¿Porqué considerar la correlación en nuestro análisis?

Caso de ejemplo:

Consideremos el vector aleatorio

$Z = [Z(s_1), \ldots, Z(s_N)]$ y que $Z \sim N(\mu, \Sigma)$

con $\mu = \mu_0 \mathbb{1}$ y $\Sigma_{ij} = \sigma^2 \rho$ si $i \neq j$ y σ^2 e.o.c.

Supongamos que deseamos estimar μ.
Recordemos que un estimador de μ es $\bar{X}$. Luego, notemos que $\bar{X}$ es insesgado:

$$\mathbb{E}(\bar{Z}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} Z(s_i)\right) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(Z(s_i)) = \frac{\cancel{N}}{\cancel{N}}\mu_0 = \mu_0$$

¿Qué sucede con su varianza?

sea $\mathbb{1} = (1, \ldots, 1)^T$

$$V(\bar{Z}) = V\left(\frac{1}{N} Z \mathbb{1}\right) = \frac{1}{N^2}\left(\mathbb{1}^T V(Z) \mathbb{1}\right) = \frac{\sigma^2}{N^2} \mathbb{1}^T \Sigma \mathbb{1}$$

$$= \frac{\sigma^2}{N^2}\left(N + \cancel{2}\rho \frac{(N-1)N}{\cancel{2}}\right) = \frac{\sigma^2}{N}\left(1 + \rho(N-1)\right)$$

$$= \sigma^2\left(\frac{1}{N} + \rho - \frac{\rho}{N}\right) \xrightarrow[N \to \infty]{} \sigma^2 \rho$$

Conclusión: ρ **no** es consistente.

OTRA CONSECUENCIA (TEST DE HIPÓTESIS, VARIANZA CONOCIDA):

$$H_0: \mu_0 = \mu_T \quad V/S \quad H_1: \mu_0 \neq \mu_T$$

$$Z^{I}_{est} = \frac{(\mu_T - \bar{Z})\sqrt{N}}{\sigma}$$

$$Z^{D}_{est} = \frac{(\mu_T - \bar{Z})\sqrt{N}}{\sigma(1 + \rho(N-1))}$$

SI $0 < \rho < 1$, ENTONCES $Z^{D}_{est} > Z^{I}_{est}$

OTRA CONSECUENCIA (INFORMACIÓN KL)

$D_{KL}(P \| Q)$ = ENTROPÍA RELATIVA DE P RESPECTO A Q

$$D(N_0 \| N_1) = \frac{1}{2}\left(TR(\Sigma_1^{-1}\Sigma_0) - N + LN\left(\frac{DET|\Sigma_1|}{DET|\Sigma_0|}\right)\right)$$

$$DET(\Sigma_0) = \sigma^{2N}\left((1-\rho)^N + N\rho(1-\rho)^{N-1}\right)$$

$$DET(\Sigma_1) = \sigma^{2N}$$

$$\frac{DET(\Sigma_1)}{DET(\Sigma_0)} = (1-\rho)^N + K\rho(1-\rho)^{N-1}$$

$$TR(\Sigma_1^{-1}\Sigma_0) = N$$

LUEGO $D(N_0 \| N_1) = -LN\left((1-\rho)^N + N\rho(1-\rho)^{N-1}\right)$

DATOS GEOESTADÍSTICOS

UN DATO GEOESTADÍSTICO SE COMPONE POR N OBSERVACIONES INDEXADAS POR LAS COORDENADAS ESPACIALES $s_1, \dots, s_N \subset \mathcal{S}$. LA CARACTERÍSTICA RELEVANTE DE ESTE TIPO DE DATO ES QUE LA OBSERVACIÓN $z_i = Z(s_i)$ PUEDE SER RECOLECTADA EN CUALQUIER POSICIÓN DE $\mathcal{S}$. NOTAMOS QUE $s_1, \dots, s_N$ SON DATOS

LA FORMA CLÁSICA DE PRESENTAR ESTOS DATOS ES MEDIANTE LAS ESTADÍSTICAS CLÁSICAS: MEDIA, VARIANZA, HISTOGRAMAS, BOXPLOTS, ETC.

