CIMAT

Centro de Investigación en Matemáticas, A.C.

Análisis espacio-temporal con modelos jerárquicos Bayesianos e INLA de la violencia en México

T E S I S

Que para obtener el grado de

Maestro en Ciencias

con especialidad en

Probabilidad y Estadística

Presenta

Mario Enrique Carranza Barragán

Director de Tesis:

Dra. Lilia Leticia Ramírez Ramírez

Co-director de Tesis:

Dr. Fernando Alarid Escudero

Autorización de la versión final

Guanajuato, Gto a 5 de diciembre de 2022

CIMAT

Centro de Investigación en Matemáticas, A.C.

Acta de Examen de Grado

Acta No.: 183

Libro No.: 002

Foja No.: 183

En la ciudad de Guanajuato, Gto., siendo las 11:00 horas del día 14 de octubre del año 2022, se reunieron los miembros del jurado integrado por los señores:

DR. ROGELIO RAMOS QUIROGA **(CIMAT)**
DR. ROMÁN AGUIRRE PÉREZ **(IIMAS-UNAM)**
DR. LILIA LETICIA RAMÍREZ RAMÍREZ **(CIMAT)**

bajo la presidencia del primero y con carácter de secretario el segundo para proceder a efectuar el examen que para obtener el grado de

MAESTRO EN CIENCIAS
CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

Sustenta

MARIO ENRIQUE CARRANZA BARRAGAN

En cumplimiento con lo establecido en los reglamentos y lineamientos de estudios de posgrado del Centro de Investigación en Matemáticas, A.C., mediante la presentación de la tesis

"ANÁLISIS ESPACIO-TEMPORAL CON MODELOS JERÁRQUICOS BAYESIANOS E INLA DE LA VIOLENCIA EN MÉXICO"

Los miembros del jurado examinaron alternadamente al (la) sustentante y después de deliberar entre sí resolvieron declararlo (a):

APROBADO

DR. ROGELIO RAMOS QUIROGA
Presidente

DR. ROMÁN AGUIRRE PÉREZ
Secretario

DRA. LILIA LETICIA RAMÍREZ RAMÍREZ
Vocal

CIMAT DIRECCIÓN GENERAL

DR. VÍCTOR MANUEL RIVERO MERCADO
Director General

000741

Dedicatoria

A Dios,

" *Nada te turbe, Nada te espante, Todo se pasa, Dios no se muda, La paciencia Todo lo alcanza; Quien a Dios tiene Nada le falta: Sólo Dios basta.*"

Santa Teresa de Ávila

"*Creer es un acto del entendimiento que asiente a la verdad divina por imperio de la voluntad movida por Dios mediante la gracia. Del mismo modo que es mejor iluminar que solamente brillar, asimismo es más grande dar a los demás las cosas contempladas que solamente contemplarlas.[...] Respecto de Dios es mejor amarlo que conocerlo, porque el conocimiento hace que las cosas vengan a nosotros y se adapten a nuestra manera de ser; pero el amor, que es la caridad, nos hace salir de nosotros y nos lanza al objeto amado.*"

Santo Tomas de Aquino

" *It isn't that they can't see the solution. It is that they can't see the problem.[...] Truth can understand error, but error cannot understand truth. [...] Poets do not go mad; but chess-players do. Mathematicians go mad, and cashiers; but creative artists very seldom....The poet only asks to get his head into the heavens. It is the logician who seeks to get the heavens into his head. And it is his head that splits. [...] Christianity never promised that it would impose universal peace. It had a great deal too much respect for personal liberty. [...] In freeing ourselves from Christianity, we have only freed ourselves from freedom. [...] Right is right, even if nobody does it. Wrong is wrong, even if everybody is wrong about it.*"

Gilbert Keith Chesterton

"*Sálvame, oh Dios, de estar completamente seguro; mantenme inseguro hasta el final, así cuando reciba la eterna bendición, podré estar completamente seguro que la tengo por gracia. Es un juego de sombras vacío para dar garantías de que uno cree que es por gracia,*

para luego estar completamente seguro. La verdad, la expresión esencial de su ser por la gracia es el mismo temor y temblor de la inseguridad. Allí yace la fe. "

Søren Kierkegaard

"*Supposing there was no intelligence behind the universe, no creative mind. In that case, nobody designed my brain for the purpose of thinking. It is merely that when the atoms inside my skull happen, for physical or chemical reasons, to arrange themselves in a certain way, this gives me, as a by-product, the sensation I call thought. But, if so, how can I trust my own thinking to be true? It's like upsetting a milk jug and hoping that the way it splashes itself will give you a map of London. But if I can't trust my own thinking, of course I can't trust the arguments leading to Atheism, and therefore have no reason to be an Atheist, or anything else. Unless I believe in God, I cannot believe in thought: so I can never use thought to disbelieve in God.* "

C.S. Lewis

Agradecimientos

Agradezco a mis directores de tesis, Leticia y Fernando, por su tiempo, paciencia, consejo y aliento en la elaboración de este trabajo.

