Notas

Adriana Pérez-Arciniega Soberón

2016

0.1 Notas Sesiones

Tenemos una muestra de microcostos de enfermedades crónicas de un cierto número de individuos los cuales tienen asociadas un número de covariables sociodemográficas, socioeconómicas y médicas. El objetivo es modelar la duración y el costo de las etapas de estos padecimientos por individuo.

Si tenemos una serie de individuos $i = 1, ..., n$ observados por un período de tiempo con costos asociados a su padecimiento, tendríamos entonces dos variables $\{(d_i, c_i)\}_{i=1}^{n_T}$ con d_i como duración y c_i como el costo, estas asociados al individuo i. Así que este modelo estaría determinado por datos de duración y costos, que se modelaría como un proceso **Poisson marcado** con:

- c_j's positivas.
- d_j's positivas.

Supongamos que empezamos el estudio de un individuo en el t_0 que es el tiempo en el que el paciente entra al panel de estudio y hacemos "cortes" en los k momentos de ocurrencia, definiremos el momento de ocurrencia como el momento que ocurra un cambio de tratamiento; tenemos el fin de nuestro horizonte de tiempo en t, esto no quiere decir que es el momento en el que cesan las observaciones, por eso tendremos datos censurados a la derecha. Por esto, definiremos $N(t)$ como la variable aleatoria de el número de cambios o cortes en el intervalo $(t_0, t]$ y a cada uno de estos k tiempos de ocurrencia están asociadas un par de variables $\{d_1, c_1\}, ..., \{d_k, c_k\}$, es decir, que a cada momento de ocurrencia tenemos variables aleatorias de duración y costo. Es importante remarcar que una muestra de n individuos, cada uno tendrá un t_0 "distinto" en tiempo calendario, pero para efectos de este estudio podemos "alinear" los t_0's.
De manera general, lo que estaríamos buscando es un **proceso estacionario de Markov** para las variables $\{d_k, c_k\}$ donde $k \in (t_0, t]$ y así poder hacer inferencias de observaciones futuras. De inicio sabemos que la variable de costo es dependiente de la duración, por lo que como un primer acercamiento tendríamos:

$$P(d_k, c_k) = P(d_k)P(c_k|d_k)$$

Si a esta idea añadimos el proceso Markov marcado vemos que:

$$P(d_k, c_k|d_{k-1}, c_{k-1}) = P(d_k|d_{k-1})P(c_k|d_k, c_{k-1})$$

Las marcas están dadas por el momento de ocurrencia, es decir, la duración en el tiempo k condicionadas a la duración del tiempo $k-1$, mientras que el costo es asociado a la duración actual y al costo de la ocurrencia pasada, logrando la estacionareidad podremos inferir la duración y el costo futuro.

Sea y_i una variable aleatoria respuesta para cada individuo i. Cada individuo i tiene asociada información adicional x_i, por lo que $x_i = (x_{i1}, ..., x_{ip})$ pertenece a $\mathbb{R}^p$.

Necesitamos una función de distribución de la variable respuesta dado las covariables tal que el valor esperado sea una función h de las covariables individuales por β.
$y_i|x_i$ con función de distribución:

$$F(y_i|x_i) \text{ tal que } E(y_i|x_i) = h(x_i'\beta)$$

Tomemos y_i en $\mathbb{R}_+$ y $x_i'\beta$ es un proyector lineal: $\mathbb{R}^p \to \mathbb{R}$ por lo que se necesita $h(\cdot)$ tal que $h : \mathbb{R} \to \mathbb{R}_+$.
Por lo cual podemos definir

$$E(y_i|x_i) = exp(x_i'\beta)$$

Dado que como sabemos, la función exponencial siempre nos arrojará un resultado positivo. Esto sería un **modelo lineal generalizado**.

Tomemos como primer acercamientos a las d_j's (duraciones) de un solo individuo. Podemos pensar en una función de densidad:

$$f(d_i) = h(d_i)S(d_i)$$

donde $h(d_i)$ es la función hazard y $S(d_i)$ es la función de supervivencia.
Ligando el proceso Markov marcado explorado anteriormente podríamos decir que la duración actual depende de la duración anterior:

$$f(d_i|d_{i-1}) = h(d_i|d_{i-1})S(d_i|d_{i-1})$$

Así, si tomamos un individuo i:
$f(d_{ji}|x_i) = h(d_{ji}|x_i)S(d_{ji}|x_i)$. Esto se define como un **Modelo de Cox**.
Entonces,
$f(d_{ji}|x_i) = h_b(d_{ji})\varphi(x_i'\beta)S(d_{ji})$
En el contexto de un **Proceso Markov Marcado**:
$f(d_{ji}|d_{j-1,i}, x_i) = h_b(d_{ji}|d_{j-1,i})\varphi(x_i'\beta)S(d_{ji})$
En donde:

- $h_b(\cdot)$ se le conoce como la función hazard base que es igual para cualquier duración de cualquier individuo, es conocido como **baseline model**.

- $\varphi(\cdot)$ es fija y conocida como la función exponencial y son comunes a **todas** las observaciones de cada individuo i.

- β es común para cualquier individuo i

Supongamos que k_j como el número de observaciones del individuo j y n el número de individuos en la muestra.

$$\begin{bmatrix} d_{11}c_{11} & d_{12}c_{12} & \dots & d_{1k}c_{1k} \\ d_{21}c_{21} & d_{22}c_{22} & \dots & d_{2k}c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1}c_{n1} & d_{n2}c_{n2} & \dots & d_{nk}c_{nk} \end{bmatrix}$$

La función que liga las duraciones del mismo individuo es la **función hazard** $h_b(\cdot)$, que podemos estimar empíricamente con el método Kaplan-Myer. Y la función que liga las duraciones entre individuos es la $\varphi(x_i'\beta)$. La β es el parámetro a estimar, pero es igual común para todos los individuos i.