NO OBSTANTE, ESTOS NO CONTIENEN LA INFORMACIÓN ESPACIAL. PARA ESTO, LA MANERA MÁS RAZONABLE DE PRESENTAR LOS DATOS ESPACIALES ES MEDIANTE MAPAS. ESTO DEBIDO A QUE, TÍPICAMENTE, LIDIAMOS CON DATOS DEL ESTILO $(s_x, s_y, Z(s_x, s_y))$.

SI LAS LOCALIZACIONES DE LOS DATOS ESPACIALES ES TAL QUE NO SE PUEDE REPRODUCIR UN MAPA, SE SUELE PROCEDER VÍA SUAVIZAMIENTO; ES DECIR:

$$\hat{Z}(s) = \frac{\sum_{i=1}^{N} K\left(\frac{s-s_i}{h}\right) Z(s_i)}{\sum_{i=1}^{N} K\left(\frac{s-s_i}{h}\right)}$$

DONDE $h > 0$ ES CONOCIDO COMO EL ANCHO DE BANDA Y $K(\cdot)$ ES UNA FUNCIÓN (KERNEL) TAL QUE:

$$K(-u) = K(u) \; ; \quad \int_{\mathbb{R}^2} u K(u)\, du = 0$$

es: $$K(u) = \frac{1}{\pi} \mathbb{I}_{(\|u\| \le 1)}(u)$$

$$K(u) = \frac{1}{2\pi} e^{-\frac{1}{2}\|u\|^2}$$

$$K(u) = \left(\frac{3}{4}\right)^2 (1 - u_x^2)(1 - u_y^2) \cdot \mathbb{I}_{(|u_x| < 1)}(u_x) \, \mathbb{I}_{(|u_y| < 1)}(u_y)$$

Naturalmente, el reporte dependerá de la elección de h. Esto típicamente se hace mediante el asumir características de los datos como independencia y gaussianidad, obteniendo una regla empírica (o "rule of thumb"). Recordar que esto es una primera aproximación.

Según Li y Racine (2006), una propuesta inicial puede ser

$$h = \sigma_x \, n^{-1/5}.$$

Para decir más sobre los datos geoestadísticos es necesario incorporar más notación.

Recordemos que un dato geoestadístico es una realización de un campo aleatorio. Asumiremos las condiciones de consistencia de Kolmogorov para identificar el campo aleatorio con muestras finitas.

Diremos que un campo aleatorio es estrictamente estacionario si, para todo $h \in \mathbb{R}^d$ se tiene que:

$$\mathbb{P}(Z(s_1) \le z_1, \dots, Z(s_n) \le z_n) = \mathbb{P}(Z(s_1+h) \le z_1, \dots, Z(s_n+h) \le z_n)$$

Para describir la observación, nos enfocamos en los momentos de primer y segundo orden:

$$\mu(s) := \mathbb{E}(Z(s)) = \int_{\mathbb{R}} z f_s(z)\, dz$$

$$C(s,s') = \mathrm{Cov}(Z(s), Z(s')) = \mathbb{E}(Z(s)Z(s')) - \mathbb{E}(Z(s))\mathbb{E}(Z(s'))$$

Proceso Gaussiano.

Diremos que Z es un campo aleatorio Gaussiano si, para todo $N \in \mathbb{N}$, y localizaciones espaciales $s_1, \ldots, s_N$, se tiene que el vector

$$[Z(s_1), \ldots, Z(s_N)]^T \sim N(\vec{\mu}, \Sigma)$$

donde $\vec{\mu} = [\mu(s_1), \ldots, \mu(s_N)]^T$ y

$$\Sigma_{ij} = C(s_i, s_j).$$

Debido a la consistencia de Kolmogorov, $\mu(s)$ es una función medible, mientras que $C(s,s')$ debe ser una función definida positiva, esto es:

Para todo $N \in \mathbb{N}$, localizaciones espaciales $s_1, \ldots, s_N$, y para toda secuencia de números reales $\alpha_1, \ldots, \alpha_N$ se tiene que:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j C(s_i, s_j) > 0.$$

ESTA HIPÓTESIS ES DISTRIBUCIONAL Y NOS PERMITE DISCUTIR SOBRE ALGUNOS ASPECTOS PRINCIPALES DE LOS CAMPOS ALEATORIOS.

