Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

Definición e intuición

- Un **outlier** es una observación que parece no seguir el patrón general de los datos.
- Puede deberse a:
    - error de medición,
    - condición experimental distinta,
    - o un valor extremo válido.
- Pregunta central: *¿Eliminamos o incorporamos el posible outlier en el análisis?*

Modelo lineal y detección clásica

Modelo lineal normal:

$$\mathbf{y} = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n).$$

- **Hat matrix:**

$$H = X(X'X)^{-1}X', \qquad \hat{\mathbf{y}} = H\mathbf{y}.$$

- Sus diagonales h_{ii} miden el **apalancamiento** de la observación i:

$$0 \leq h_{ii} \leq 1, \quad \sum_i h_{ii} = p.$$

- **Residuales:** $\mathbf{e} = (I - H)\mathbf{y}$, con $(e_i) = \sigma^2(1 - h_{ii})$.
- **Residuales estandarizados:**

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

- Reglas prácticas:
  - $|r_i| > 2$ o 3 $\Rightarrow$ posible atipicidad.
  - $h_{ii} > 2p/n$ $\Rightarrow$ posible apalancamiento inusual.

Idea bayesiana

- En lugar de una decisión *binaria* (outlier / no outlier), modelamos la **posible atipicidad** dentro del marco probabilístico.
- Supongamos una mezcla:

$$y_i \sim (1-\alpha)\mathcal{N}(\mu, \sigma^2) + \alpha\mathcal{N}(\mu, k^2\sigma^2), \quad k > 1.$$

- Cada observación tiene una variable latente z_i:

$$z_i = \begin{cases} 0 & \text{si proviene de la componente "regular"} \\ 1 & \text{si proviene de la componente "outlier"} \end{cases}$$

- La inferencia bayesiana nos da la **probabilidad posterior**:

$$\Pr(z_i = 1 \mid y_i, \text{datos}) \quad \Rightarrow \quad \text{medida continua de atipicidad.}$$

- No se eliminan observaciones: se *repondera* su contribución según la evidencia que aportan.

Modelo simple de mezcla normal

Especificación (West, 1984)

Para $y_1, \ldots, y_n \in \mathbb{R}$:

$$y_i \sim (1-\alpha)\mathcal{N}(\mu, \sigma^2) + \alpha\mathcal{N}(\mu, \tau^2), \qquad 0 < \alpha < 1, \ \tau^2 \gg \sigma^2.$$

- **Componente central:** distribución normal con varianza pequeña, $\mathcal{N}(\mu, \sigma^2)$.
- **Componente de outliers:** distribución normal con la misma media pero varianza inflada, $\mathcal{N}(\mu, \tau^2)$.
- Restricción: $\tau^2 > \sigma^2$ para garantizar la interpretación.
- La mezcla permite que cada observación pueda ser generada por cualquiera de las dos componentes.

Verosimilitud

Densidad marginal de cada observación:

$$f(y_i \mid \mu, \sigma^2, \tau^2, \alpha) = (1-\alpha)\,\phi(y_i \mid \mu, \sigma^2) + \alpha\,\phi(y_i \mid \mu, \tau^2),$$

Verosimilitud (datos observados):

$$L(\mu, \sigma^2, \tau^2, \alpha \mid \mathbf{y}) = \prod_{i=1}^{n} \Big[(1-\alpha)\,\phi(y_i \mid \mu, \sigma^2) + \alpha\,\phi(y_i \mid \mu, \tau^2)\Big].$$

Reformulación con variables latentes z_i:

$$p(y_i, z_i \mid \cdot) = \big[(1-\alpha)\,\phi(y_i \mid \mu, \sigma^2)\big]^{1-z_i} \cdot \big[\alpha\,\phi(y_i \mid \mu, \tau^2)\big]^{z_i},$$

$$\Rightarrow \quad p(\mathbf{y}, \mathbf{z} \mid \cdot) = \prod_{i=1}^{n} p(y_i, z_i \mid \cdot).$$

Distribuciones a priori

Para completar el modelo bayesiano, especificamos distribuciones a priori sobre los parámetros:

$$\mu \sim \mathcal{N}(m_0, s_0^2), \qquad \sigma^2 \sim \text{Inv-Gamma}(a_1, b_1),$$
$$\tau^2 \sim \text{Inv-Gamma}(a_2, b_2), \qquad \alpha \sim \text{Beta}(c, d).$$

- **Media:** prior normal para μ, centrado en m_0 con varianza s_0^2.
- **Varianzas:** priors Inversa-Gamma para σ^2 (componente central) y τ^2 (componente outlier). Se impone la restricción $\tau^2 > \sigma^2$ para mantener la identificación.
- **Proporción de outliers:** prior Beta para α, flexible en $[0, 1]$.

