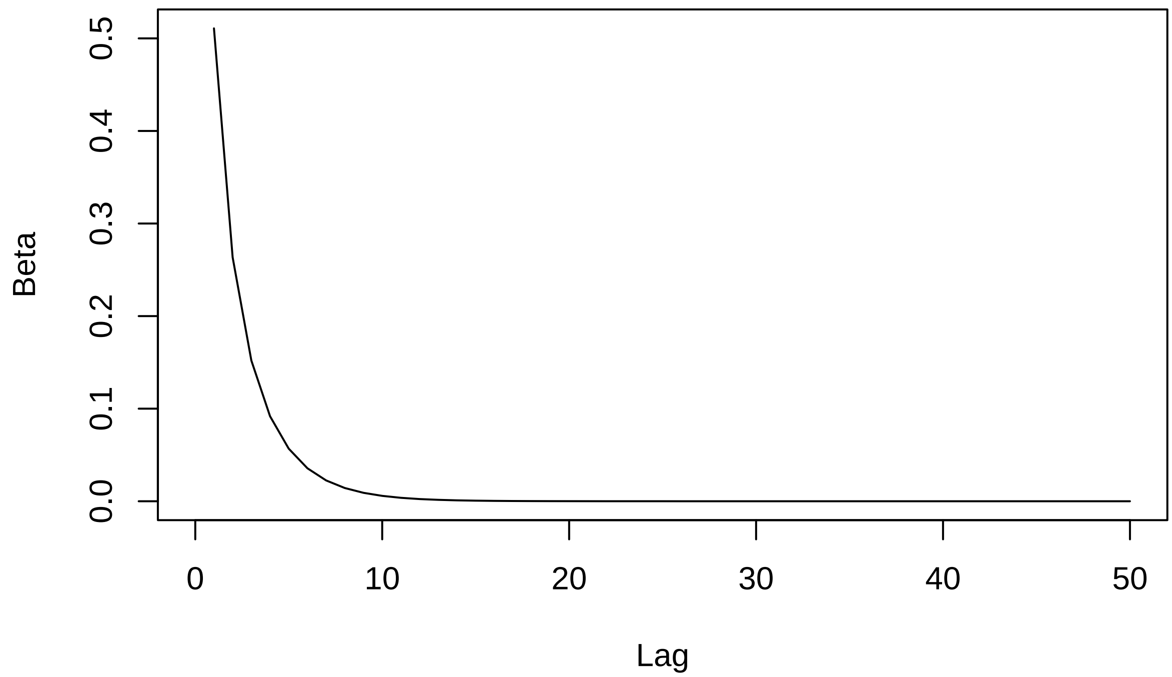
$$\Sigma = \left[\begin{array}{cc} \gamma(0) & \gamma(h) \\ \gamma(h) & \gamma(0) \end{array}\right].$$

- So we get

$$\beta_X(h) = -\frac{1}{2}\log(1 - \rho(h)^2).$$

- These mixing coefficients are non-negative, and zero only if $\rho(h) = 0$ (which corresponds to no dependence at lag h).
- Consider the case of a Gaussian AR(1), with $\phi_1 = .8$. We display the entropy mixing coefficients.

```
phi1 <- .8
lags <- 50
beta <- -.5*log(1 - phi1^(2*seq(0,lags)))
plot(ts(beta,start=0),xlab="Lag",ylab="Beta")
```



Lesson 8-3: Maximum Entropy

- We discuss the maximum entropy principle.

Paradigm 8.3.1. Maximum Entropy Principle

- If the parameters of a distribution are chosen so as to maximize entropy, we guard against a worst-case scenario for the state of nature.
- So we seek to maximize entropy subject to the observed data.

Example 8.3.2. Bernoulli Maximum Entropy

- In Example 8.1.10 with a Bernoulli random variable, if we have no data the maximum entropy principle yields $p = 1/2$.
- If we observed $X = 1$, we would instead say $p = 1$.

Definition 8.3.5.

Given two continuous random variables X and Y with probability density functions p and q respectively, the **relative entropy** of X to Y is

$$H(X;Y) = -\int p(x)\,\log\left(\frac{q(x)}{p(x)}\right)\,dx = -\int p(x)\log q(x)\,dx - H(X).$$

By Jensen's inequality, $H(X;Y) \geq 0$ and equals zero iff X and Y have the same distribution.

Example 8.3.8. Gaussian has Maximum Entropy given its Variance

- Suppose that X is a continuous random variable with pdf p, and has mean zero and variance σ^2.
- Let $Y \sim \mathcal{N}(0, \sigma^2)$, but with pdf q.
- Then $\log q(x) = -.5 \log(2\pi) - .5x^2/\sigma^2$, and

$$-\int p(x) \log q(x)\, dx = .5 \log(2\pi) + .5\sigma^{-2} \int x^2 p(x) dx = .5(1 + \log(2\pi)).$$

- The relative entropy is

$$H(X; Y) = -\int p(x) \log q(x)\, dx - H(X) = .5(1 + \log(2\pi)) - H(X).$$

- Since relative entropy is non-negative, we find that $H(X) \leq .5(1 + \log(2\pi))$. This is a bound on any such X, and the Gaussian attains this upper bound (since $H(Y) = .5(1+\log(2\pi))$). Hence the Gaussian has maximum entropy.