Ya que tenemos las funciones comunes a todos los individuos, podemos hacer predicción con la entrada de cualquier individuo x_{i+1} en un tiempo fututo f. Entonces tendríamos

$$\mathbb{P}(d^f_{i+1,1}, d^f_{i+1,2}, ..., d^f_{i+1,n}) = h_b(d^f_{i+1})\varphi(x^f_{i+1}\beta) \prod_{j=2}^{N(t)} h(d^f_j | d^f_{j-1})\varphi(x^f_{i+1}\beta)$$

.

Suponemos una sucesión de variables aleatorias $\{X_t\}_{t=1}^T$ donde el orden de las observaciones no importa:
Bajo un supuesto de **independencia** tendríamos que la distribución conjunta es la multiplicación de las distribuciones marginales.

$$\mathbb{P}(X_1, ..., X_T) = \prod_{t=1}^{T} \mathbb{P}(X_t)$$

Por lo que la distribución de la observación $T+1$ dado las observaciones pasadas era también la distribución de la observación $T+1$.

$$\mathbb{P}(X_{T+1}|X_1, ..., X_T) = \mathbb{P}(X_T + 1)$$

Es decir, las variables aleatorias tienen una función de probabilidad común. Bajo un supuesto de **intercambiabilidad** tendríamos que la distribución conjunta de la sucesión de v.a.'s con una permutación σ (donde $\{\sigma(1), ..., \sigma(T)\}$ es una permutación de $\{1, ..., T\}$) es igual a la distribución conjunta original.

$$\begin{aligned} \mathbb{P}(X_1, ..., X_T) &= \mathbb{P}(X_{\sigma(1)}, ..., X_{\sigma(T)}) \\ &= \int \prod_{t=1}^{T} \mathbb{P}(X_t|\theta)\pi(\theta)\mathrm{d}\theta \end{aligned}$$

donde θ es una variable aleatoria no observable y $\pi(\theta)$ es una medida de probabilidad común a las variables aleatorias.

Bajo un supuesto de **estacionareidad** estaríamos diciendo que el orden de las observaciones importa y que las observaciones pasadas ayudan a construir la variable aleatoria, lo cual lo hace un supuesto crucial para poder predecir.

$$\begin{aligned}\mathbb{P}(X_T, ..., X_1) = \mathbb{P}(X_T|X_{T-1}, ..., X_1) \\ *\mathbb{P}(X_{T-1}|X_{T-2}, ..., X_1) \\ \vdots \\ *\mathbb{P}(X_2|X_1) * \mathbb{P}(X_1)\end{aligned}$$

Usando este mismo concepto, podríamos predecir la observación $T+1$

$$\mathbb{P}(X_{T+1}|X_1, ..., X_T) = \mathbb{P}(X_{T+s+1}|X_{1+s}, ..., X_{T+s}) \quad \forall s \geq 0$$

Suponemos que no conocemos la relación entre las observaciones $(X_t)'s$, pero podemos suponer que hay un parámetro (variable aleatoria) no observable que es lo que conecta las observaciones.
De tal manera se construye el modelo:

$$X_t|\theta_t \sim \mathbb{P}(X_t|\theta_t) \tag{1}$$

donde X_t es la parte observable y la θ_t lo no observable que tendría la siguiente forma de un modelo de Markov estacionario:

$$\theta_t|\theta_{t-1} \sim \mathbb{P}(\theta_t|\theta_{t-1})$$

Por lo que la distribución conjunta de $\{X_t\}_{t=1}^T$

$$\mathbb{P}(X_1, ..., X_T) = \int_{\mathbb{R}^T} \prod_{t=1}^{T} \mathbb{P}(X_t|\theta_t)\mathbb{P}(\theta_t|\theta_{t-1})\mathrm{d}\theta_1\mathrm{d}\theta_2...\theta_T$$

- $\{X_t|\theta_t\}$ son condicionalmente independientes.
- $\{\theta_t\}$ son markovianos estacionarios.
- $\{X_t\}_{t=1}^T$ son Markov marginalmente.

Usaríamos una metodología más general que el modelo ARIMA (p,q,r) que es mejor para predecir, estos son los **modelos State-Space**. De este modo,

$$\mathbb{P}(X_{T+1}|X_T, ..., X_1) = \int_{\mathbb{R}} \mathbb{P}(X_{T+1}|\theta_{T+1})\pi(\theta_{T+1}|X_1, ..., X_T)\mathrm{d}\theta_{T+1}$$

El primer término de la integral se refiere a (1) y la segunda parte sería la distribución de aprendizaje pues el parámetro no observable se construye a partir de las observaciones anteriores.

En nuestro contexto con las d_t's:

$$\mathbb{P}(d_1, ..., d_T) = \mathbb{P}(d_1) \prod_{t=2}^{T} \mathbb{P}(d_t|d_{t-1}) = \Gamma(d_1|\alpha_0, \beta_0) \prod_{t=2}^{T} \Gamma(d_t|\alpha d_{t+1}, \beta)$$

Este modelo difiere del propuesto por Russell y Engel pues ellos proponen un modelo:

$$\log d_t = \alpha_0 + \alpha_1 \log d_{t-1} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2)$$

es decir,

$$\log d_t | \log d_{t-1} \sim N(\log d_t | \alpha_0 + \alpha_1 d_{t-1}, \sigma^2)$$

ó

$$d_t|d_{t-1} \sim log - Normal(d_t| \exp\{\alpha_0 + \alpha_1 d_{t-1}\}, \tilde{\sigma}^2)$$

El problema con este modelo es que no satisface una de las características cruciales para la resolución de nuestro problema crucial de la predicción de observaciones futuras que es la estacionareidad. En el paper de Russell y Engel se resuelve este problema condicionando los parámetros α_0 y α_1 a $|\alpha_0 + \alpha_1| \leq 1$. Por lo que el modelo $d_t|d_{t-1} \sim \Gamma(d_t|\alpha d_{t-1}, \beta)$ es preferible dado que es estacionario por construcción logrando que la estimación sobre α a β es directa.