Modelo bayesiano completo:

$$p(\mu, \sigma^2, \tau^2, \alpha, \mathbf{z} \mid \mathbf{y}) \;\propto\; L(\mu, \sigma^2, \tau^2, \alpha \mid \mathbf{y}, \mathbf{z})\, p(\mu)\, p(\sigma^2)\, p(\tau^2)\, p(\alpha).$$

Probabilidad posterior de outlier

Definamos variables latentes $z_i \in \{0, 1\}$:

$$z_i = \begin{cases} 0 & \text{si } y_i \text{ proviene de la componente central,} \\ 1 & \text{si } y_i \text{ proviene de la componente outlier.} \end{cases}$$

La probabilidad posterior de que la observación i sea un outlier es:

$$\Pr(z_i = 1 \mid y_i, \mu, \sigma^2, \tau^2, \alpha) = \frac{\alpha\,\phi(y_i \mid \mu, \tau^2)}{(1-\alpha)\,\phi(y_i \mid \mu, \sigma^2) + \alpha\,\phi(y_i \mid \mu, \tau^2)}.$$

- Es una **medida continua de atipicidad**: valores cercanos a 1 sugieren que y_i es un outlier; valores cercanos a 0 sugieren que es regular.
- Este enfoque evita clasificaciones binarias rígidas.

Discusión

- El modelo de mezcla $\mathcal{N}(\mu, \sigma^2)$ / $\mathcal{N}(\mu, \tau^2)$ introduce la **incertidumbre** en la detección de atípicos.
- Cada observación recibe una **probabilidad posterior** de ser outlier, en lugar de una clasificación rígida.
- No se descarta información: los datos atípicos siguen influyendo, pero con *peso reducido*.
- Este enfoque motiva **extensiones más realistas**:
  - Regresión bayesiana robusta (mezclas condicionadas a X).
  - Modelos con errores Student-t.
  - Modelos dinámicos para series temporales.
- En particular, el siguiente paso es pasar de una mezcla en y a una **mezcla de regresiones**, donde distintas relaciones lineales explican subpoblaciones en los datos.

Modelo de mezcla de dos regresiones

$$y_i \ \sim \ \pi\,\mathcal{N}(x_i^\top \beta_1, \sigma_1^2) + (1-\pi)\,\mathcal{N}(x_i^\top \beta_2, \sigma_2^2).$$

- Cada observación proviene de una de dos **relaciones lineales** posibles.
- Variables latentes $z_i \in \{1, 2\}$ indican de qué regresión proviene cada y_i.
- Parámetros del modelo:
  - β_1, σ_1^2: coeficientes y varianza de la primera regresión.
  - β_2, σ_2^2: coeficientes y varianza de la segunda regresión.
  - π: peso de mezcla (proporción esperada de la primera componente).
- Interpretación:
  - Una recta explica la mayoría de los datos.
  - La otra recta puede capturar **subpoblaciones** o comportarse como mecanismo para identificar posibles **outliers estructurados**.

Identificabilidad y re-etiquetado

- En modelos de mezcla, la verosimilitud es **invariante** a la permutación de etiquetas de las componentes.
- Consecuencia: durante el muestreo MCMC puede ocurrir el **label switching**, es decir, que las cadenas intercambien las etiquetas de las rectas.
- Esto no afecta a la bondad del ajuste, pero sí dificulta la **interpretación** de parámetros y probabilidades posteriores.

Radiocarbono y outliers

- En paleoecología y arqueología, las cronologías se construyen con fechamientos de radiocarbono.
- Problema común: algunas dataciones son **atípicas** debido a
  - contaminación del material,
  - problemas de laboratorio,
  - mezcla de sedimentos (retrabajo, bioturbación).
- Estas observaciones pueden sesgar fuertemente la cronología si se usan sin corrección.

Tratamiento clásico

- Enfoque tradicional: **excluir manualmente** fechas sospechosas.
- Limitaciones:
  - Decisión subjetiva y difícil de justificar.
  - Puede eliminar información útil.
  - No cuantifica la **incertidumbre** en la clasificación.
- Ejemplo: en cronologías de lagos o turberas, se eliminan fechas "demasiado jóvenes" o "demasiado antiguas" respecto a la estratigrafía.

