Stat 153

Leo Villani

Fall 2024

1 Times Series and Measures of Dependence

1.1 Time Series vs Batch

In a time series data points are not I.I.D. In a batch the classic assumption is that it is I.I.D.
For example when considering climate forecasting if it is raining one second odds are it will be next second.
A markov process is a specific time series, order one markov processes only depend on yesterday.
We will study Time Domain vs Frequency Domain Methods

1.2 White Noise

$X_t, t = 1, 2, 3, ...$, R.V. representing time series
White Noise is a time series satisfying:

1. $Cov(X_s, X_t) = 0$ for all $s \neq t$
2. $E(X_t) = 0$, $Var(X_t) = \sigma^2$ for all t

Note: $Cov(X_s, X_t) = E[(X_s - E(X_s))(X_t - E(X_t))]$, $Cor(X_s, X_t) = \frac{Cov(X_s, X_t)}{\sqrt{(Var(X_s)Var(X_t))}}$
IID White Noise (White noise that is also IID)
Note: independence (IID) implies zero correlation but not other way
Gaussian White Noise
White noise that is jointly Gaussian (IID implied by this assumption), meaning if I give you $X_t, t = 1, ..., T$ as a vector that will be jointly Gaussian.
Note: 0 correlation and jointly gaussian implies independence but not in general.
Main Point: White noise has no trend (unpredictable) and is 0-mean so just a building block

1.3 Random Walk

$X_t = X_{t-1} + w_t$ where w_t is white noise (now there is dependence but still no trend)
Now with drift (trend): $X_t = \delta + X_{t-1} + w_t$, $\delta > 0$
Hence $X_t = t\delta + X_0 + w_1 + ... + w_t$. Assuming we start the random walk at $X_0 = 0$, clearly we are moving with δt and we have a sum of white noise. As t increases, variance increases because of the sum of white noise variances while $E(X_t) = \delta t$
$Var(X_t) = \sum_{i=1}^{t} Var(w_i) + \sum_{i \neq j} Cov(w_i, w_j) = \sigma^2 t + 0$
Note: Mean function denoted as $\mu_{X,t} = E(X_t)$ and variance function denoted $\sigma^2_{X,t} = Var(X_t)$
Also note: Autocovariance function denoted $\delta_X(s,t) = Cov(X_s, X_t)$, autocorrelation $\rho_X(s,t) = Cor(X_s, X_t) = \frac{\delta_X(s,t)}{\sigma_s \sigma_t}$

Using random walk with drift lets calculate the autocov, and autocorr:
$\delta_X(s,t) = Cov(X_s, X_t) = Cov(\delta s + \sum_{i=1}^{s} w_i, \delta t + \sum_{j=1}^{t} w_j)$, if $s < t$ then we get $Cov(\sum_{i=1}^{s} w_i, \sum_{j=1}^{s} w_j) + Cov(\sum_{i=1}^{s} w_i, \sum_{j=s+1}^{s} w_j) = \sigma^2 s + 0 = \sigma^2 s$ as the first term is just the variance of a random walk with drift.

In other words $\delta_X(s,t) = \sigma^2 s$ for $s \leq t$ and $\delta_X(s,t) = \sigma^2 t$ for $s > t = \sigma^2 * min(s,t)$
Then $\rho_X(s,t) = \frac{\sigma^2 min(s,t)}{\sigma\sqrt{(s)}\sigma\sqrt{(t)}} = \frac{min(s,t)}{\sqrt{(s)}\sqrt{(t)}}$
As s, t get large this will approach 1 as a lot of history between them so the r.v's are highly correlated.

Fun Example (drunk bird problem):
$X_t = X_{t-1} + w_t, w_t \sim N(0, \sigma^2)$ lives in $\mathbf{R}^d$ (just a random walk in d-dimensions)
If $d = 1$ the probability of return is 1. Same with $d = 2$. $d \geq 3$ then we return with probability 0.
This phenomenon is equivalent to the James-Stein paradox.

1.4 Linear Filtering

Given a time series x_t, $t = 1, 2, 3, ...$
Linear Filtering uses weights: $..., a_{-2}, a_{-1}, a_0, a_1, ... \in \mathbf{R}$
$y_t = \sum_{i=-\infty}^{\infty} a_i x_{t-i}$
Examples:
Moving Averages

1. Trailing Average: $y_t = \frac{1}{3}(x_t + x_{t-1} + x_{t-2})$, where $a_0 = a_1 = a_2 = \frac{1}{3}$ and all other weights are 0.