Gracias al Centro de Investigación en Matemáticas (CIMAT), profesores, administrativos y mantenimiento.

Agradezco también al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo económico con una beca de maestría.

A mis padres, Marlond y Alicia, porque siempre han confiado en mi y por mantenerme con los pies en la tierra.

A mi hermano, Marlond, por el apoyo incondicional que siempre me ha brindado y compartir los buenos momentos como grandes amigos y brindarnos mutuo apoyo.

A mis abuelos, Alicia y Samuel, por el cariño y aliento que siempre me dan.

A mi tío Samuel, por su apoyo y animo en todo momento.

A mis sinodales, Román y Rogelio, por el tiempo dedicado a la revisión de este trabajo y por sus valiosos comentarios.

A mis profesores, Leticia, Rolando, Miguel, Johan, Andrés, Eloisa, Rogelio, Enrique por todo su apoyo en durante mi estancia en CIMAT.

A mis amigos, Verónica, Salim, Diego, Gustavo, Joshue, Isaias, Melina y Juan por su amistad y apoyo durante la maestría.

Resumen

En esta tesis nuestro primer objetivo es proponer el uso de herramientas y modelos estadísticos que ayuden a entender la dinámica y aumento de la violencia en México provocada por las organizaciones criminales durante el periodo de interés (2007-2011). Para ello emplearemos herramientas del análisis exploratorio y la ciencia de datos, estimación no paramétrica, modelos espaciales (geoestadísticos y procesos puntules), así como de modelos jerárquicos con componente espacial y temporal (de Campo Aleatorio Gaussiano Markoviano Latente).

Nuestro segundo objetivo consiste en desarrollar la teoría detrás del método conocido como aproximación de Laplace anidada integrada (o por sus siglas en inglés INLA) que permite la inferencia aproximada de ciertas clases modelos Bayesianos. INLA incluye un conjunto de herramientas estadísticas y computacionales basadas en el uso iterado de la aproximación de Laplace, que a diferencia de otros métodos de inferencia como MCMC su ejecución es mucho más rápida pero sus soluciones no son exactas. Para aplicar este método debemos limitarnos a los modelos con Campo Aleatorio Gaussiano Markoviano Latente (o por sus siglas en inglés LGMRF). Esta restricción, que es también considerada en muchos de los modelos más usados, permite integrar de forma eficiente el campo aleatorio latente y con ello obtener una función de densidad posterior de los llamados hiperparámetros. Esta integración emplea la llamada aproximación de Laplace y resulta especialmente eficiente, incluso si el LGMRF es muy grande, debido a la estructura (Markoviana y Gaussiana) que tiene.

El CIDE ha proporcionado los registros geocodificados de todos los eventos violentos (Ejecuciones, Enfrentamientos y Agresiones) ocurridos durante la administración de Felipe Calderón, a los que se aplicaron diversas herramientas de análisis exploratorio. Entre ellos, para distintos intervalos de tiempo, se consideró la aparición de eventos violentos como un proceso puntual y se estimó la función de intensidad mediante kernels. También, como parte de un análisis exploratorio y preliminar, se empleo Kriging para hacer predicción lineal de la incidencia por número de habitantes. Adicionalmente, se consideraron series temporales (por semestre) para cada estado y se aplicaron herramientas de análisis de conglomerados

(jerárquico), reescalamiento multidimensional, PCA y k-medias.

Hecho el análisis exploratorio, se propusieron modelos jerárquicos Poisson para modelar el número de eventos por estado. En estos modelos se considera un efecto espacial (definido por los estados vecinos con los que comparten frontera) y una componente temporal (semestre inmediato anterior y siguiente). Inicialmente se analizaron 3 distintos tipos de verosimilitud (Poisson, Zero Inflated Poisson y Hurdle) y dos tipos de efecto temporal (Autoregresivo de orden 1 y Random Walk) y se realizó su inferencia a través de R-INLA (Aproximación de Laplace Anidada Integrada implementado desde R). Al ver que no existía mucha diferencia entre estos modelos finalmente solo comparamos modelo Poisson contra ZIP con efecto temporal autoregresivo de orden 1. Se emplearon distintos criterios Bayesianos para comparar estos modelos, tales como la verosimilitud marginal y sus variantes (BIC, DIC y WAIC). Se observó que en general los criterios tradicionales no capturan la idoneidad de estos modelos con cero inflado. Los criterios de tipo validación cruzada, en especial el criterio Transformación Integral Prediciva (PIT), parecen ser flexibles e interpretables para comparar los diferentes modelos propuestos. Se propone un resumen de la información de los PIT a través de estadísticos pruebas de bondad de ajuste.