Remark 8.3.10. Redundancy Lowers Entropy

- By Fact 8.3.9, entropy increases linearly in sample size for a random sample.
- When dependence is full, then $H(\underline{X}) = H(X_1)$ instead.
- So redundancy (full dependence) lowers entropy.
- By the maximum entropy principle, serial independence is favored over dependence on a priori grounds.

Definition 8.3.11.

- A transformation Ξ that maps $\underline{X}$ to $\Xi[\underline{X}]$ is *entropy-increasing* if

$$H\left(\Xi[\underline{X}]\right) > H(\underline{X}).$$

- For instance: Ξ decorrelates (reduces dependence), Ξ preserves variance while transforming marginal structure to Gaussian.

Example 8.3.12. Whitening as an Entropy-Increasing Transformation

- Suppose $\underline{X} \sim \mathcal{N}(0, \Sigma)$. We want to decorrelate $\underline{X}$, and see how entropy changes.
- Let D denote the diagonal entries of Σ. Let $LL' = \Sigma$ be the Cholesky decomposition. Then $\underline{Y} = D^{1/2} L^{-1} \underline{X} \sim \mathcal{N}(0, D)$.
- So $\underline{Y} = \Xi[\underline{X}]$ has been decorrelated.
- Comparing entropies:

$$H(\underline{X}) - H(\underline{Y}) = .5 \log \det \Sigma - .5 \log \det D = .5 \log \det \left(D^{-1/2} \Sigma D^{-1/2} \right).$$

- The matrix $R = D^{-1/2} \Sigma D^{-1/2}$ is a correlation matrix, and has determinant between 0 and 1. Hence $H(\underline{X}) \leq H(\underline{Y})$, and inequality is strict unless $\det R = 1$.

Illustration of Example 8.3.12.

- Consider the bivariate normal $\underline{X}$ with mean zero and variance

$$\Sigma = \left[\begin{array}{cc} 2 & 3 \\ 3 & 5 \end{array} \right].$$

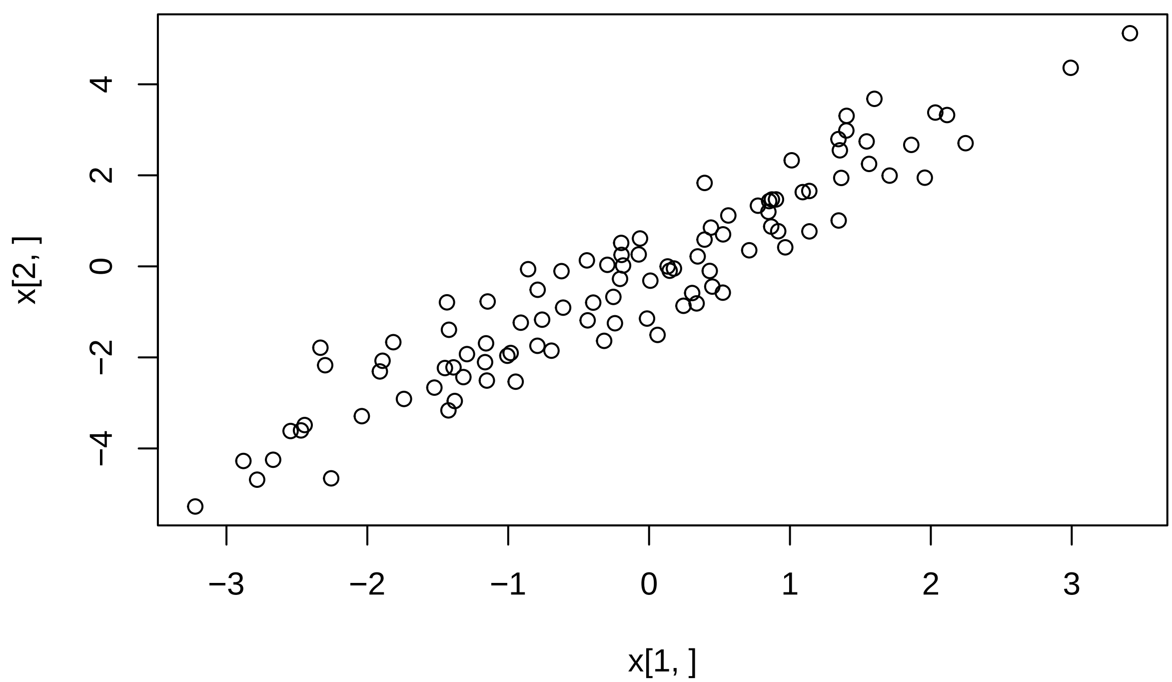
- So $D = \text{diag}[2, 5]$.
- We print the entropies of $\underline{X}$ and $\underline{Y} = D^{1/2} L^{-1} \underline{X}$.

```r
Sigma <- rbind(c(2,3),c(3,5))
D <- diag(Sigma)
L <- t(chol(Sigma))
ents <- c(log(1 + 2*pi ) + .5*log(det(Sigma)), log(1 + 2*pi ) + .5*sum(log(D)))
print(ents)
```