0.2 Regression Models and Life-Tables, Cox y Autoregressive Conditional Duration, Engel y Russell

Si consideramos una población de individuos y para cada individuo observamos el "tiempo de ocurrencia" o duración. Sea T una variable aleatoria que representa el tiempo de ocurrencia y definamos la función de supervivencia $\mathbb{F}(t)$:

$$\mathbb{F}(t) = \mathbb{P}(T \geq t)$$

y sea $\lambda(t)$ la función hazard según la edad.

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | t \leq T)}{\Delta t}$$

Es decir, que la función hazard es la función que nos dice la probabilidad de que el tiempo de ocurrencia suceda en el tiempo t.
Supongamos una población de n individuos y que cada uno de los individuos tiene p variables asociadas $z_1, ..., z_p$, así para el j-ésimo individuo tiene el vector $\mathbf{z}_j = (z_{1j}, ..., z_{pj})$, supongamos que las z's son asociadas al tiempo. Entonces,

tendríamos una función hazard:

$$\lambda(t; z) = \exp(z\beta)\lambda_0(t)$$

En esta función β es un vector de parámetros desconocidos y $\lambda_0(t)$ es la función hazard para un conjunto de condiciones **z=0**. Ligándolo con lo anterior en **Notas de Sesiones**, podemos ver que $\lambda_0(t)$ es equivalente a la $h_b(\cdot)$ y la $exp(z\beta)$ es equivalente a $\varphi(x_i'\beta)$.
Tomemos los distintos tiempos de ocurrencia como $t_1 < ... < t_k$, sea m_i el número de fracasos en t_i tal que $\sum m_i = n$ y r_i el número de individuos en riesgo al tiempo t_i; entonces tendríamos un estimador de la función hazard:

$$\tilde{\lambda}(t) = \sum_{i=1}^{k} \frac{m_i}{r_i}\delta(t - t_i)$$

Esta fórmula sería análoga a lo que se presenta en Engel y Russell como el tiempo de ocurrencia modelando el tiempo entre eventos y se podría estimar mediante el método de Kaplan-Meier.

En el paper de Engel y Russell exploran un modelo donde el tiempo entre los eventos es tratado como un proceso estocástico y propone una nueva clase de proceso marcado con tasas de llegada dependientes. Como este modelo se enfoca en la duración esperada entre eventos, este modelo se llama **Autorregresión con Duración Condicional (ACD)**. El paper de Engel y Russell habla de microfinanzas y de cómo es imposible analizar el trading de alta frecuencia en un espacio de tiempo fijo, por lo que se propone tomar el tiempo de ocurrencia como variable aleatoria que sigue un proceso puntual; asociado al tiempo de ocurrencia hay otras variables aleatorias a las que llamamos **marcas**, en el caso de microfinanzas estas marcas corresponderían al volumen, a la diferencia en el precio de oferta y demanda o el precio; en el contexto de la tesis, estas marcas sería el costo del padecimiento.
La intensidad condicional se parametriza en términos de eventos pasados, por lo que la formulación básica de este modelo es la dependencia de la intensidad condicional en duraciones pasadas, es por eso que lo llamamos el modelo de autorregresión con duración condicional. Asociado con la intensidad está la esperanza condicional del tiempo de espera al siguiente evento. Es importante notar que el modelo está formulado en **tiempo de ocurrencia** pero modela la frecuencia y distribución del tiempo calendario entre eventos.
Consideremos un proceso estocástico $\{t_0, ..., t_n\}$ con $t_0 < t_1 < ... < t_n$ que es una secuencia de tiempos de ocurrencia dependientes en el tiempo, asociado a estos tiempos de ocurrencia tenemos la función $N(t)$ que es el número de eventos que han ocurrido para el tiempo t, si hay características asociadas al tiempo de ocurrencia se le conoce como un **proceso puntual marcado**. Hay dos generalizaciones de un proceso puntual:

- Proceso puntual que evoluciona sin efectos secundarios:
  Si para cualquier $t > t_0$ la realización de puntos durante $[t, \infty)$ no depende

de la secuencia de puntos en el intervalo $[t_0, t)$.

- Proceso puntual condicionalmente ordenado:
  Si en $t \geq t_0$ hay un intervalo de tiempo suficientemente corto y condicional a cualquier evento P definido por la realización del proceso en $[t_0, t)$, que la probabilidad de dos o más eventos es infinitesimalmente relativa a la probabilidad de un evento.

Para este trabajo nos concentraremos en procesos puntuales que evolucionen con efectos secundarios y que sean condicionalmente ordenados.
Una función de intensidad de un proceso puntual "self-exciting", esto es, un proceso donde el pasado impacta la posible estructura de eventos futuros tiene la siguiente forma:

$$\lambda(t|N(t), t_1, ..., t_{N(t)}) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(N(t + \Delta t) > N(t)|N(t), t_1, ..., t_{N(t)})}{\Delta t}$$

Hay dos maneras de interpretar esta función de intensidad:

- Por tiempo calendario:

$$\lambda(t|N(t), t_1, ..., t_{N(t)}) = \omega + \sum_{i=1}^{N(t)} \pi(t - t_i)$$

  Cada t_i que pasa contribuye con $\pi(t - t_i)$ a la intensidad del tiempo t, π es una medida infectológica. Este modelo tiene el efecto marginal de que un evento que ocurrió hace X tiempo es independiente de lo que sucede sin importar cuántos eventos hayan sucedido.