2. Centered Average: $y_t = \frac{1}{3}(x_{t+1} + x_t + x_{t-1})$
   Calculation Example:
   Assume x_t is white noise $\sim N(0, \sigma^2)$. Then $\mu_{y,t} = E(y_t) = \frac{1}{3}(0 + 0 + 0) =$

0. $\sigma_{y,t}^2 = Var(y_t) = \frac{1}{9}(Var(x_{t+1}) + Var(x_t) + Var(x_{t-1})+$ Covariances between each distinct pair) $= \frac{1}{9}(Var(x_{t+1}) + Var(x_t) + Var(x_{t-1})) = \frac{\sigma^2}{3}$
$Cov(y_s, y_t) = Cov(\frac{1}{3}(y_{s-1}+y_s+y_{s+1}), \frac{1}{3}(y_{t-1}+y_t+y_{t+1}))$. We can do this by casework. When then are no overlapping indices then the $\delta_y(s,t) = 0$. There will be a lot of other cases but you can imagine when there is more overlap there will be a stronger covariance. The $s = t$ case is just $\frac{\sigma^2}{3}$ as computed via the variance. For $s = t+1, t-1$ we get $\frac{2\sigma^2}{9}$ and $s = t+2, t-2$ we get $\frac{\sigma^2}{9}$.

1.5 General Facts

$E(\sum_i a_i x_i) = \sum_i a_i E(x_i)$
$Cov(\sum_i a_i x_i, \sum_j b_j y_j) = \sum_{i,j} a_i b_j Cov(x_i, y_j)$

1.6 Stationarity

Strong vs Weak
$x_t, t = 1, 2, 3, ...$ is called **strongly stationary** if $(x_{t_1}, ..., x_{t_k}) \stackrel{\text{d}}{=} (x_{t_1+l}, ..., x_{t_k+l})$ $\forall l, k, t_1, ..., t_k$
This definition implies the following:

1. $x_s \stackrel{\text{d}}{=} x_t \ \forall s, t$ hence $\mu_x(t) = \mu$
2. $(x_s, x_t) \stackrel{\text{d}}{=} (x_{s+l}, x_{t+l}) \ \forall s, t, l$ hence $Cov(x_s, x_t) = Cov(x_{s+l}, x_{t+l})$ or $\delta_x(s,t) = \delta_x(s+l, t+l)$ thus δ_x only depends on $t - s$ (gap)
3. all triplets are equal in distribution
4. ...

$x_t, t = 1, 2, 3, ...$ is called **weakly stationary** if $\mu_x(t) = \mu, \forall t$ and $\delta_x(s,t)$ depends only on $t-s$ the gap (these are the two implications of strong stationarity)
Note: **strong stationarity** $\Rightarrow$ **weak stationarity**
Counterexample for "only if": Suppose we have a time series of independent random variables so covariance is always 0 hence depending only on gap. Assume we are sampling our random variable values from different distributions all with mean 0. Then we definitely are not "immune" to time shifts in distribution.
Under weak stationarity we write $\delta_x(h)$ as the auto-covariance function where $h = t - s$ as it only depends on the lag/gap
Note: $\sigma_x^2(t) = \delta_x(0) = \sigma^2$, i.e. constant in time so variance is constant
Remark: for a Gaussian process (jointly Gaussian x_t), **weakly stationary** $\Rightarrow$ **strongly stationary**
Intuitive Proof: the joint gaussian density function looks something like $exp(-(x-$

$\mu)^T\Sigma^{-1}(x-\mu))$ which only depends on the μ and Σ so the joint distribution is invariant to lag

1.7 Covariance Estimation

how to estimate auto-covariance $\delta_x(h)$ First for lag $h = 0$, $\delta_x(0) = \sigma^2$ so given samples $x_1, ..., x_n$ we would just compute the sample variance given by $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, hence $\hat{\delta}_x(0) = \hat{\sigma}^2$
Now for $h = 1$, we have:

$$Cov(x_t, x_{t+1}) = E((x_t - \mu)(x_{t+1} - \mu)) \approx E((x_t - \bar{x})(x_{t+1} - \bar{x}))$$
$$Cov(x_1, x_2) = E((x_1 - \mu)(x_2 - \mu)) \approx E((x_1 - \bar{x})(x_2 - \bar{x}))$$
$$...$$

so we can just average these approximations and get the following

$$\hat{\delta}_x(1) = \frac{1}{n}\sum_{t=1}^{n-1}(x_t - \bar{x})(x_{t+1} - \bar{x})$$