EJ: SI $s, s' \in \mathbb{R}_0^+$:

- $C(s,s') = \operatorname{Min}(s,s')$ ES COVARIANZA

- $C(s,s') = \frac{1}{2\beta} e^{-\beta \| s-s' \|} \qquad s,s' \in \mathbb{R}^d$

- $C(s,s') = \sigma^2 e^{-\beta \| s-s' \|^2} \qquad s,s' \in \mathbb{R}^d$

SIN EMBARGO, PARA DECIR ALGO DE LA ASOCIACIÓN ESPACIAL, ES NECESARIO INCORPORAR ESTRUCTURAS DE 2^{DO} ORDEN:

ESTACIONARIEDAD DE 2^{DO} ORDEN:

Z ES ESTACIONARIO DE 2^{DO} ORDEN SI $\mathbb{E}(Z(s)) = m$ Y SI EXISTE $R: \mathbb{R}^d \to \mathbb{R}$ Y $\sigma^2 > 0$ TAL QUE $C(s,s') = \sigma^2 R(s-s')$.

SE DIRÁ <u>ISOTRÓPICO</u> SI EXISTE $R_0 : \mathbb{R}_0^+ \to \mathbb{R}$ Y $\sigma^2 > 0$ TAL QUE $C(s,s') = \sigma^2 R_0(\| s-s' \|)$

UN PROCESO ESTACIONARIO DE 2DO ORDEN SATISFACE QUE:

1.- $C(s,s) = \sigma^2 R(0) \geq 0$; PARA TODO s

2.- $R(s) = R(-s)$; PARA TODO s

3.- $|R(s)| \leq R(0)$; PARA TODO s

DEMOSTRACIÓN:

1.- $C(s,s) = V(Z(s)) \geq 0$

2.- $$C(s,s') = COV(Z(s), Z(s')) = COV(Z(s'), Z(s)) = C(s', s)$$

3.- SIN PÉRDIDA DE GENERALIDAD, ASUMIMOS QUE $\sigma^2 = 1$. LUEGO:

$$|R(s-s')|^2 = |COV(Z(s), Z(s'))| \underset{\text{CAUCHY-SCHWARZ}}{\leq} V(Z(s))\,V(Z(s')) = (\sigma^2)^2$$

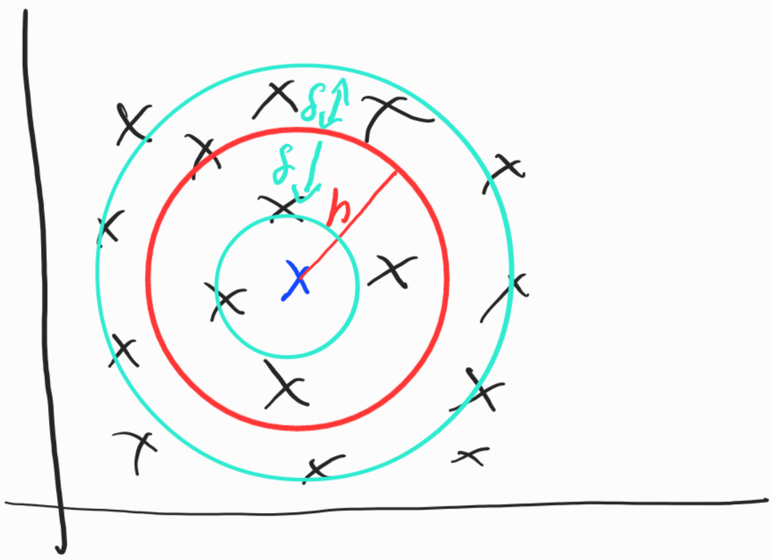
La estacionariedad de 2do orden nos permite generar descriptores de los datos mediante la función de covarianza. En efecto, se puede estimar mediante el método de los momentos:

$$\hat{C}(s, s+h) = \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} (Z(s_i)-\hat{m})(Z(s_j)-\hat{m})$$