Palabras Clave

Violencia y Drogas, Modelos jerárquicos, Estadística espacial, Aproximación de Laplace, Inferencia Bayesiana, selección de modelos Bayesianos

Índice

Referencias **131**

Índice de figuras

Índice de tablas

Introducción y organización

El tema general de este trabajo es el uso de modelos de estadística espacial, en particular modelos jerárquicos espacio-temporales Bayesianos, para analizar la violencia en México asociada al crimen organizado.

Se desea comprender mejor, a través de un modelo estadístico, el fenómeno de la violencia asociado al crimen organizado para así justificar o evaluar las políticas publicas que se toman para combatirla, ya sean medidas económico-sociales o bien intervención de las fuerzas armadas federales o estatales. Deseamos poder proporcionar elementos estadísticos con los que contestar cuál es la dinámica de la violencia asociada al crimen organizado en México y qué factores están relacionados a su intensificación.

El incremento de los niveles de violencia e inseguridad en México se ha atribuido a la Guerra contra las Drogas. Este aumento en la violencia tiene repercusiones muy importantes en la actividad económica, la cohesión social y orden político en diferentes escalas. Es de capital importancia comprender la dinámica de los eventos violentos, su evolución en el tiempo y espacio, así como la asociación que tienen con variables económico y socio-demográficas, ya que este entendimiento puede apuntar hacia una mejor medición en el impacto de los controles implementados. La incorporación de un modelo estadístico-matemático se convierte entonces en una herramienta para la mejor toma de decisiones de diferentes agentes (familias, empresas, gobiernos). Hasta ahora, no se ha considerado un modelo que integre las componentes espacial y temporal, que sería muy útil para la comprensión de la dinámica.

Mediante la propuesta y ajuste de un modelo estadístico espacio-temporal se pretende ayudar a identificar las trayectorias de enfrentamientos y asesinatos a nivel municipal y regional que se presentó durante la guerra contra las drogas en el sexenio de Felipe Calderón, así como los patrones donde los enfrentamientos se volvieron más letales y sus tendencias, mediante variables o características (económicas, geográficas y políticas) con los que puede asociarse. Se emplearán modelos jerárquicos Bayesianos y la inferencia se hará mediante aproximaciones de Laplace anidadas integradas.

Este trabajo busca cumplir dos objetivos. El primero es justificar y aplicar un conjunto

de modelos y herramientas estadísticas, especialmente del área de la estadística espacial, a un problema de vinculación real. No solamente se busca desarrollar la capacidad de hacer análisis de datos con una estructura especial (conteos relacionados espacio temporalmente) sino profundizar en las etapas de comprensión del problema así como la interpretación útil al investigador de los resultados de los distintos modelos.

Como segundo objetivo, desarrollamos y explicamos los distintos pasos, tanto estadísticos/ inferenciales y numéricos/ computacionales que componen la estrategia de inferencia Bayesiana conocida como Aproximación de Laplace Anidada Integrada o INLA, por sus siglas en inglés. La aproximación de Laplace es una técnica clásica de para aproximar integrales y muy recurrida en la inferencia Bayesiana antes de la llegada de los métodos de integración numérica con Monte Carlo vía Cadenas de Markov (MCMC). A diferencia de otros métodos de integración numérica vía cuadratura o Monte Carlo (simulación), la aproximación no esta justificada por un resultado de convergencia respecto al número de términos (sumandos) o de número de pasos en la cadena de Markov sino que la aproximación se justifica por el tamaño de muestra sea suficientemente grande.

Si bien la aproximación de Laplace no es nueva, relativamente lo es la herramienta INLA. Para poder aplicarse, uno debe limitarse a considerar los modelos con Campo Latente Gaussiano Markoviano. Resulta que esta restricción no es muy fuerte ya que muchos de los modelos más empleados, y otros que nos interesan como los de componente espacio temporal, pertenecen a esta clase de modelos. La idea es poder especificar la dependencia (espacial-temporal) de las variables observadas mediante los campos latentes (que son variables latentes con una estructura de dependencia). El primer uso de la aproximación de Laplace (función de verosimilitud de los hiperparámetros) es usada para integrar respecto al campo latente, y demostramos que esta aproximación es equivalente a emplear una aproximación Gaussiana (mediante media y varianza) de a distribución posterior (condicional a los datos) del campo latente. Esta aproximación se simplifica si trabajamos con las matrices de precisión tanto del campo Markoviano y una matriz diagonal que depende del modelo (condicional al campo latente) de la variable de salida. Aplicada la aproximación de Laplace por primera vez, tenemos una función marginal (respecto al campo latente) sobre los hiperparámetros $p(\boldsymbol{\theta}|\boldsymbol{y})$ que sólo dependen de los datos.