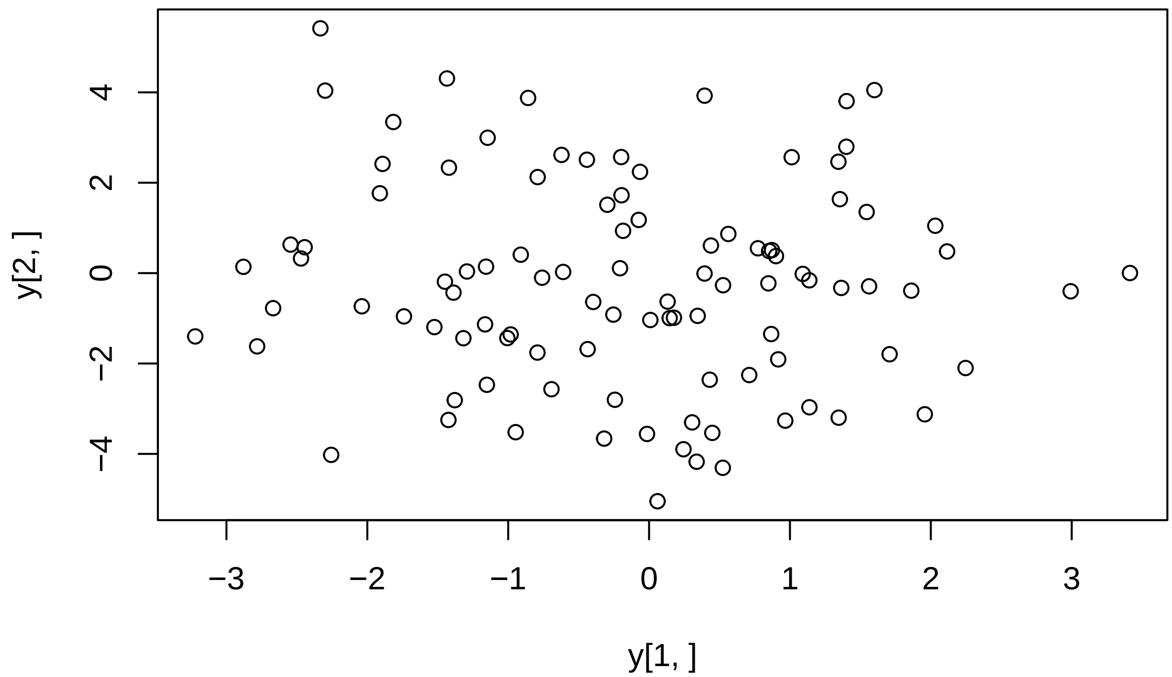
```
## [1] 1.985568 3.136861
```

- We simulate 100 draws of $\underline{X}$, construct the decorrelated $\underline{Y}$, and generate both scatterplots.

```r
x <- L %*% matrix(rnorm(2*100),nrow=2)
y <- diag(sqrt(D)) %*% solve(L) %*% x
plot(x[1,],x[2,])
```




```r
plot(y[1,],y[2,])
```



Lesson 8-4: Time Series Entropy

- We now extend the concept of entropy to time series.
- We also define conditional entropy.

Definition 8.4.1.

The **entropy rate** of a strictly stationary time series $\{X_t\}$ is

$$h_X = \lim_{n\to\infty} n^{-1} H(\underline{X}),$$

where $\underline{X} = [X_1, \ldots, X_n]$.

Example 8.4.2. Gaussian Entropy Rate

- We determine the entropy rate for a stationary Gaussian time series with mean zero and spectral density f.
- Using Theorem 6.4.5,

$$\det \Gamma_n \approx \det \Lambda = \prod_{\ell=[n/2]-n+1}^{[n/2]} f(\lambda_\ell).$$

- Taking logs and using Riemann sums, we obtain

$$h_X = .5 \left(1 + \log(2\pi) + (2\pi)^{-1} \int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda\right).$$

Definition 8.4.5. Conditional Entropy

- The conditional entropy of X given $\underline{Z}$ is

$$H(X|\underline{Z}) = -\mathbb{E}[\log p_{X|\underline{Z}}(X|\underline{Z})].$$

- Conditioning always lowers entropy: $H(X|\underline{Z}) \leq H(X)$.
- More information means that future outcomes are less uncertain.

Proposition 8.4.8.

- The entropy rate of a strictly stationary time series $\{X_t\}$ is

$$h_X = H(X_0|X_{-1}, X_{-2}, \ldots).$$

- This is the entropy of X_0 conditional on its infinite past.