Enfoques bayesianos

- Los modelos de mezcla permiten representar explícitamente la **posibilidad de outliers**.
- Cada fechamiento recibe una **probabilidad posterior** de ser atípico.
- Aplicaciones concretas:
  - OxCal: modelo de mezcla para detectar outliers en series de radiocarbono.
  - Bacon / Bchron: permiten incluir *priors* sobre outliers.
  - Plum / PyPlum: integración bayesiana de ^{210}Pb y radiocarbono con mecanismos de detección de fechas problemáticas.
- Ventaja: en lugar de borrar fechas, se *reponderan* de acuerdo a su consistencia con el modelo cronológico.

Ejemplo real: Outliers en radiocarbono (Bronk Ramsey, 2009)

- Problema: en cronologías de radiocarbono siempre aparecen **fechas atípicas**, por contaminación, problemas de contexto o sesgos sistemáticos.
- Estrategia tradicional: exclusión manual de fechas sospechosas.
- Propuesta: modelos bayesianos de mezcla que asignan a cada datación una **probabilidad de ser outlier**.
- Implementación en **OxCal v4**:
  - Priors para la probabilidad de outlier (ej. 5%).
  - Fechas problemáticas se *downweightean* en vez de borrarse.
  - Soporta distintos tipos: errores de medida (s-type), offsets de reservorio (r-type), desplazamientos temporales (t-type).

Ejemplo real: Outliers en radiocarbono (Bronk Ramsey, 2009)

- Cada datación y_i se modela como una mezcla:

$$f(y_i) \; = \; (1 - p_i)\, \phi(y_i \mid \theta_i, \sigma_i^2) \; + \; p_i\, g(y_i),$$

  donde
  - $\phi(\cdot)$: distribución normal asociada a la edad verdadera θ_i y su error σ_i.
  - $g(\cdot)$: distribución alternativa de outliers (ej. uniforme sobre un rango amplio).
  - p_i: probabilidad a priori de que la fecha sea outlier (ej. 5%).
- De esta manera, las fechas inconsistentes no se eliminan, sino que su peso se *reduce* automáticamente en el modelo.
- Implementado en **OxCal**, ampliamente usado en cronologías arqueológicas y paleoambientales.

Ejemplo alternativo: Robustez vía Student-t (Christen & Pérez, 2009)

- Problema: la varianza reportada σ_j^2 de cada datación 14C suele estar subestimada y no refleja la dispersión real observada en interlaboratorios.
- Propuesta: introducir un **multiplicador de varianza** desconocido $\alpha > 0$:

$$y_j \sim \mathcal{N}\big(\mu(\theta), \alpha(\sigma_j^2 + \sigma^2(\theta))\big).$$

- Priorizando $\alpha \sim \text{Inv-Gamma}(a, b)$, el modelo marginal resulta en una **distribución Student-t** para y_j.
- Ventaja: el modelo hereda colas más pesadas que la Normal y, por lo tanto, es *naturalmente robusto* a valores atípicos.
- Aplicaciones:
  - Fechados simulados y reales (incluyendo la Sábana Santa de Turín).
  - Intervalos HPD más amplios y estables en presencia de outliers.
  - Sin necesidad de eliminar fechas: se ajusta automáticamente el grado de incertidumbre.

Discusión

- La detección bayesiana de outliers es hoy una herramienta estándar en la construcción de cronologías.
- Se alinea con la idea central: **incorporar la incertidumbre** en lugar de ignorarla.
- Ejemplo real: en núcleos de sedimentos marinos y lacustres, unos pocos fechamientos atípicos pueden desplazar décadas o siglos la cronología si no se modelan adecuadamente.
- Modelos de mezcla proporcionan un mecanismo formal y replicable para tratar estos casos.
- Existen distintos enfoques:
  - Mezclas explícitas con componentes "outlier" (Bronk Ramsey, 2009).
  - Modelos robustos basados en Student-t y varianzas infladas (Christen & Pérez, 2009).
- Estas ideas inspiran extensiones más generales, como las **mezclas de regresiones**, donde diferentes rectas explican subpoblaciones en los datos.

Referencias

- West, M. (1984). *Outlier models and prior distributions in Bayesian linear regression*. JRSS B, 46(3), 431–439.
- Box, G. E. P. & Tiao, G. C. (1968). *A Bayesian approach to some outlier problems*. Biometrika, 55(1), 119–129.
- West, M. & Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.
- Bronk Ramsey, C. (2009). *Dealing with outliers and offsets in radiocarbon dating*. Radiocarbon, 51(3), 1023–1045.
- Christen, J. A. & Pérez, S. (2009). *A new robust statistical model for radiocarbon data*. Radiocarbon, 51(3), 1047–1059.