Caso de Estudio 7
IELE-4010 Procesos Estocásticos 2021-II
Profesor: Luis Felipe Giraldo

Experimento MLE: datos reales cambio climático

1. Se tienen que resolver todos los enunciados planteados en este ejercicio para obtener puntos en el mismo. La falta de intento de solución de alguno de los enunciados de este problema implicará automáticamente una calificación de 0 puntos en todo el ejercicio. Las rutinas se tienen que implementar o en Python o en Matlab. Adjunte en un archivo aparte el código fuente

El cambio climático ha sido debatido por años, y una pieza central del debate ha sido su causa: se debe al ciclo natural del sistema, o se debe a la actividad humana? Usted puede encontrar adjunto el reporte sobre cambio climático publicado en 2007 por si quiere entender mejor el contexto de este problema.

Adjunto puede encontrar los siguientes datos reales:

- `SunRad.txt`: **Radiación solar mensual** (kWh/m^2/dia) en la ciudad de Columbus, Ohio, EU (periodo: 1961- 1990)
- `TempCO2.txt`:
  Anomalías de temperatura anual (°C). Estas anomalías se calculan como la diferencia entre la temperatura observada y la temperatura estimada por un modelo (periodo: 1961-1990).
  Concentración anual de CO2 (ppm) medido en Mauna Loa Observatory, Hawaii (periodo: 1961-1990).

a) (2 puntos) Convierta los datos de radiación solar a un promedio anual.

b) (5 puntos) Grafique cada una de las tres señales contra el tiempo. Analice las gráficas. No se le olvide colocar las unidades de cada eje.

c) (5 puntos) Grafique temperatura vs radiación solar, y grafique temperatura vs concentración de CO2.

d) (8 puntos) Utilizando la estimación ML de la matriz de covarianza, estime el coeficiente de correlación entre las parejas de variables: (temperatura, radiación solar), y (temperatura, concentración de CO2). Analice estos resultados y los obtenidos en a)- c). **No puede utilizar funciones predefinidas para hacer estas estimaciones.** Recuerde que un estimador de la matriz de covarianza está dado por la ecuación

$$\hat{Q} = \frac{1}{M-1} \sum_{i=1}^{M} (x_i - \mu)(x_i - \mu)^T,$$

donde $x, \mu \in \mathbb{R}^n$, y μ es la estimación del vector de valores esperados dada por la ecuación

$$\mu = \frac{1}{M} \sum_{i=1}^{M} x_i.$$

e) (25 puntos) Empezando por los primeros 10 años y terminando en los últimos 10 años, deslice una ventana de 10 años y estime el coeficiente de correlación de las parejas (temperatura, radiación solar), y (temperatura, concentración de CO2) para cada ventana. Como resultado usted debe obtener dos señales de correlación con 20 puntos cada una. Grafíquelas contra el tiempo.

f) (5 puntos) Basados en cada uno de los resultados obtenidos en a) - e), intente responder a la pregunta planteada inicialmente en este problema.

Experimento MLE: datos reales invernadero

2. (50 puntos) Se tienen que resolver todos los enunciados planteados en este ejercicio para obtener puntos en el mismo. La falta de intento de solución de alguno de los enunciados de este problema implicará automáticamente una calificación de 0 puntos en todo el ejercicio. Las rutinas se tienen que implementar o en Python o en Matlab. Adjunte en un archivo aparte el código fuente

   Las mediciones de temperatura T (unidades en $^{\circ}C$) y humedad relativa H (unidades en %) en un invernadero real de un cultivo de rosas [1] se modelan como variables aleatorias continuas a través del vector aleatorio $Z = [T, H]^T$. Nos indican que asumir que Z es un vector Gaussiano no es suficiente para modelar el comportamiento de la temperatura y la humedad conjuntamente. Nos piden estimar una distribución que se ajuste mejor al verdadero comportamiento del invernadero. Para esto, nos dan en el archivo `greenhouseCaso7.txt` 250 observaciones conjuntas de las variables de temperatura T (primera columna) y de humedad H (segunda columna).