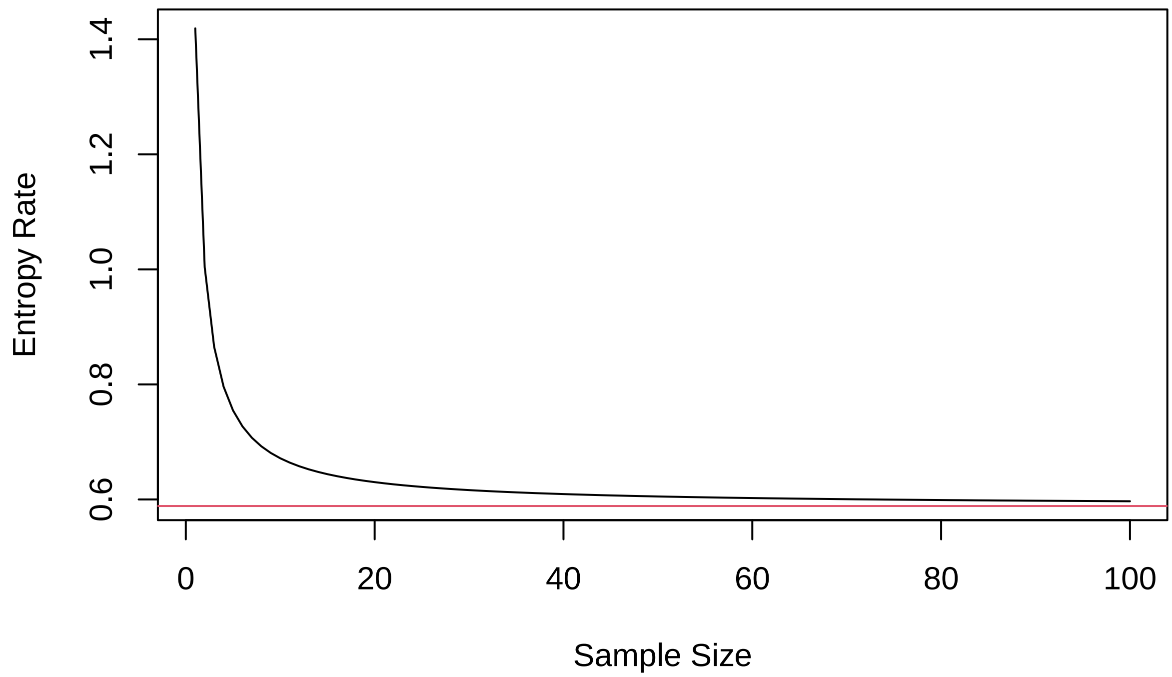
Exercise 8.29. Entropy of a Gaussian AR(1).

- We compute the entropy for a sample of an AR(1), and compare to the entropy rate.

```
gauss.ent.ar1 <- function(phi,sig2,n)
{
    Sigma <- toeplitz(phi^(seq(0,n-1,length=n))*sig2/(1-phi^2))
    ent <- n/2*(1 + log(2*pi)) + log(det(Sigma))/2
    return(ent)
}
```

- We consider an AR(1) with $\phi = .9$ and variance 1. So $\sigma^2 = 1 - \phi^2$, and the entropy rate is $.5(1 + \log(2\pi) + \log \sigma^2)$.
- We plot the entropy divided by n.

```
phi <- .9
sig2 <- 1-phi^2
ent.rate <- .5*(1 + log(2*pi) + log(sig2))
ents <- NULL
for(n in 1:100)
{
    ents <- c(ents,gauss.ent.ar1(phi,sig2,n))
}
plot(ts(ents/seq(1,100)),xlab="Sample Size",ylab="Entropy Rate")
abline(h = ent.rate,col=2)
```



Lesson 8-5: Markov Time Series

- We introduce the class of Markov processes, and show that they have a maximum entropy property.

Definition 8.5.1.

- A process $\{X_t\}$ is **Markov** of order p if

$$p_{X_t|X_{t-1},\ldots,X_{t-m}} = p_{X_t|X_{t-1},\ldots,X_{t-p}}$$

  for any t and $m \geq p$.
- So the conditioning on the past only involve the past p observations.

Example 8.5.2. Causal AR(p)

- Consider a causal AR(1) given by $X_t = \phi X_{t-1} + Z_t$.
- Conditional on $X_{t-1} = x$, we have X_t given by $\phi x + Z_t$, so

$$p_{X_t|X_{t-1}=x}(y) = p_Z(y - \phi x).$$

- Hence, $p_{X_t|X_{t-1},\ldots,X_{t-m}} = p_{X_t|X_{t-1}}$ and the the process is Markov(1).
- Generalizing, a causal AR(p) is Markov(p).
- The converse is true for Gaussian processes: if it is Markov(p), then it is causal AR(p).

Proposition 8.5.5.

- A Gaussian AR(p) process has maximum entropy rate among strictly stationary processes with given $\gamma(0), \ldots, \gamma(p)$.

- Why: maximize the Gaussian entropy rate formula subject to the unknowns $\gamma(k)$ for $k > p$. So differentiate h_X with respect to $\gamma(k)$:

$$\frac{\partial}{\partial \gamma(k)} h_X = (4\pi)^{-1} \int_{-\pi}^{\pi} \frac{\partial}{\partial \gamma(k)} \log f(\lambda)\, d\lambda = (4\pi)^{-1} \int_{-\pi}^{\pi} \frac{2\cos(k\lambda)}{f(\lambda)}\, d\lambda = \xi_k,$$

  since $f(\lambda) = \gamma(0) + 2\sum_{k\geq 1} \gamma(k)\cos(k\lambda)$.
- Setting these derivatives to zero, we see that the inverse autocovariances are zero for $k > p$. Hence the process must be an AR(p).
- So a Markov(p) is maximum entropy among Gaussian processes with given autocovariances up to lag p.

Lesson 8-6: Modeling via Entropy

- We can attempt to model data so as to increase entropy.

Definition 8.6.2.

Given a sample $\underline{X}$ with pdf $p_{\underline{X}}$, a **model** is the composition Π of successive entropy-increasing transformations, such that the **residuals** $\underline{Z} = \Pi[\underline{X}]$ has maximum entropy among the class of transformations.