Para explicar la estructura de la tesis seguiremos el esquema en la Figura 1. En este diagrama cada rectangulo de color corresponde a un capítulo de la tesis referido por el número romano en blanco. En el primer cápitulo desarrollamos la teoría y construcción de los procesos estocásticos necesarios para las herramientas de análisis espacial. En particular se justifican las herramientas de estimación kernel para la función de intensidad de un proceso puntual y la predicción con Kriging. En el segundo capítulo introducimos varios conceptos preeliminares para INLA. Entre ellos están los campos latentes, la aproximación clásica de Laplace y los criterios de selección de modelos. En el tercer capítulo entramos propiamente a la metodología de INLA donde hablamos de la construcción de la matriz de precisión del Campo Gaussiano Markoviano Latente (CGML), su actualización, su diagonalización y cómo a partir de este podemos obtener las densidades marginales posteriores de los hiperparámetros, el campo latente y las variables de salida. En el cuarto capítulo realizamos

un análisis exploratorio de los datos, empleamos las herramientas de análisis espacial y aplicamos modelos INLA sin covariables. En el quinto capítulo proponemos un análisis exploratorio de las covariables que utilizamos para los modelos INLA con covariables. Explicamos las pruebas de Kormogorov-Smirnov y χ^2 de Pearson para evaluar conjuntamente las Transformadas Integrales Predictivas (PIT) para evaluar la calidad de ajuste del modelo.

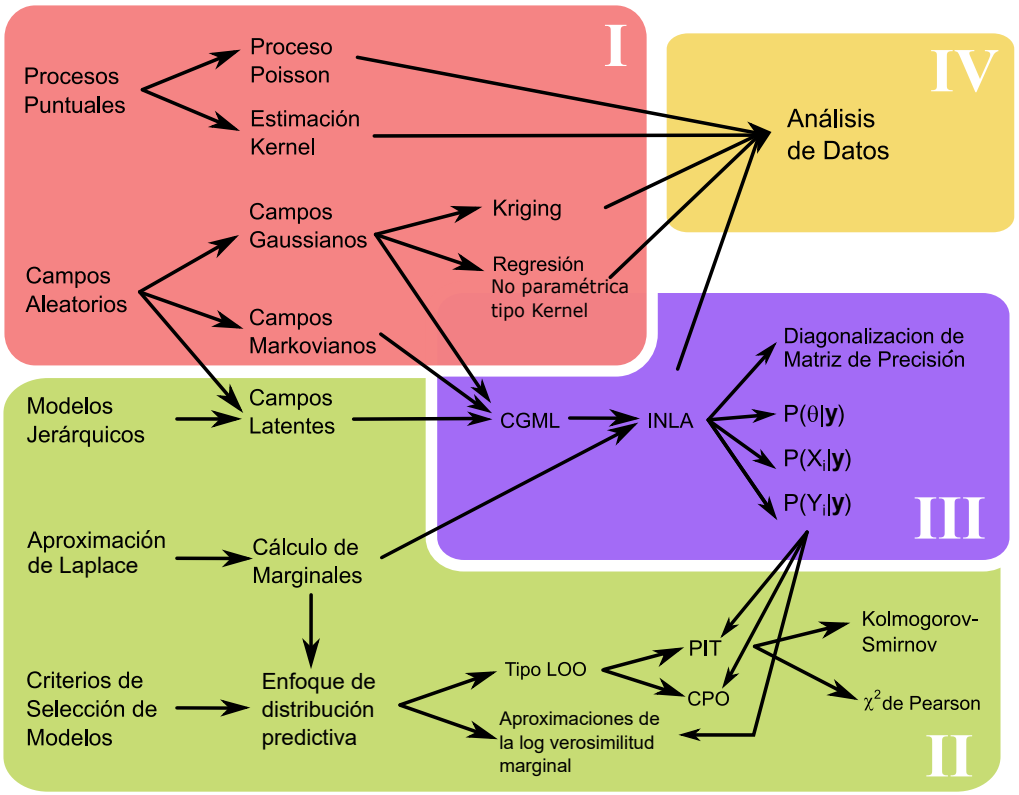


Figura 1: Mapa de temas

CAPÍTULO 1

Preeliminares de modelos espaciales

1.1. Introducción a los modelos espaciales

Conviene, en primer lugar, dar una breve introducción a los modelos espaciales para así poder enmarcar dentro de esta gran variedad de modelos con los que podemos analizar los datos de eventos violentos.

En problemas espaciales, las observaciones vienen de procesos $X = \{X_s, s \in S\}$ indexados por un conjunto espacial S (numerable o no dependiendo del tipo de datos), y donde X_s toma valores en un espacio de estados E, en general $E \subseteq \mathbb{R}$ o $E \subseteq \mathbb{C}$. El caso más usual es que $S \subseteq \mathbb{R}^2$, sin embargo también se da que S sea unidimensional o bien sea $S \subseteq \mathbb{R}^3$ como se da en ciencias de la tierra y mineralogía. Más aun, en estadística Bayesiana y en simulación la dimensión puede ser aún mayor que 3 o incluso pueden ser otros espacios topológicos no vectoriales. Las tres estructuras principales de datos en estadística espacial son

- *Datos geoestadísticos*

  En este tipo de datos S es un subespacio continuo de $\mathbb{R}^d$ y el campo aleatorio $\{X_s, s \in S\}$ se observa en n puntos fijos $\{s_1, ..., s_n\} \subset S$ que toman valores en E, que es real-valuado. Usualmente se desea poder reconstruir X_s sobre todo el espacio S. Este enfoque es el que usamos al hacer predicciones como con Kriging pues son predicciones de $X_{s'}$ en cualquier punto de $s' \in S$, pero para visualizar usualmente elegimos una celosía.