   a) Grafique las observaciones de temperatura vs humedad, y realice una inspección visual de los datos analizando su comportamiento.

   b) Estime la mejor distribución Gaussiana de Z (una sola Gaussiana) utilizando MLE, e indique la función resultante. Grafique las observaciones de temperatura vs humedad, y traslape el contorno de la función de densidad. Comente los resultados obtenidos. Recuerde que para estimar la distribución Gaussiana basta con estimar el vector de valores esperados y la matriz de covarianza. **No puede utilizar funciones predefinidas para estimar el vector de medias y la matriz de covarianza. Implemente su propia rutina para hacer este cálculo.**

   c) Implemente el algoritmo EM, y estime la función de densidad Z como una combinación de dos funciones de densidad Gaussianas. Es decir, $f_Z(z) = \sum_{k=1}^{2} \alpha_k N(z; \mu_k, Q_k)$. Grafique de nuevo las observaciones, y traslape el contorno de la función de densidad combinada $f_Z(z)$ (NO el contorno de cada función de densidad base por separado) para varias de las iteraciones del algoritmo. Comente los resultados obtenidos. **No puede utilizar funciones predefinidas para obtener la combinación de Gaussianas. Implemente su propia rutina del algoritmo EM.**

   d) Repita el punto c) ahora con 3 funciones Gaussianas. Indique cuál de los tres resultados (i.e., 1, 2, o 3 Gaussianas) usted consideraría genera un mejor ajuste de los datos.

   e) Se tienen 50 nuevos datos de temperatura en el archivo `greenhouse50temp.txt` para los cuales se desconoce el valor asociado de humedad.

      (i) Con el modelo obtenido en el enunciado b), obtenga un estimador MMSE sin restricciones de H basado en las observaciones de T. En este caso, como se tiene sólo una Gaussiana conjunta, este estimador sin restricciones en es mismo lineal. Muestre claramente la ecuación del estimador y estime las observaciones de H asociadas a las 50 observaciones recolectados de temperatura. Grafique T vs $\hat{H}$, donde $\hat{H}(t)$ se refiere a la estimación de la humedad basado en la observación de temperatura t.

[1] Tomado de IEEE Data Port

(ii) Sea $f_Z(t,h) = \sum_{k=1}^{3} \alpha_k N(t,h;\mu_k,Q_k)$ la distribución encontrada en el enunciado d). Escriba una rutina que, dada una observación de temperatura t, calcule $\hat{H}(t) = E[H|T = t] = \int_{-\infty}^{\infty} h f_H(h|T = t)dh$. Utilizando esta rutina, encuentre la estimación de la humedad basado en las 50 observaciones de temperatura. Grafique T vs $\hat{H}$, donde $\hat{H}(t)$ se refiere a la estimación de la humedad basado en la observación de temperatura t.
Recuerde que $f_H(h|T = t) = f_Z(t,h)/f_T(t)$, y que $f_T(t) = \int_{-\infty}^{\infty} f_Z(t,h)dh$.

(iii) Eventualmente se pudo recuperar las observaciones reales de humedad asociadas a las 50 observaciones de temperatura. Éstas se encuentran en el archivo `greenhouse50hum.txt`. Se quiere determinar cuál de los dos estimadores de humedad fue el más preciso: el encontrado en el enunciado (i) o el encontrado en el enunciado (ii). Para eso, en ambos casos, calcule el error: $error = \frac{1}{50}\sum_{i=1}^{50}(h_i - \hat{H}(t_i))^2$, donde h_i y t_i se refieren a los valores reales de humedad y temperatura, y $\hat{H}(t_i)$ se refiere a la estimación de h_i. Basado en este error, determine cuál de los estimadores sería el más adecuado de utilizar con respecto a precisión y tiempo de cómputo.