Example 8.6.7. Log Difference

- The growth rate transformation (log differencing) can sometimes increase entropy for economic time series.
- Consider $X_t = \exp Z_t$ where $\{Z_t\}$ is a random walk. Log differencing is a model Π that maps the process to white noise:

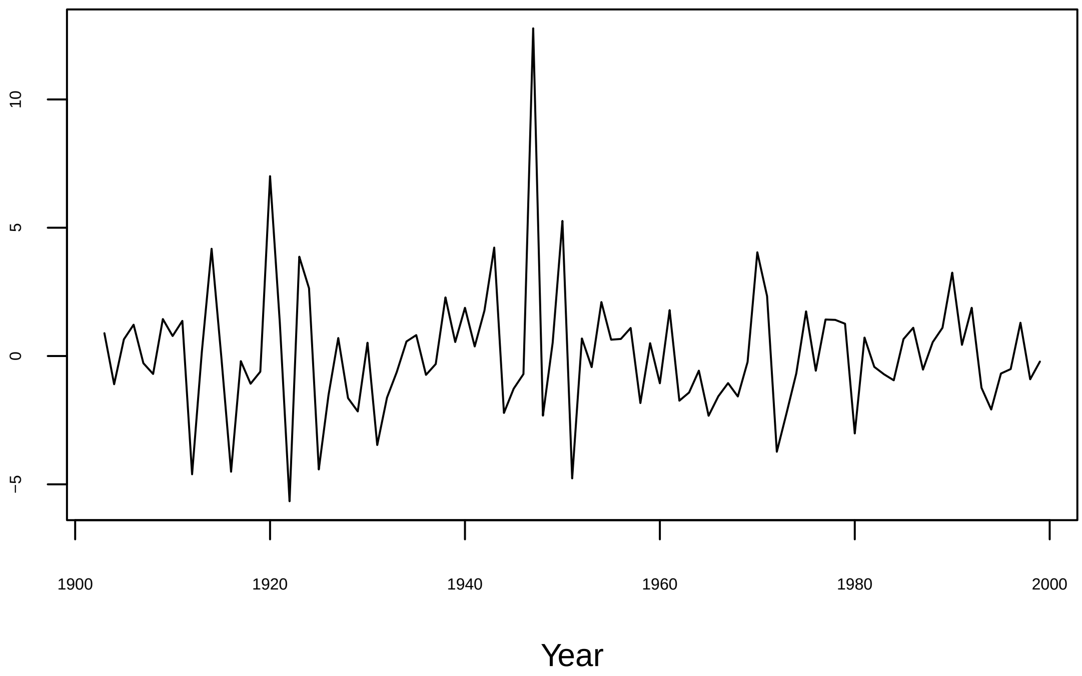
$$\log X_t - \log X_{t-1} = Z_t - Z_{t-1} = \epsilon_t.$$

Example 8.6.8. Entropy-Increasing Transformation for U.S. Population

- The raw data of U.S. population has a lot of structure (low entropy).
- We know that first differences remove much of the trend structure. So $1 - B$ is a model for the data.
- We can also consider the model $(1 - B)^2$.

```r
pop <- read.table("USpop.dat")
pop <- ts(pop, start = 1901)
diff.pop <- diff(pop*10e-6)
diffdiff.pop <- diff(diff(pop*10e-6))
plot(diff.pop,xlab="Year",ylab="",yaxt="n",xaxt="n")
axis(1,cex.axis=.5)
axis(2,cex.axis=.5)
```

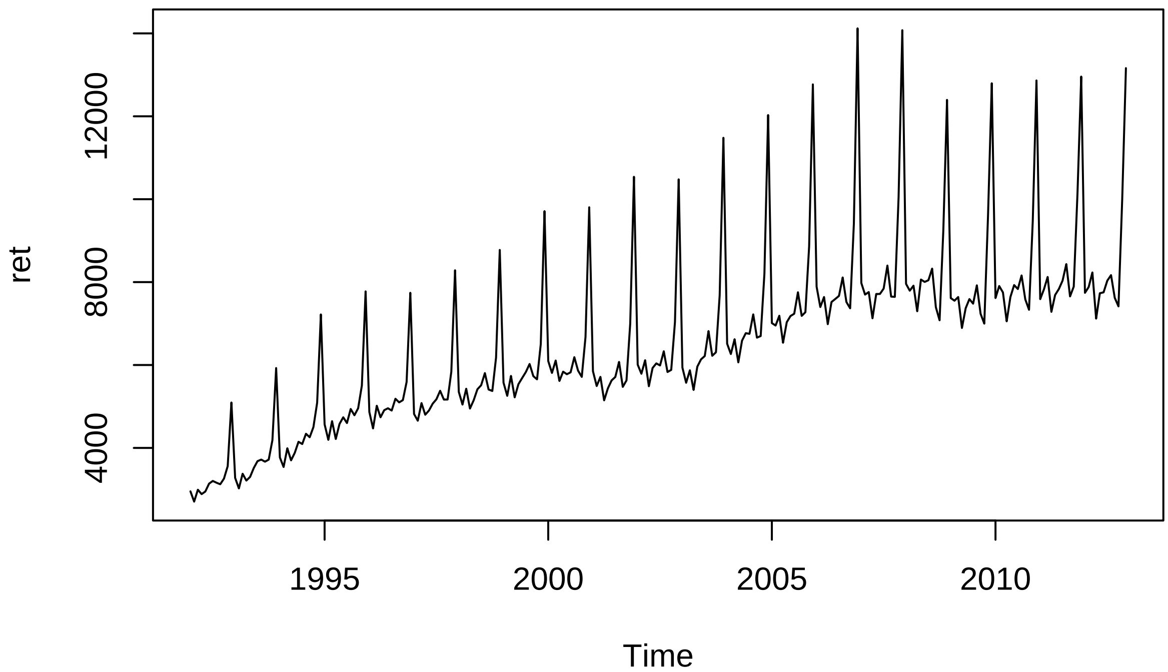

```r
plot(diffdiff.pop,xlab="Year",ylab="",yaxt="n",xaxt="n")
axis(1,cex.axis=.5)
axis(2,cex.axis=.5)
```