Donde $N(h) = \{ (i,j) : \| s_i - s_j \| = h \}$

Generalmente, $N(h)$ es vacío, por lo que se suele reemplazar por un aproximado:

$$N_\delta(h) = \{ (i,j) : \big| \| s_i - s_j \| - h \big| \le \delta \}$$



El estimador de la covarianza suele ser evitado ya que m se debe estimar. Es por esto que se introduce la siguiente definición:

ESTACIONARIEDAD INTRÍNSECA:

SE DICE QUE Z ES INTRÍNSECAMENTE ESTACIONARIO SI:

$$\mathbb{E}\big(Z(s+h) - Z(s)\big) = 0 \quad h \in \mathbb{R}^d$$

$$\mathbb{V}\big(Z(s+h) - Z(s)\big) = 2\gamma(h)$$

LA FUNCIÓN $2\gamma(h)$ SE CONOCE COMO VARIOGRAMA Y MIDE LA DISIMILARIDAD DEL PROCESO Z CON SU TRASLADO EN h UNIDADES.

EJ:

$$\gamma(h) = \frac{1}{2} \|h\|^{2H} \quad ; \quad 0 \leq H \leq 1$$

Obs: ESTACIONARIEDAD DE 2^{do} ORDEN IMPLICA INTRÍNSECAMENTE ESTACIONARIO. EL RECÍPROCO NO ES CIERTO. POR EJ, SI $\gamma(h)$ ES COMO EN EL EJEMPLO ANTERIOR, SE TIENE QUE

$$\begin{aligned} 2\gamma(h) &= \mathbb{V}\big(Z(s+h) - Z(s)\big) \\ &= \mathbb{V}\big(Z(s+h)^2\big) + \mathbb{V}\big(Z(s)\big) - 2\,\mathbb{V}\big(Z(s+h)\,Z(s)\big) \\ &= C(s+h, s+h) + C(s,s) - 2\,C(s+h, s) \end{aligned}$$

SI $h \to \infty$, ENTONCES $\gamma(h) \to \infty$. ESTO IMPLICA QUE $C(s+h, s)$ NO ES ACOTADO O EN ALGUNA DIRECCIÓN

AL IGUAL QUE LA COVARIANZA, UN VARIOGRAMA DEBE SATISFACER CONDICIONES ESPECÍFICAS.

UNA FUNCIÓN $\gamma(h)$ ES VARIOGRAMA SI:

1.- $\gamma(h) \geq 0$
2.- $\gamma(0) = 0$
3.- $\gamma(h) = \gamma(-h)$

4.- $\gamma(h)$ ES SEMIVARIOGRAMA SI Y SOLAMENTE SI PARA TODO $N \in \mathbb{N}$, LOCALIZACIONES ESPACIALES $s_1, \dots, s_N$, PESOS $w_1, \dots, w_N$ TALES QUE $\sum_{i=1} w_i = 0$, SE CUMPLE QUE:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \gamma(s_i - s_j) \leq 0$$

<u>DEMOSTRACIÓN:</u>

SEA $N \geq 1$ Y CONSIDEREMOS UN CAMPO ALEATORIO INTRÍNSECAMENTE ESTACIONARIO. LUEGO:

$$\left(\sum_{i=1}^{N} w_i Z(s_i) \right)^2 = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \left(Z(s_i) - Z(s_j) \right)^2$$