- *Datos en una celosía o en redes fijas*

  En este contexto, $S = \{s_1, \ldots, s_n\}$ es un conjunto discreto no aleatorio, usualmente $S \subset \mathbb{R}^d$ y X_s se observa en todo S. Es posible que los puntos sean regiones geográficas cuyas adyacencias están representadas por un grafo. Un caso particular, que

además es clásico, es el análisis de imágenes donde S es usualmente un conjunto equiespaciado de zonas de mismo tamaño o pixeles. En estos modelos los objetivos son comúnmente cuantificar la correlación espacial y realizar predicción de X_s, que en este contexto también se le llama restauración de imágenes. En este trabajo este es el enfoque que usamos con los modelos por área (estado o municipio). Las X_s se distribuyen como campos aleatorios Markovianos Gaussianos donde el grafo de adyacencia especifica la propiedad de Markov para estas variables aleatorias.

- *Datos puntuales*

  Aquí, el conjunto de lugares de observación $\boldsymbol{s} = \{s_1, s_2, ..., s_n\}, s_i \in S \subset \mathbb{R}^d$ es aleatorio, así como también lo es el número $n = n(\boldsymbol{s})$ de sitios de observaciones. Así, $\boldsymbol{s}$ es la observación de un proceso puntual observado en el espacio S. Destacamos que a diferencia de los modelos para datos geoestadísticos y en redes fijas, aquí las s_i no son simplemente índices, sino que son variables aleatorias. Se dice que el proceso $\boldsymbol{S}$ esta *marcado* si cada s_i tiene asociado un valor. Una pregunta central en el análisis estadístico de procesos puntuales es saber si la distribución de puntos es esencialmente aleatoria o si tiene propiedades agregativas o repulsivas entre los puntos. A diferencia de los otros enfoques, más observaciones no significa una mejor estimación o predicción sino que habla de la intensidad de los eventos en ciertas áreas.

Esta categorización de los tipos de datos espaciales es la que aparece en Cressie, 1992. Para más detalles sobre cualquiera de estos tipos de datos se recomienda ver la sección correspondiente en el libro antes mencionado.

1.2. Procesos puntuales

Conviene, en primer lugar, dar una descripción formal los modelos espaciales para así poder enmarcar dentro el modelo Poisson no homogéneo.

Procesos puntuales

Un proceso puntual modela la distribución aleatoria de puntos en un espacio. Para definir formalmente este concepto, supongamos que todos los puntos cuya distribución queremos estudiar viven en un espacio S, el cual para efectos de la definición del proceso Poisson no homogéneo, será siempre un subconjunto de algún espacio euclidiano de dimensión finita.

Definición medida puntual

Denotemos por $\mathcal{B} := \mathcal{B}(S)$ la σ-álgebra de Borel asociada a S. Para $s \in S$, $B \in \mathcal{B}$ definimos la medida $z_s : \mathcal{B} \to \mathbb{R}_+$ como

$$I_s(B) = \begin{cases} 1 & s \in B \\ 0 & s \notin B \end{cases}$$

Sea $\boldsymbol{s} = \{s_j, j \in I \subseteq \mathbb{N}\}$ una colección numerable de puntos en S, los cuales no necesariamente son distintos. Sea

$$\phi(\cdot) := \sum_{j \in I} I_{s_j}(\cdot).$$

Es fácil ver que μ es una medida y que, si $K \subseteq S$ es compacto y elemento de la σ-álgebra de B, entonces $\phi(K) < \infty$.

La medida ϕ definida anteriormente se denomina medida puntual en S. Para todo $s \in S$, $\phi(\{s\})$ es la multiplicidad de s y decimos que s es simple si su multiplicidad es a lo más 1. Una medida puntual ϕ en S es simple si todo $s \in S$ es simple (bajo ϕ). Notemos que para todo $s \in \boldsymbol{s}$ la multiplicidad es al menos 1. Recordemos $\mathcal{B} := \mathcal{B}(S)$ es la σ-álgebra de Borel y los conjuntos singoletes son cerrados, por lo que están en $\mathcal{B}$.