- Por tiempo de ocurrencia modelando el tiempo entre eventos:

$$\lambda(t|N(t), t_1, ..., t_{N(t)}) = \omega + \sum_{i=1}^{N(t)} f_i(t_{N(t)+1-i} - t_{N(t)-i})$$

  así el impacto de la duración entre eventos sucesivos depende del número de eventos que intervienen.

Este tipo de modelos es estudiado por Cox en el marco de modelo de hazard proporcionales, donde la función de intensidad condicional de este modelo sería la descrita al inicio de esta sección donde z_i es el vector de variables explicatorias asociadas al **tiempo de llegada** i y el tiempo de falla, en este caso, esta condicionado por el vector de variables asociadas al número de eventos ocurridos para el tiempo t.

$$\lambda(t|z_{N(t)}, ..., z_1) = \lambda(t) \exp(\beta' z_{N(t)})$$

El modelo ACD está especificado en términos de la densidad condicional de las duraciones. Sea $x_i = t_i - t_{i-1}$ el intervalo de tiempo entre dos realizaciones,

es decir, la duración. La densidad de x_i condicionado en las x's pasadas sean especificadas directamente.
Sea ψ_i la esperanza de la duración i dado por:

$$E[x_i|x_{i-1},...,x_1] = \psi_i(x_{i-1},...,x_1) = \psi_i$$

Así, la esperanza condicional de la duración depende de las duraciones pasadas. Asumimos, también, que:

$$x_i = \psi_i \epsilon_i; \quad \epsilon_i \sim p(\epsilon, \phi)$$

p es cualquier función de densidad y ϕ tiene varianza libre.
Sea p_0 la función de densidad de ϵ y sea S_0 su función de supervivencia asociada. Definimos como riesgo basal:

$$\lambda_0(t) = \frac{p_0(t)}{S_0(t)}$$

$$\Rightarrow \lambda(t|N(t), t_1, ..., t_{N(t)}) = \lambda_0(\frac{t - t_{N(t)}}{\psi_{N(t)+1}})\frac{1}{\psi_{N(t)+1}}$$

Este modelo es conocido como modelo de vida acelerada dado que la información pasada influencia el ritmo en el que el tiempo pasa. La velocidad del tiempo depende de los tiempos de ocurrencia pasados a través de la función ψ. La versión más simple del modelo ACD asume que las duraciones son exponenciales condicionalmente.

$$\lambda(t|x_{N(t)}, ..., x_1) = \psi^{-1}_{N(t)+1}$$

Con una memoria de m, la intensidad condicional implica que solo los momentos más recientes han influenciado la duración:

$$\psi_i = \omega + \sum_{j=0}^{m} \alpha_j x_{i-j}$$

En general,tomaremos el modelo ACD (m, q) que se refiere al orden de desfases, este modelo es conveniente porque permite calcular varios momentos.

$$\psi_i = \omega + \sum_{j=0}^{m} \alpha_j x_{i-j} + \sum_{j=0}^{q} \beta_j \psi_{i-j}$$

El modelo ACD es propuesto como un modelo con tiempos de ocurrencia correlacionados intertemporalmente. Para examinar la dependendia, calculamos las autocorrelaciones y las autocorrelaciones parciales en el tiempo de espera entre eventos.
En el contexto de la tesis y de las duraciones, veríamos el modelo autorregresivo grado 1 como:

$$P(d_k|d_{k-1}) = exp\{\alpha_0 + \alpha_1 d_{k-1} + \epsilon\}$$

con las condiciones de que

$$|\alpha_0| < 1, \quad |\alpha_1| < 1, \qquad |\alpha_0 + \alpha_1| < 1$$

Con estas condiciones aseguramos la misma estacionareidad que necesitamos para hacer inferencia sobre observaciones futuras, al igual que en el modelo del proceso de Markov marcado.

0.3 The Analysis of Time Series, Chatfield

Una serie de tiempo es una colección de observaciones hechas de manera secuencial en el tiempo. Un proceso puntual es una serie de tiempo donde se considera una serie de eventos que ocurre de manera aleatoria en el tiempo. Una característica muy importante del análisis de series de tiempo es que las observaciones sucesivas no son independientes unas de otras por lo que para hacer el análisis debemos tomar en cuenta el orden de las observaciones y al ser dependientes podemos predecir observaciones futuras tomando en cuenta las observaciones pasadas. Existen varios objetivos por los cuales estudiar series de tiempo, por ejemplo:

- **Descripción** El primer paso a seguir una vez obtenidos los datos es graficarlos para obtener las principales características de la serie, pues nos puede decir si la serie muestra estacionalidad, tiene datos atípicos, puntos de inflexión, etc.
- **Explicación** Cuando estamos estudiando más de una variable, podemos usar la serie de tiempo para explicar el comportamiento de otra variable. En esta parte del análisis nos ayudaremos con los modelos lineales, que son aquellos que toman la serie insumo y a través de una operación lineal nos devuelve un resultado.
- **Predicción** Dada una serie observada es de nuestro interés predecir valores futuros de esta misma serie. Muchas veces la predicción está muy relacionada con problemas de control.
- **Control** Cuando la serie de tiempo mide la calidad de un proceso de manufactura, el objetivo del análisis puede ser controlar el proceso. Podemos usar modelos estocásticos para predecir valores futuros y las variables se ajustan para mantener el proceso en el objetivo.

Hay muchas herramientas que nos ayudan a analizar una serie de tiempo como técnicas descriptivas, modelos de probabilidad aunque el mayor instrumento de análisis son los modelos de autocorrelación que describe la evolución del proceso a través del tiempo y la inferencia se hace con un análisis de dominio del tiempo, con un análisis de densidad espectral podríamos ver como las variaciones de la serie del tiempo pueden ser explicadas por componentes cíclicos a diferentes frecuencias y la inferencia a partir de este método se llama análisis en el dominio

de frecuencia.