Exercise 8.40. Entropy-Increasing Transformation of Electronics and Appliance Stores

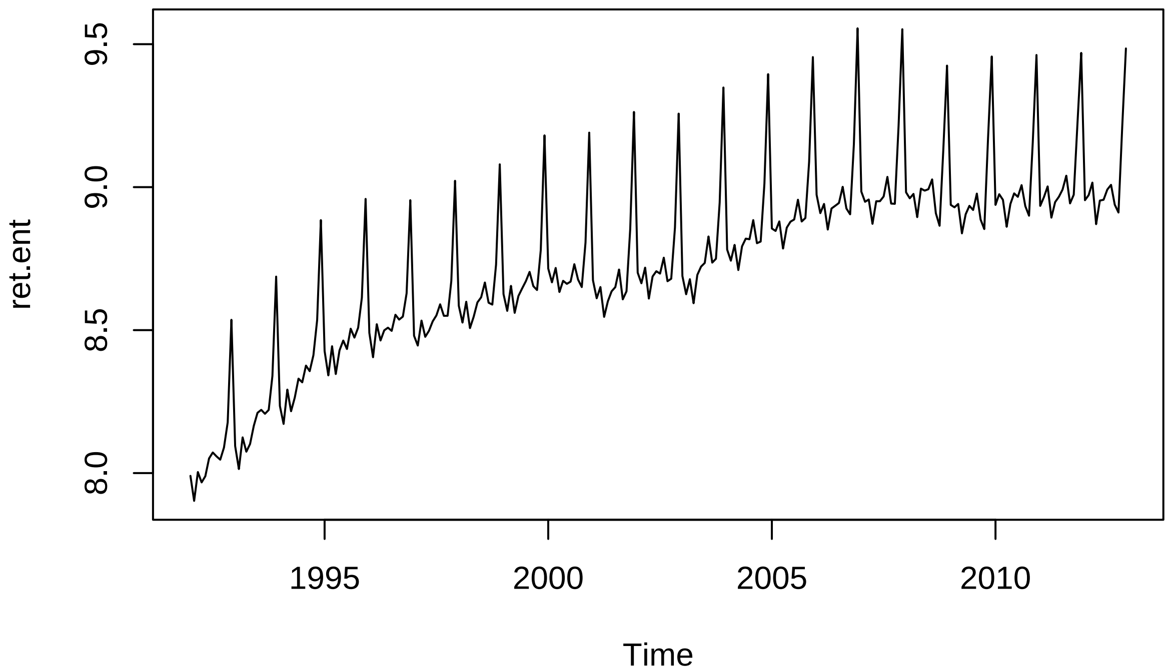
- What is an entropy-increasing transformation (or model) for the dataset of Electronics and Appliance Stores?
- First plot the data.

```
ret <- read.table("retail443.b1",header=FALSE,skip=2)[,2]
ret <- ts(ret,start=1992,frequency=12)
plot(ret)
```