LUEGO:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \gamma(s_i - s_j) = -2 \mathbb{V}\left(\sum_{i=1}^{N} w_i Z(s_i) \right) \leq 0$$

PARA RELACIONAR EL VARIOGRAMA CON LA FUNCIÓN DE COVARIANZA SE DEBE INCORPORAR UNA HIPÓTESIS ADICIONAL:

ERGODICIDAD: UN CAMPO ALEATORIO Z CON FUNCIÓN DE MEDIAS $\mu(s) = \mu_0$ SE DICE QUE ES ERGÓDICO SI

$$\lim_{V \to \infty} \frac{1}{|V|} \int_V Z(s, w)\, ds = \mu_0$$

TEOREMA ERGÓDICO DE SLUTSKY:

SEA Z UN PROCESO ESTACIONARIO DE SEGUNDO ORDEN. LUEGO, Z ES ERGÓDICO SI PARA TODA REGIÓN $V \subset \mathbb{R}^d$ SE CUMPLE QUE:

$$\lim_{V \to \infty} \frac{1}{|V|} \int_V C(h)\, dh = 0$$

COROLARIO: SEA Z UN PROCESO ESTACIONARIO DE 2DO ORDEN. LUEGO Z ES ES ERGÓDICO SI $C(h) \xrightarrow[h \to \infty]{} 0$

RESUMEN: → ESTACIONARIEDAD ⟹ INTRÍNSECAMENTE ESTACIONARIO

→ INTRÍNSECO ESTACIONARIO + ERGÓDICO ⟹ ESTACIONARIO 2DO ORDEN

LUEGO $\gamma(h) = C(0) - C(h)$

Para estimar $\gamma(h)$ se suele proceder mediante su estimador de momentos:

$$\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} \left(Z(s_i) - Z(s_j)\right)^2$$

Propiedades de covarianzas estacionarias y variogramas

Sobre la construcción de covarianzas estacionarias:

Para construir funciones de covarianza debemos recordar que satisfacen la definición de positivo definido. Luego,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j C(s_i, s_j) \geq 0 \qquad \forall N\in\mathbb{N}, \alpha_1,\dots,\alpha_N\in\mathbb{R}, s_1,\dots,s_N \in D$$

Si consideramos la noción de estacionariedad de 2do orden se tiene que:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j C_0(s_i - s_j) \geq 0$$

Luego, sería razonable establecer un resultado de caracterización que explote esta propiedad.

UNA ALTERNATIVA ES ESCRIBIR C_0 EN FUNCIÓN DE SU DENSIDAD ESPECTRAL Y ENCONTRAR CONDICIONES SOBRE ESTA. ESPECÍFICAMENTE:

$$C(h) = \int_{\mathbb{R}^d} e^{\hat{i}\langle w, h\rangle} f(w)\,dw \,, \quad h \in \mathbb{R}^d$$

LUEGO:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j C(S_i - S_j) = \int_{\mathbb{R}^d} \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j e^{\hat{i}\langle w, S_i - S_j\rangle} f(w)\,dw$$

$$= \int_{\mathbb{R}^d} \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j e^{\hat{i}\langle w, S_i\rangle} e^{-\hat{i}\langle w, S_j\rangle} f(w)\,dw$$

$$= \int_{\mathbb{R}^d} \left| \sum_{k=1}^{N} \alpha_k e^{-\hat{i}\langle w, S_k\rangle} \right|^2 f(w)\,dw$$

LUEGO, SI f ES UNA FUNCIÓN REAL Y POSITIVA, C ES DEFINIDA POSITIVA.
ESTE RESULTADO NOS ENTREGA UNA IDEA SOBRE UN RESULTADO MÁS GENERAL.

<u>TEOREMA DE BOCHNER</u>: SUPONGA QUE $C_0 : \mathbb{R}^d \to \mathbb{R}$ ES UNA FUNCIÓN CONTINUA. LUEGO C_0 ES FUNCIÓN DE COVARIANZA DE ALGÚN CAMPO ALEATORIO GAUSSIANO ESTACIONARIO DE 2° ORDEN SÍ Y SOLAMENTE SI

$$C_0(h) = \int e^{\hat{i}\langle w, h\rangle} d\mu(w)$$

PARA ALGUNA MEDIDA μ SIMÉTRICA FINITA Y NO-NEGATIVA.