Generalizing to estimate the auto-covariance $\delta_x(h)$ for any gap we get

$$\hat{\delta}_x(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_t - \bar{x})(x_{t+h} - \bar{x})$$

Now for the sample auto-correlation $\hat{\rho}(h) = \frac{\hat{\delta}(h)}{\hat{\delta}(0)}$
Note: In R we use $acf()$ with type = "covariance" or "correlation"

1.8 Cross covariance/correlations

Auto: one series vs. cross: two series
Letting $x_t, t = 1, 2, 3, ...$ and $y_t, t = 1, 2, 3, ...$ be two time series then

1. Cross-covariance: $\delta_{xy}(s, t) = Cov(x_s, y_t)$
2. Cross-correlation: $\rho_{xy}(s, t) = Cor(x_s, y_t) = \frac{\delta_{xy}(s,t)}{\sigma_x(s)\sigma_y(t)}$

Note: in general not symmetric functions as we are using two different time series
If $\delta_{xy}(s, t)$ is consistently large when $s < t$ then we say "x leads y", intuitively knowing something about x_t at some value helps you understand y_t in the future.

1.9 Gaussian Processes

Given $x_t, t = 1, 2, 3, ...$ and $(x_{t_1}, ..., x_{t_k}) \sim JointlyGaussian$ where $\mu_i = E(x_{t_i})$, $\Sigma_{ij} = Cov(x_{t_i}, x_{t_j})$. The density $N(\mu, \Sigma)$ where μ is a $kx1$ vector and Σ is a kxk constant matrix. The desnity is entirely determined by μ, Σ. Hence weak stationarity for GP's implies strong stationarity.

2 Regression

2.1 Simple Linear Regression

y: response, x: covariate, predictor, feature
$y \approx \beta_0 + \beta_1 x$
Population-Level Suppose we dont have just samples but we have access to the random variables X,Y distributions and joint distribution.
Find β_0^*, β_1^* to minimize:

$$E[(y - \beta_0 - \beta_1 x)^2] := Q(\beta_0, \beta_1)$$

If convex we can take derivatives and set equal to 0 to find minimzers. The expectation above is convex so:

$$\frac{dQ}{d\beta_0} = E[\frac{d}{d\beta_0}(y - \beta_0 - \beta_1 x)^2] = 2E[\beta_1 x + \beta_0 - y] = 0$$
$$\beta_0 = E[y - \beta_1 x] = E[y] - \beta_1 E[x]$$
$$\frac{dQ}{d\beta_1} = E[\frac{d}{d\beta_1}(y - \beta_0 - \beta_1 x)^2] = 2E[x(\beta_1 x + \beta_0 - y)] = 0$$
$$\beta_1 = \frac{E[xy] - \beta_0 E[x]}{E[x^2]} = \frac{E[xy] - E[x]E[y] + \beta_1 E[x]^2}{E[x^2]}$$
$$\beta_1(E[x^2] - E[x]^2) = Cov(x, y)$$

which implies that $\beta_0^* = E[y] - \beta_1 E[x], \beta_1^* = \frac{Cov(x,y)}{Var(x)} = Cor(x, y)\frac{\sqrt{Var(y)}}{\sqrt{Var(x)}}$

Note: regression is inherently "assymetric", so we say "regression of y on x"

Sample-Level Now we have data points instead of the random variables: $(x_i, y_i), i = 1, ..., n$ with the goal of finding $\hat{\beta}_0, \hat{\beta}_1$ to minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

after taking derivatives, setting equal to 0, and solving we get

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 = \frac{\hat{Cov}(x,y)}{\hat{Var}(x)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

which is the "same" as the population level values. So this is actually a plug in estimator because I could have just taken sample estimations and plugged them into the population level β^* values.

Prediction Types

$$\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

"ex-ante": true forecast which means it was made with information available at the time
"ex-post": prediction made with feature available later