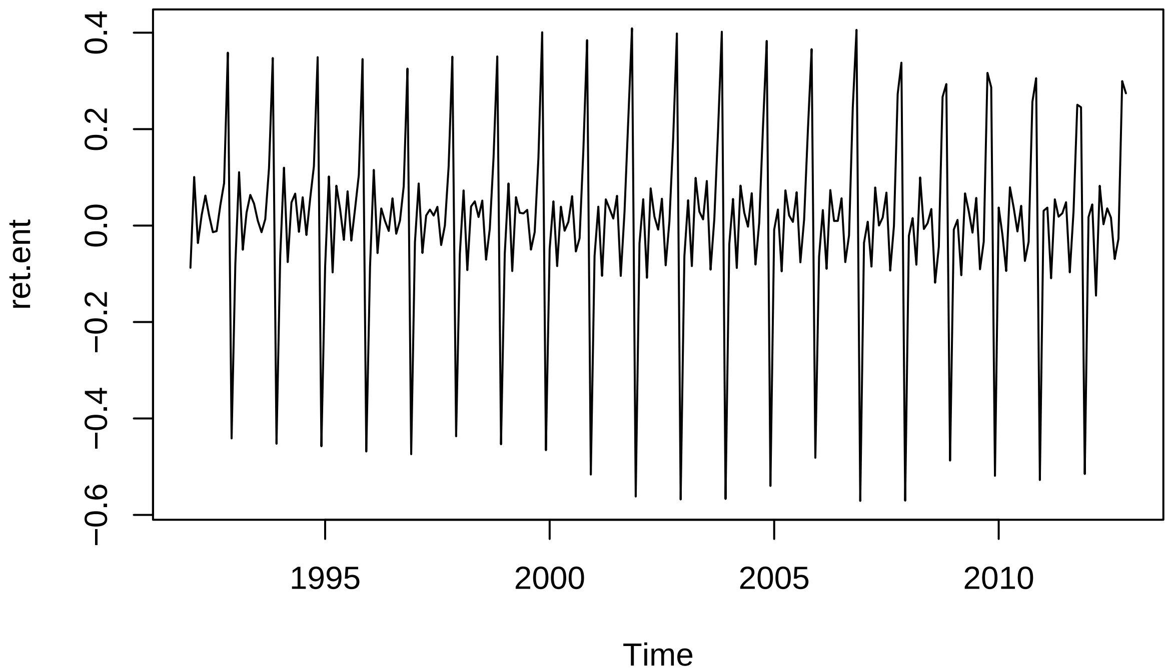
- Examine the log transformation.

```
ret.ent <- ts(log(ret),start=start(ret),frequency=12)
plot(ret.ent)
```



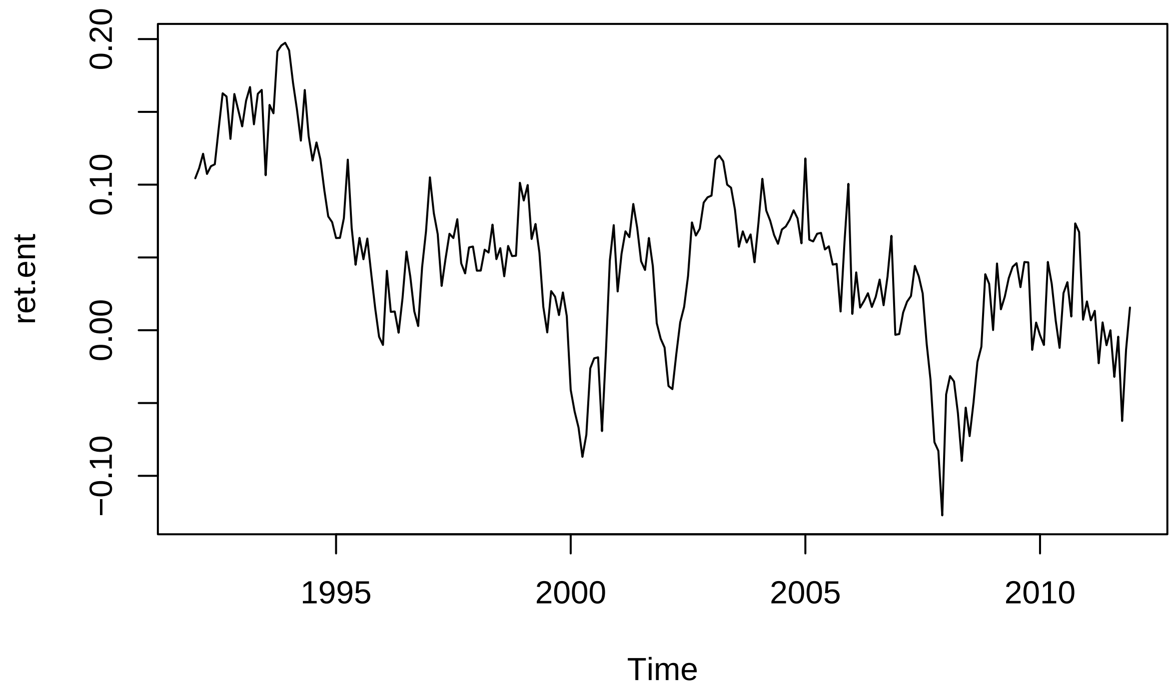
- Examine log differences.

```
ret.ent <- ts(diff(log(ret)),start=start(ret),frequency=12)
plot(ret.ent)
```