NOTAMOS QUE C ES LA TRANSFORMADA DE FOURIER DE f. POR EL TEOREMA DE INVERSIÓN DE FOURIER SE TIENE ENTONCES QUE:

$$f(w) = \int_{\mathbb{R}^d} C(h)\, e^{-\hat{i}\langle w, h\rangle}\, dh$$

Ej:

$$C(h) = \sigma^2 e^{-\beta \|h\|^2} \quad ; \ \beta > 0, \ \sigma^2 > 0$$

$$f(w) = \sigma^2 \left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d} e^{-\hat{i}\langle w, h\rangle} e^{-\beta \|h\|^2}\, dh$$

Recordamos que:

$$\int_{-\infty}^{\infty} e^{-\hat{i}wt} e^{-\beta t^2}\, dt = \int_{-\infty}^{\infty} e^{-\beta\left(t + \frac{\hat{i}w}{(2\beta)}\right)^2}\, dt = e^{-\frac{w^2}{(4\beta)}} \left(\frac{\pi}{\beta}\right)^{1/2}$$

Luego:

$$f(w) = \frac{\sigma^2}{2^d (\beta\pi)^{d/2}}\, e^{-\frac{w^2}{4\beta}} \qquad w \in \mathbb{R}^d$$

Ej 2:

Sea

$$f(\omega) = \frac{\sigma^2 \Gamma\left(\nu + \frac{d}{2}\right)}{\pi^{d/2+\nu} \Gamma(\nu) \phi^{\nu}} \left(\frac{1}{\pi^2 \phi^2} + \|\omega\|^2 \right)^{-\left(\nu + \frac{d}{2}\right)}$$

Luego se obtiene la familia de covarianzas de Matern:

$$C_0(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{h}{\phi}\right)^{\nu} K_{\nu}\left(\frac{h}{\phi}\right)$$

Donde $K_\nu(\cdot)$ es la función de Bessel de 2^{da} especie de orden ν

Obs: bajo isotropía la transformada de Fourier toma la siguiente forma:

$$f(\rho) = 2\pi \rho^{1-\frac{d}{2}} \int_0^{\infty} h^{d/2} J_{\frac{d}{2}-1}(2\pi \rho h) C_0(h)\, dh$$

$$C_0(h) = 2\pi h^{1-\frac{d}{2}} \int_0^{\infty} \rho^{d/2} J_{\frac{d}{2}-1}(2\pi \rho h) f(\rho)\, d\rho$$

Donde $J_\nu(\cdot)$ es la función de Bessel de 1^{era} especie de orden ν

Para las dimensiones de interés se tiene que:

$d=1$

$$C_0(h) = \int_0^\infty \cos(2\pi h \rho) f(\rho) d\rho$$

$d=2$

$$C_0(h) = \int_0^\infty J_0(2\pi h \rho) f(\rho) d\rho$$

$d=3$

$$C_0(h) = \int_0^\infty \rho \sin(2\pi \rho h) f(\rho) dr$$

Propiedades

Sea $C(h)$ una función de covarianza estacionaria, entonces se tiene que:

- Si $C_1, \dots, C_K$ son funciones de covarianza estacionarias y $\alpha_1, \dots, \alpha_N \in \mathbb{R}_0^+$ entonces

$$\sum_{i=1}^{K} \alpha_i C_i(h) \quad y \quad \prod_{i=1}^{K} C_i(h) \quad \text{son}$$

funciones de covarianza estacionarias

Procedimientos de estimación

- Regresión lineal
- LS
- Maximum likelihood

→ **Selección de modelos:**

AIC, BIC, Cross validatio
Leave-one-out

→ Predicción (kriging)