Definición de proceso puntual

Sea Φ el conjunto de todas las medidas puntuales definidas en S y sea $\mathcal{N}$ la menor σ-álgebra de subconjuntos de Φ que contiene a todos los subconjuntos de la forma $\{\phi \in \Phi : \phi(B) = n\}$ para todo $n \in \{0, 1, 2, \dots\}$ y $B \in \mathcal{B}$. Esta σ-álgebra también puede entenderse como la menor σ-álgebra que hace medibles a los mapeos (de Φ a $\{0, 1, 2, \dots\}$) dados por $\phi \to \phi(B)$, para todo $B \in \mathcal{B}$.

Un proceso puntual en S es un mapeo medible de un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ a $(\Phi, \mathcal{N})$. Es decir, un proceso puntual es un elemento aleatorio de Φ. Si tomamos $\omega \in \Omega$, entonces $N(\omega, \cdot)$ es una medida puntual en S y $N(\omega, B)$ es el número de puntos en $B \subseteq S$, para la realización ω. La medida $N(\omega, \cdot)$ es aleatoria. Por ejemplo, si consideramos $K : \Omega \to \mathbb{N}$ (los naturales) y X tal que $X(\omega)$ dado $K(\omega) = k$ es un conjunto de puntos $\{s_1, s_2, ..., s_k\} \subseteq S$, entonces $N(\omega, \cdot) = \sum_{j=1}^{k} I_{s_j}(\cdot) \in \Phi$.

En este ejemplo, es claro que si cambiamos de ω cambiamos el valor k y también cambian los puntos $s_1, s_2, ..., s_k$. La siguiente proposición presenta un criterio simple para verificar si un mapeo $N : \Omega \to \Phi$ es un proceso puntual.

Proposición: N es un proceso puntual si y sólo si el mapeo $\omega \to N(\omega, B)$ para $B \in \mathcal{B}$ fijo, es medible de (Ω, F) a $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$.

El enunciado de la proposición es equivalente a decir que N es un proceso puntual si y sólo si, para $B \in \mathcal{B}$ fijo, $N_B := N(\cdot, B)$ es una variable aleatoria.

Sean $X_1, X_2, ..., : (\Omega, \mathcal{F}, P) \to (S, \mathcal{B})$. El conjunto de variables aleatorias $\{N_B, B \in S\}$ donde $N_\cdot = \sum_{j=1}^{\infty} I_{X_j}(\cdot)$, es un proceso puntual en $(S, \mathcal{B})$.

Definición medida de intensidad

Dado un proceso puntual N, se define su medida de intensidad (o simplemente, su intensidad) como la medida λ dada por

$$\lambda(B) = \mathrm{E}[N_B] = \int_\Omega N(\omega, B) P[d\omega], B \in \mathcal{B}.$$

Lema: Sea N un proceso puntual en $(S, \mathcal{B})$. La distribución de los vectores

$$(N_{B_1}, N_{B_2}, ..., N_{B_m}),$$

para todo $m \in \mathcal{N}$ y cualesquiera $B_1, B_2, ..., B_m \in \mathbb{B}$ caracteriza completamente a la distribución del proceso N.

1.2.1. Proceso Poisson homogéneo (CSR) y no homogéneo

Procesos estacionarios e isotrópicos

Decimos que un proceso puntual en $\mathbb{R}^d$ es estacionario si para cada $\xi \in \mathbb{R}^d$, la distribución del proceso puntual $X_\xi = \{X + \xi\}$ es el mismo que el de X. Decimos que X es isotrópico si la distribución de ρX, obtenida al rotar X mediante cualquier matriz de rotación ρ, tiene la misma distribución que X.

En el caso particular de un Campo Aleatorio Gaussiano o GRF (ver Definición 1.4.2) es estacionario si:

- Tiene media cero.
- La covarianza entre dos puntos depende sólo de la distancia y la dirección entre ambos puntos.

Además, un GRF será isotrópico si la covarianza solo depende de la distancia entre los puntos.

Procesos agregativos y repulsivos

Los procesos puntuales se pueden clasificar según las interacciones que existen entre sus puntos. Los procesos agregativos tienden a tener sus eventos en racimos, como los cúmulos estelares donde las estrellas se encuentran juntas. Los puntos de los procesos repulsivos se encuentran alejados entre sí y es muy raro encontrarlos en grupos, como en especies de árboles grandes cuyas raíces impiden el crecimiento de árboles semejantes en espacios cercanos.

Por último, los procesos aleatorios son aquellos que no presentan un patrón definido. Tienen grupos creados por azar, como los procesos de Poisson.

Definición del proceso Poisson

Sea λ una medida de Radón en $\mathcal{B}$ y sea N un proceso puntual. Diremos que N es un proceso puntual de Poisson con intensidad λ si se cumplen las siguientes condiciones:

1. Para todo $B \in \mathcal{B}, N_B \sim$ Poisson$(\lambda(B))$.

2. Si $B_1, B_2 \subset S$ son disjuntos, las variables aleatorias N_{B_1} y N_{B_2} son independientes.

El caso cuando $\lambda \equiv cL$, donde $c \in (0, \infty)$ y L es la correspondiente medida de Lebesgue, diremos que el proceso puntual de Poisson es homogéneo. En caso contrario diremos que el proceso puntual de Poisson es no homogéneo .