Las series de tiempo pueden tener variaciones de la forma de efectos estacionales, cambios cíclicos, tendencias o cualquier fluctuación irregular, todas estas variaciones deben eliminarse y/o ajustarse para que no influyan con los resultados del análisis de la serie. Desde un punto de vista intuitivo, una serie de tiempo es **estacionaria** si no hay un cambio sistemático en la media (la serie no tiene tendencia), no hay cambios sistemáticos en la varianza y si las variaciones estrictamente periódicas han sido eliminadas; es decir, la serie se mueve dentro de una "banda". Para poder identificar si la serie tiene algún tipo de variación el primer paso es graficar la serie de observaciones contra el tiempo, pues con este análisis las características principales de la serie se muestran y también, si la serie es aproximadamente estacionaria es útil calculas la media y la desviación estándar de las observaciones.
Una vez que hemos graficado los datos, podemos darnos cuenta si la serie necesitaría una transformación, ya sea para estabilizar la varianza o bien para hacer aditivo el efecto estacional y el error.

Si observáramos una **tendencia** en nuestra serie de datos tendríamos que tomar una decisión sobre que es lo que se va a estudiar: el estudio de la tendencia misma o eliminar la tendencia para analizar las fluctuaciones locales, para lo que se necesitarían las técnicas adecuadas. Una de estas técnicas es ajustar una curva a la serie como una curva de Gompertz, lo cual nos permitiría medir la tendencia mientras los residuales nos darían un estimado de las fluctuaciones locales, donde los residuales son las diferencias entre las observaciones y los valores de la curva. Otra técnica seían los "filtros lineales" en la que se convierte una serie $\{x_t\}$ en otra mediante una operación lineal con unos ponderadores que suavizan las fluctuaciones y estiman la media local, esta técnica se recomienda para remover la variación estacional mas no para estudiar la tendencia de la serie. Por último, una manera de eliminar la tendencia es a través del método de diferencias para volver la serie estacionaria, que en el contexto de la tesis es la más probable a ser usada.

De igual modo, si al analizar la serie de tiempo ésta demostrará tener **fluctuaciones estacionales** tendríamos que decidir que es lo que se quiere hacer: medir los efectos estacionales o bien, eliminarlos. Para series que muestran una ligera tendencia, simplemente se calcula el promedio del período (mensual, trimestral, etc.) y se compara con el promedio general. Para series con una tendencia más marcada, la manera de eliminar el efecto estacional sería calcular el siguiente ponderador:

- Con información mensual:

$$Sm(x_t) = \frac{1/2x_{t-6} + x_{t-5} + x_{t-4} + ... + x_{t+5} + 1/2x_{t+6}}{12}$$

- Con información trimestral:

$$Sm(x_t) = \frac{1/2x_{t-2} + x_{t-1} + x_t + x_{t+1} + 1/2x_{t+2}}{4}$$

La suma de estos coeficientes debe ser igual a 1. Así, el efecto estacional puede ser estimado con $x_t - Sm(x_t)$ o $\frac{x_t}{Sm(x_t)}$ dependiendo si el efecto estacional es aditivo o multiplicativo; si el efecto estacional se queda del mismo tamaño independientemente de la media se dice que es aditivo, si este incrementa directamente proporcional a la media entonces es multiplicativo.
Los efectos estacionales también pueden ser eliminados por el método de diferencias, donde con información mensual se tendría:

$$\Delta_{12}x_t = x_t - x_{t-12}$$

Los modelos de probabilidad que corresponden a las series de tiempo son llamadas procesos estocásticos, un proceso estocástico es definido como una colección de variables aleatorias $\{X(t), t \in T\}$, donde T es el conjunto de puntos en el tiempo en el cual el proceso está definido. Se define un conjunto infinito de series de tiempo como *ensamble* y cada miembro de este ensamble es una posible *realización* del proceso estocástico y la serie de tiempo observada es una realización. Una manera de describir un proceso estocástico es especificar la distribución de probabilidad conjunta de $X_{t_1}, ..., X_{t_n}$ para cualquier conjunto de tiempos de ocurrencia $t_1, ..., t_n$ para cada valor de n. Definimos los momentos del proceso:

- **Media** La media $\mu(t)$ se define como:

$$\mu(t) = E(X_t)$$

- **Varianza** La varianza $\sigma^2(t)$ se define como:

$$\sigma^2(t) = Var(X_t)$$

- **Autocovarianza** La autocovarianza $\gamma(t_1, t_2)$ se define como:

$$\gamma(t_1, t_2) = Cov(X_{t1}, X_{t2}) = E\{[X_{t_1} - \mu(t_1)][X_{t_2} - \mu(t_2)]\}$$

Una serie de tiempo es *estrictamente estacionario* si la distribución conjunta de $X(t_1), ..., X(t_n)$ es la misma que la distribución conjunta de $X(t_1+\tau), ..., X(t_n+\tau)$ para toda $t_1, ..., t_n, \tau$; es decir, que cambiando el tiempo de origen por un monto τ no tiene efecto en la distribución conjunta que solo depende de los intervalos entre $t_1, ..., t_n$, esto para cualquier valor de n.
El tamaño de los coeficientes de la autocovarianza dependen de las unidades en las cuales $X(t)$ esté medida, así tenemos la *función de autocorrelación* que mide la autocorrelación entre $X(t)$ y $X(t+\tau)$:

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

Las propiedades de la función de autocorrelación son:

- **Propiedad 1** La función de autocorrelación es una función simétrica en el incremento:
$$\rho(\tau) = \rho(-\tau)$$

- **Propiedad 2** $|\rho(\tau)| \leq 1$

- **Propiedad 3** Para una función de autocovarianza particular existe solo un posible proceso estacionario normal y este está totalmente determinado por su media, varianza y función de autocorrelación y sin embargo, pueden existir varios procesos no-normales que compartan función de autocorrelación.