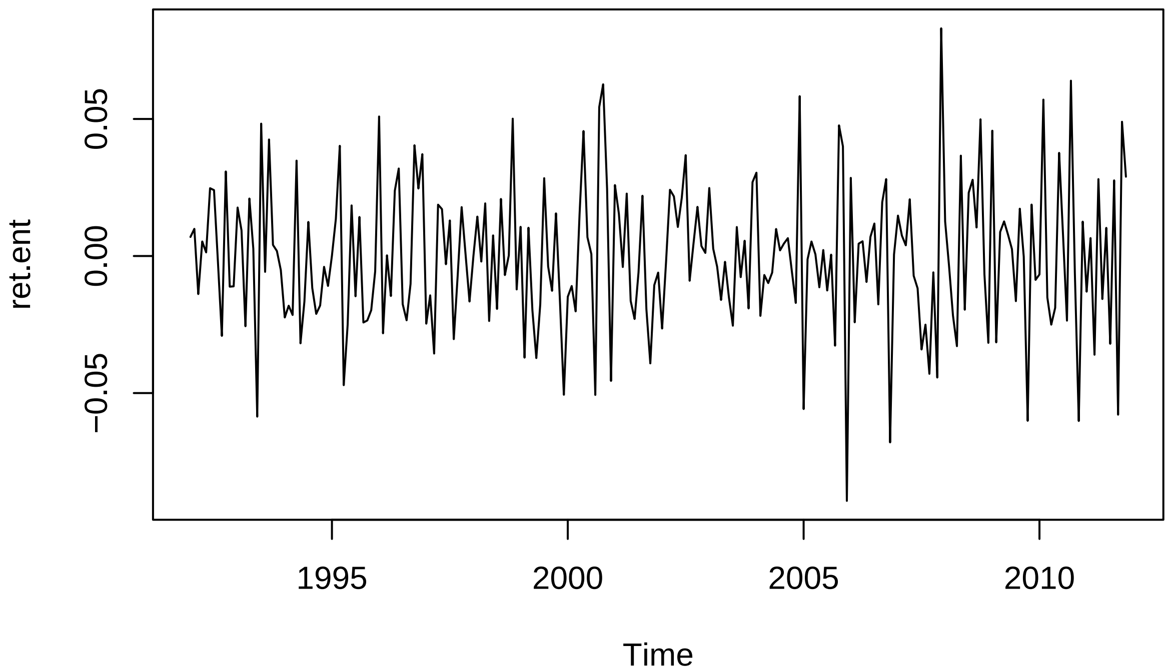
- Examine logs with seasonal differencing.

```
ret.ent <- ts(diff(log(ret),lag=12),start=start(ret),frequency=12)
plot(ret.ent)
```



- Examine logs with seasonal and nonseasonal differencing.

```
ret.ent <- ts(diff(diff(log(ret),lag=12)),start=start(ret),frequency=12)
plot(ret.ent)
```



Lesson 8-7: Kullback-Leibler Discrepancy

- We extend the idea of relative entropy, as a tool for modeling time series.

Example 8.7.2. Gaussian Relative Entropy

- Suppose $\underline{X}$ and $\underline{Y}$ are each samples of size n from stationary Gaussian time series, respectively with spectral densities f_x and f_y.
- It can be shown that their relative entropy, divided by n, has limiting value

$$n^{-1} H(\underline{X}; \underline{Y}) \to .5 \left(-1 + (2\pi)^{-1} \int_{-\pi}^{\pi} f_x(\lambda)/f_y(\lambda) d\lambda - (2\pi)^{-1} \int_{-\pi}^{\pi} \log[f_x(\lambda)/f_y(\lambda)] d\lambda \right).$$

- This is the analogue of entropy rate for relative entropy, for two time series.

Definition 8.7.3.

- The **Kullback-Leibler Discrepancy** between two stationary time series $\{X_t\}$ and $\{Y_t\}$ with spectral densities f_x and f_y is

$$h(f_x; f_y) = (2\pi)^{-1} \int_{-\pi}^{\pi} f_x(\lambda)/f_y(\lambda) d\lambda + (2\pi)^{-1} \int_{-\pi}^{\pi} \log[f_y(\lambda)] d\lambda.$$

- This looks like the relative entropy rate, but with the $\log f_x$ term omitted.
- Small values of this discrepancy correspond to closely aligned f_x and f_y.
- Think of $\{X_t\}$ as data process, and $\{Y_t\}$ gives a model; we try to describe given f_x with f_y drawn from a nice class (e.g., AR(p) spectral densities).

Example 8.7.4. The KL Distance for AR and MA Models.

- Suppose we try to model an MA(1) with an AR(1).
- So $f_x(\lambda) = |1 + \theta e^{-i\lambda}|^2 \sigma_x^2$, and $f_y(\lambda) = |1 - \phi e^{-i\lambda}|^{-2} \sigma_y^2$.
- Then
$$h(f_x; f_y) = \log \sigma_y^2 + \frac{\sigma_x^2}{\sigma_y^2} (2\pi)^{-1} \int_{-\pi}^{\pi} |1 + \theta e^{-i\lambda}|^2 |1 - \phi e^{-i\lambda}|^2 d\lambda.$$
- The expression in the integral is the spectral density of the MA(2) with polynomial $(1 + \theta z)(1 - \phi z) = (1 + (\theta - \phi) z - \theta \phi z^2)$, and so we obtain
$$h(f_x; f_y) = \log \sigma_y^2 + \frac{\sigma_x^2}{\sigma_y^2} (1 + (\theta - \phi)^2 + \theta^2 \phi^2).$$
- By calculus, the minimum value is $\phi = \theta/(1 + \theta^2)$. That is the best AR(1) approximation (via KL) to a given MA(1).
- Also the best σ_y^2 is $\sigma_x^2 (1 + (\theta - \phi)^2 + \theta^2 \phi^2)$. Plugging back in, the KL is then $\log \sigma_y^2 + 1$.

```
theta <- .5
sigma2.x <- 1
phi <- seq(-1,1,.01)
sigma2.y <- sigma2.x * (1 + (theta - phi)^2 + theta^2*phi^2)
my.kl <- log(sigma2.y) + 1
plot(ts(my.kl,start=-1,frequency=100),xlab="phi",ylab="KL")
phi.opt <- theta/(1 + theta^2)
abline(v = phi.opt,col=2)
```