1.2.2. Pruebas para contrastar *Complete Spatial Randomness* (CSR)

La importancia de la hipótesis completa aleatoriedad espacial o CSR radica en que es un paso primordial en cualquier análisis de procesos puntuales. Un proceso que satisface CSR es equivalente a un proceso de Poisson homogéneo. Si no se rechaza CSR, no tenemos información para decir que el proceso estudiado tenga un comportamiento distinto al de un proceso de Poisson homogéneo, y un análisis más hondo carecería de sentido. Si se rechaza CSR, quiere decir que el proceso tiene una estructura distinta a la de un proceso de Poisson homogéneo, dando lugar a estudios más complejos como indagar sobre la repulsividad o la formación de grupos del proceso. Es de interés obtener pruebas que disciernan con mayor eficiencia cuándo un proceso satisface CSR.

Este proceso tiene la propiedad de que, condicionado en el número de eventos en una región acotada $A \subset \mathbb{R}^d$ $N(A)$, los eventos del proceso son independientes y uniformemente distribuidos sobre A. Es decir, dado $N(A) = n$, la tupla ordenada de eventos $(s_1, ..., s_n)$ en A^n satisface

$$P(s_1 \in B_1, ..., s_n \in B_n) = \prod_{i=1}^{n} (|B_i|/|A|),$$

donde $B_1, ..., B_n \subset A$ son ajenos y $|X| \equiv \int_X ds$.

Este resultado justifica algunos métodos para plantear las pruebas de hipótesis contrastar CSR. Definido el modelo por regiones se puede usar, por ejemplo, la prueba χ^2 de Pearson.

Método de cuadrantes

El método de cuadrantes consiste en contar el número de eventos en subconjuntos del área de estudio A. Tradicionalmente, estos subconjuntos son rectangulares, pero en principio cualquier forma es válida. Pueden colocarse aleatoriamente o bien de forma contigua.

Bajo CSR, el número de eventos en A_i, de área $|A_i|$ tiene distribución Poisson con media $\lambda|A_i|$, donde λ es la intensidad del proceso Poisson. Así, una prueba para CSR es la prueba de bondad de ajuste con estadístico χ^2 de Pearson.

$$\chi^2 = \sum_{i=1}^{n} \frac{(E_i - O_i)^2}{E_i}$$

con $O_i = n_i$, el número de eventos observados en A_i, y $E_i = \lambda|A_i|$. El estadístico se distribuye asintóticamente como una χ^2 con $n-1$ grados de libertad.

1.2.3. Función de intensidad y estimación vía kernels

Sea $x_1, \ldots, x_n$ una muestra de variables aleatorias independientes e idénticamente distribuidas con función de densidad f. El estimador kernel se define

$$\hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)$$

donde $h > 0$ es el ancho de banda o ventana. Un ejemplo muy usado en la práctica kernel es el kernel Gaussiano, de modo que

$$K_h(x - x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-x_i)^2}.$$

Este tipo de estimadores de densidad tienen varias propiedades asintóticas que justifican su uso en la práctica.

El problema de la elección del ancho de banda puede ser complicado en la práctica. El criterio de optimalidad más comúnmente usado es a través del MISE o Error Cuadrado Integrado Medio

$$\text{MISE}(h) = \text{E}\left[\int (\hat{f}_h(x) - f(x))^2 dx\right].$$

Una alternativa es emplear la llamada regla de pulgar (Rule of thumb) de Silverman, $\sqrt{h} = \left(\frac{4}{3}\right)^{\frac{1}{5}} n^{\frac{'1}{5}} \sigma$ donde σ es la desviación estándar de X.

1.2.3.1. Estimación de densidades multivariadas mediante kernels

A continuación describiremos las herramientas de estimación de densidades multivariadas mediante kernels. Sea $x_1, \ldots, x_n$ una muestra de un vector aleatorio d-variado cuya distribución conjunta esta determinada por la función de densidad f. El estimador kernel se define

$$\hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - x_i)$$

donde $x = (x_1, \cdots, x_d)^T$, $x_i = (x_{i1}, \cdots, x_{id})^T, i = 1, 2, ..., n$ con vectores de dimensión d. H es una matriz $d \times d$ simétrica y definida positiva, llamada matriz de ancho de banda o de suavizamiento. K es la función kernel que es una función de densidad simétrica multivariada y $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$.

La elección de la función de kernel K no es crucial para la precisión del estimador kernel, por lo que es común emplear la distribución Normal multivariada de modo que: $K(H)x = (2\pi)^{-d/2}|H|^{-1/2}e^{-\frac{1}{2}x^T H^{-1}x}$ donde en este caso H será la matriz de covarianza de la Normal multivariada.