Existen *distribuciones de equilibrio* conforme $t \to \infty$ en donde la distribución de probabilidad de $X(t)$ tiende a un límite que no depende de las condiciones iniciales, así si se especifica que las condiciones iniciales son idénticas a aquellas especificadas en la distribución de equilibrio el proceso es estacionario en el tiempo y esta es la distribución estacionaria del proceso.
Un proceso es conocido como *estacionario en segundo orden* o debilmente estacionario si la media es constante y la función de covarianza depende solamente del incremento (τ) y ambas, la media y la varianza, son constantes y finitas. Es decir,

$$E[X(t)] = \mu \quad \text{y} \quad Cov[X(t), X(t+\tau)] = \gamma(\tau)$$

Hay varios procesos estocásticos útiles como:

- **Proceso aleatorio puro** Un proceso discreto $\{Z_t\}$ es considerado un proceso aleatorio puro si sus variables son una secuencia de variables aleatorias independientes e indénticamente distribuidad, este proceso es llamado también *ruido blanco.*

- **Caminata aleatoria** Supongamos que $\{Z_t\}$ es un proceso aleatorio puro discreto con media μ y varianza σ_Z^2, entonces un proceso $\{X_t\}$ es de caminata aleatoria si
$$X_t = X_{t-1} + Z_t$$
Y mientras la media y la varianza cambien con t el proceso es no estacionario.

- **Proceso de medias móviles** Supongamos que $\{Z_t\}$ es un proceso aleatorio, puro con media cero y varianza σ_Z^2, entonces el proceso $\{X_t\}$ es un proceso de medias móviles de orden m si
$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + ... + \beta_m Z_{t-m}$$
donde $\{\beta_i\}$ son constantes y las Z's están escalas tal que $\beta_0 = 1$.

- **Proceso Autorregresivo** Supongamos que $\{Z_t\}$ es un proceso aleatorio, puro con media cero y varianza σ_Z^2, entonces el proceso $\{X_t\}$ es un proceso autorregresivo de orden m si

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_m X_{t-m} + Z_t$$

Es como un modelo de regresión múltiple solo que la regresión de X_t no se hace sobre variables independientes sino sobre valores pasados de X_t. En su forma más simple vemos que el proceso autorregresivo puede escribirse también como un modelos de medias móviles.

- **Procesos Mezclados** Un proceso mezclado autorregresivo con medias móviles (proceso ARMA por sus siglas en inglés) contiene p términos del modelo autorregresivo y q términos del modelo de medias móviles y está definido como

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + ... + \beta_q Z_{t-q}$$

La importancia de este modelo es que las series estacionarias pueden ser descritas mediante un modelo ARMA involucrando menos parámetros que en un proceso autorregresivo o de medias móviles por si mismos.

- **Modelos integrados** En la práctica, la mayoría de las series de tiempo no son estacionarias, por lo que para poderlas estudiar con los métodos descritos debemos eliminar las fuentes de variación no-estacionaria. Si X_t es reemplazado por $\Delta^d X_t$ en la definición de los modelos mezclados tendríamos un modelo capaz de describir una serie no-estacionaria. Este modelo es conocido como integrado porque es un modelo estacionario al que se la ajustan datos no estacionarios, llamando el modelo proceso integrado autorregresivo con medias móviles (ARIMA por sus siglas en inglés. Definimos

$$W_t = \nabla^d X_t$$

Entonces, el proceso ARIMA

$$W_t = \alpha_1 W_{t-1} + ... + \alpha_p W_{t-p} + Z_t + ... + \beta_q Z_{t-q}$$

- **Proceso lineal generalizado** Un proceso de medias móviles de orden infinito y con media distinta de cero es definido como

$$X_t - \mu = \sum_{i=0}^{\infty} \beta_i Z_{t-i}$$

se llama proceso lineal generalizado porque se obtiene al pasar un proceso aleatorio puro por un sistema linear, así tanto el proceso de medias móviles como el autorregresivo son casos particulares de este proceso.

0.4 Constructing First Order Stationary Autoregressive Models via Latent Processes, Chatfield

Tenemos un modelo de autorregresión de la siguiente manera:

$$(Y_t - \mu) = \rho(Y_{t-1} - \mu) + \epsilon_t$$

Se trata de demostrar que se puede definir una función de densidad de la transición $f(y_t|y_{t-1})$ que vamos a expresar temporalmente como $f_{Y|Z}(y|z)$ donde $Y = Y_t$ y $Z = Y_{t-1}$:

$$f_Y(y) = \int f_{Y|Z}(y|z) f_Y(z) \mathrm{d}z$$

Por lo tanto:

$$\mathbb{E}(y|z) = \int y f_{Y|Z}(y|z) \mathrm{d}y = \mathbb{E}(y_t|y_{t-1})$$

Pues para lograr un modelo estacionario de Markov necesitamos exigir una relación lineal con respecto a la media, lo cual nos permite estacionareidad en la distribución marginal de Y_t y en la distribución de transición $(Y_t|Y_{t-1})$.
Ahora bien, para lograr esta relación lineal necesitamos introducir una tercera variable llamada latente o parámetro no observable ligándolo con el apartado 1, en este caso llamado x. Entonces tendríamos la densidad de transición como:

$$f_{Y|Z}(y|z) = \int f_1(y|x) f_2(x|z) \mathrm{d}\lambda(x)$$

En este caso, la $\lambda(x)$ es una medida de probabilidad (Lebesgue). El punto crucial es construir una distribución conjunta $f_{Y,X}(y, x)$ con una marginal $f_Y(y)$ tal que $f_1(y|x) = f_{Y|X}(y|x)$ y $f_2(x|z) = f_{X|Y}(x|z)$ como densidades condicionales y $f_Y(y) = \int f_{Y,X}(y, x) \mathrm{d}\lambda(x)$, así llegamos a $f_Y(y) = \int f_{Y|Z}(y|z) f_Y(z) \mathrm{d}z$.
Nótese que que la densidad de transición se puede realizar con un proceso latente (o de variables no observables) X_t tal que $Y_{t+1}|X_t \sim f_{Y|X}(\cdot|X_t)$ y $X_t|Y_t \sim f_{X|Y}(\cdot|Y_t)$.
Para lograr la linealidad del modelo debemos especificar $f_{X|Y}(x|y)$ para asegurar que $f_{Y|X}(y|x) \propto f_{X|Y}(x|y) f_Y(y)$ pertenezcan a la familia de $f_Y(y)$.
Para estimar futuras observaciones es relativamente sencillo construir la función de máxima verosimilitud, así tomamos $\{X_t\}$ como datos latentes o no observables y tenemos la función de la siguiente forma:

$$L = \prod_{t \geq 0} f_{Y|X}(y_t|x_{t-1}) f_{X|Y}(x_{t-1}|y_{t-1})$$

Dado que la estimación puede no ser tan sencillo como la construcción de la función, utilizaremos el algoritmo EM.

0.5 Maximum Likelihood from Incomplete Data via the EM Algorithm, Dempster.

En el paper se trata de explicar el estimador de máxima verosimilitud a través del algoritmo EM y dado que cada iteración del algoritmo consiste en un paso de esperanza seguido de uno de maximización, de ahí el nombre EM; este algoritmo sirve mucho cuando los datos están incompletos.
Cuando hablamos de "datos incompletos" implica la existencia de dos espacios muestrales $\mathcal{Y}$ y $\mathcal{X}$ y muchos mapeos que van solo de $\mathcal{X}$ a $\mathcal{Y}$. De nuevo vemos las variables aleatorias latentes o parámetros no observables, pues y es una r ealización de $\mathcal{Y}$; sin embargo, la realización de x en $\mathcal{X}$ se observa indirectamente a través de y, es decir, observamos el mapeo $x \to y(x)$, donde y son los datos observables.
Se postula una familia de densidades muestrales $f(x|\phi)$ que dependen del parámetro ϕ y se deriva en la correspondiente familia de densidades muestrales $g(y|\phi)$, es decir, las especificaciones de datos completos $f(\cdot|\ldots)$ se relaciona con los datos incompletos $g(\cdot|\ldots)$ mediante:

$$g(y|\phi) = \int_{\mathcal{X}(y)} f(x|\phi)\mathrm{d}x \tag{2}$$

El algoritmo EM se dedica a encontrar un valor de ϕ que maximice $g(y|\phi)$ dado una y observada, pero para lograr esto es esencial el uso de la familia asociada de $f(x|\phi)$.
Suponiendo que $f(x|\phi)$ tiene una forma usual de familia exponencial

$$f(x|\phi) = \frac{b(x)\exp(\phi t(x)^T)}{a(\phi)} \tag{3}$$

donde ϕ es un vector de parámetros de $1xr$, $t(x)$ es un vector de $1xr$ de estadísticos suficientes de datos completos, el parámetro ϕ es una transformación lineal no singular de dimensiones rxr arbitraria a $t(x)$. Una de las caracterizaciones más simples del algoritmo EM supone que $\phi^{(p)}$ denota el valor actual de ϕ después de p iteraciones, entonces la próxima iteración se puede desglosar en los siguentes dos pasos:

- El paso E: Estimar los estadísticos suficientes de los datos completos

$$t^{(p)} = \mathbb{E}[t(x)|y, \phi^{(p)}] \tag{4}$$

- El paso M: Determinar $\phi^{(p+1)}$ como la solución de la siguiente ecuación

$$\mathbb{E}[t(x)|\phi] = t^{(p)} \tag{5}$$

Si suponemos que $t^{(p)}$ es un estadístico suficiente calculado de x observada de la distribución (3), entonces la ecuación (5) se define como el estimador de máxima verosimilitud de ϕ. Para maximizar cualquier x basados en la distribución (3), podríamos maximizar

$$\log f(x|\phi) = -\log a(\phi) + \log b(x) + \phi t(x)^T \equiv -\log a(\phi) + \phi t(x)^T$$

Es decir, que la verosimilitud de x solo depende de la misma a través de $t(x)$, por lo que la ecuación (5) utilizaría como condición de maximización $-\log a(\phi) + \phi t(x)^{(p)T}$ aunque la $t^{(p)}$ calculada en la ecuación (4) represente o no un valor de $t(x)$ asociado a cualquier x en $\mathcal{X}$.
Siguiendo los pasos E y M del algoritmo, al final tendremos el valor de ϕ^* tal que maximice ϕ en

$$L(\phi) = \log g(y|\phi)$$

Para hacer más claros, primero se toma el concepto de la densidad condicional de $x|y, \phi$

$$k(x|y, \phi) = \frac{f(x|\phi)}{g(y|\phi)}$$

Así, podemos reescribir

$$L(\phi) = \log f(x|\phi) - \log k(x|y, \phi)$$

Como estamos trabajando con familias exponenciales, escribimos la función k como

$$k(x|y, \phi) = \frac{b(x)\exp(\phi t(x)^T)}{a(\phi|y)}$$

donde

$$a(\phi|y) = \int_{\mathcal{X}(y)} b(x)\exp(\phi t(x)^T)\mathrm{d}x$$

La diferencia de la función k con la f es que el denominador de la segunda está condicionado por y, por lo tanto ambas distribuciones pertenecen a la familia exponencial y tienen el mismo parámetro ϕ y el mismo estadístico suficiente $t(x)$ pero tienen soportes distintos, una en el espacio muestral $\mathcal{X}$ y el otro en $\mathcal{X}(y)$. Es por esto que podemos escribir la verosimilitud como

$$L(\phi) = -\log a(\phi) + \log a(\phi|y)$$

Derivando,

$$\mathbb{D}\log a(\phi) = \frac{\partial}{\partial\phi}\log a(\phi) = \mathbb{E}[t(x)|\phi]$$

$$\mathbb{D} \log a(\phi|y) = \frac{\partial}{\partial \phi} \log a(\phi|y) = \mathbb{E}[t(x)|\phi, y]$$

Por lo tanto, la verosimilitud quedaría

$$\mathbb{D}L(\phi) = -\mathbb{E}[t(x)|\phi] + \mathbb{E}[t(x)|\phi, y]$$

Esta última formula es crucial para el desarrollo del algoritmo EM, pues si el límite $\phi^* = \phi^{(p)} = \phi^{(p+1)}$ y combinando los pasos E y M, tendríamos que $\mathbb{E}[t(x)|\phi^*] = \mathbb{E}[t(x)|y, \phi^*]$.
Ahora supondremos que los datos completos se distribuyen como una familia exponencial curva. En este caso, la representación (3) sigue siendo útil pero ahora el parámetro ϕ está en una subvariedad curveada Ω_0 de la región convexa y con r dimensiones de Ω, el paso E puede ser definido como en la ecuación (4) pero debemos sustituir el paso M tal que se determine una $\phi^{(p+1)}$ como un valor de ϕ en Ω_0 que maximice $-\log a(\phi) + \phi t^{(p)T}$. Es decir, el paso de maximización se trata de maximizar una verosimilitud suponiendo que x produce estadísticos suficientes $t^{(p)}$.
En un nivel aún más general se omite cualquier referencia a familias exponenciales, se toma la función

$$Q(\phi'|\phi) = \mathbb{E}[\log f(x|\phi')|y, \phi] \tag{6}$$

que se asume que existe para toda pareja (ϕ', ϕ). Definimos la iteración EM como:

- El paso E: Calcular $Q(\phi, \phi^{(p)})$.
- El paso M: Determinar un $\phi^{(p+1)} \in \Omega$ tal que maximice $Q((\phi, \phi^{(p)})$.

La idea central es que se tiene que tomar una ϕ^* que maximice $\log f(x|\phi)$, dado que esto no necesariamente se conoce se puede maximizar la esperanza de los datos observados y y la $\phi^{(p)}$.
Análogamente se define también

$$H(\phi'|\phi) = \mathbb{E}[\log k(x|y, \phi')|y, \phi] \tag{7}$$

Con los resultados anteriores podemos concluir

$$Q(\phi'|\phi) = L(\phi') + H(\phi'|\phi) \tag{8}$$

Lema 1. Para cualquier par $(\phi'.\phi)$ en $\Omega x \Omega$

$$H(\phi'|\phi) \leq H(\phi)\phi$$

si y solo si $k(x|y, \phi') = k(x|y, \phi)$.

Definiremos un "algoritmo iterativo" como un mapeo de $\phi \to M(\phi)$ y de $\Omega \to \Omega$ tal que en cada paso $\phi^{(p)} \to \phi^{(p+1)}$, se define como

$$\phi^{(p+1)} = M(\phi^{(p)})$$

Definición 1 Un algoritmo iterativo con mapeo $M(\phi)$ es un algoritmo EM generalizado (GEM) si

$$Q(M(\phi)|\phi) \geq Q(\phi'|\phi)$$

es decir, que el valor que maximiza $Q(\phi'|\phi)$ es $\phi' = M(\phi)$.

Teorema 1 Para cada algoritmo GEM

$$L(M(\phi)) \geq L(\phi) \qquad \forall \phi \in \Omega$$

si y solo si

$$Q(M(\phi)|\phi) = Q(\phi|\phi)$$
$$k(x|y, M(\phi)) = k(x|y, \phi)$$

Corolario 1 Supongamos que $\phi^* \in \Omega$, $L(\phi^*) \geq L(\phi) \qquad \forall \phi \in \Omega$. Entonces para cada algoritmo GEM:

- $L(M(\phi^*)) = L(\phi^*)$.
- $Q(M(\phi^*)|\phi^*) = Q(\phi^*|\phi^*)$.
- $k(x|y, M(\phi^*)) = k(x|y, \phi^*)$.

Corolario 2 Si para $\phi^* \in \Omega, L(\phi^*) > L(\phi) \qquad \phi \in \Omega$ tal que $\phi = \phi^*$ entonces, para cada algoritmo GEM

$$M(\phi^*) = \phi^*$$

Teorema 2 Supongamos que $\phi^{(p)}$ para $p = 0, 1, ...$ es un argumento del algoritmo GEM tal que

- La secuencia $L(\phi^{(p)})$ está acotada.
- $Q(\phi^{(p+1)}|\phi^{(p)}) - Q(\phi^{(p)}|\phi^{(p)}) \geq \lambda(\phi^{(p+1)} - \phi^{(p)})(\phi^{(p+1)} - \phi^{(p)})^T$ para cualquier escalar $\lambda > 0, p$.

Entonces la secuencia $\phi^{(p)}$ converge a ϕ^* en Ω.

Estos corolarios indican que la máxima verosimilitud estimada es un punto fijo en el algoritmo GEM. Asumimos que Ω es una región en un espacio real y existe continuidad, así se asegura la convergencia de los estimadores de máxima verosimilitud.