Así como en el caso univariado, incluso aún más, el problema de la elección de la matriz de suavizamiento suele ser complicado en la práctica. Como en el caso univariado, el criterio de optimalidad más comúnmente usado es el MISE o Error Cuadrado Integrado Medio

$$\text{MISE}(H) = \text{E}\left[\int (\hat{f}_H(x) - f(x))^2 dx\right].$$

También, una alternativa es emplear la llamada regla de pulgar de Silverman para el caso multivariado, $\sqrt{H_{ii}} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{1}{d+4}}\sigma_i$ donde σ_i es la desviación estándar de la i-ésima variable y $H_{ij} = 0, i \neq j$.

Función de intensidad

Sea $N(s, a)$ el número de eventos en un cuadrado de dimensiones $a \times a$ en la localización s. Consideremos

$$\lambda_a(s) \equiv P(N(s, a) > 0)/a^2.$$

Si $\lambda_a(s) \to \lambda(s)$ cuando $a \to 0$, para todo $s \in A$, llamamos a $\{\lambda(s); s \in A\}$ la función intensidad. Para un nodo s de la celosía cuadrada de dimensión $a \times a$, el conteo de cuadrante por unidad de área, $N(s, a)/a^2$, es un estimador insesgado de $\int_0^a \int_0^a \lambda(\boldsymbol{u}+s)d\boldsymbol{u}/a^2$. Cuando $\lambda(\cdot)$ no varia mucho en el cuadrado $a \times a$, esta integral es aproximadamente igual a $\lambda(s)$.

Por lo tanto, el conjunto de conteos de cuadrante por unidad de área $\{N(s, a)/a^2$: con s un punto del cuadrado $a \times a$ de la celosía de $A\}$, estima $\{\lambda(s) : s \in A\}$.

Estimación vía kernels de la función de intensidad

Este problema se parece mucho al de estimación de función de densidad multivariada. De hecho, los estimadores kernel pueden extenderse para obtener estimadores no paramétricos de $\lambda(\cdot)$.

Sea $(s_1, s_2, ..., s_n)$ las localizaciones espaciales de $n = N(A)$ eventos en la región acotada de estudio $A \subset \mathbb{R}^d$. Considere el estimador de la forma

$$\hat{\lambda}_h(s) \equiv \frac{1}{p_h(s)}\left\{\sum_{i=1}^{n} \kappa_h(s - s_i)\right\}, s \in A$$

donde $\kappa_h(\cdot)$ es una función kernel simétrica en el origen, $h > 0$ determina el suavizamiento y $p_h(s) \equiv \int_A \kappa_h(s - u)du$ es un factor de corrección.

Más detalles sobre los procesos puntuales, las pruebas de hipótesis y estimación de la función de intensidad pueden verse en la Parte III de Cressie, 1992.

Estimación de función de intensidad

A continuación se definen algunos miembros en la familia de kerneles a utilizar para la estimación de la función de intensidad y se mencionan algunas de sus propiedades asintóticas.

Recordemos la forma del estimador de la función de intensidad. Sea $(s_1, s_2, ..., s_n)$ las localizaciones espaciales de $n = N(A)$ eventos en una región acotada de estudio $A \subset \mathbb{R}^d$. El estimador de la función de intensidad es de la forma

$$\hat{\lambda}_h(s) \equiv \frac{1}{p_h(s)} \left\{ \sum_{i=1}^{n} \kappa_h(s - s_i) \right\}, s \in A$$

donde $\kappa_h(\cdot)$ es una función kernel simétrica en el origen, $h > 0$ determina el suavizamiento y $p_h(s) \equiv \int_A \kappa_h(s - u)du$ es un factor de corrección.

En particular en esta tesis emplearemos kerneles Gaussianos, es decir $\kappa_h(\cdot) = \phi(\cdot/h)$ donde ϕ es la función de densidad de una normal estándar (bivariada). La elección del ancho de banda h se escoge usualmente minimizando alguna medida de error esperada, como el error medio cuadrado integrado (*mean integrated squared error*)

$$\text{MISE}(h) = \text{E}\left[\int (\hat{\lambda}_h(x) - \lambda(x))^2 \, dx \right].$$

En el caso univariado en que la densidad subyacente a estimar es normal se sabe que el mínimo del MISE se alcanza en

$$\tilde{h} = \left(\frac{4\hat{\sigma}^5}{3n} \right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5}.$$

con $\hat{\sigma}$ la desviación estándar muestral. Sin embargo, en la práctica este valor es usado como valor provisional siempre que la distribución sea univariada y no diste mucho de la forma de densidad normal.

Algunas propiedades asintóticas sobre los estimadores no paramétricos kernel pueden consultarse en Parzen, 1962 y Nakayama, 2011.

Teorema 1.2.1. *Consistencia puntual de los estimadores kernel*

Supongamos que λ es continua en s, y κ satisface las condiciones

$$\int \kappa(s)ds = 1,$$

y $h = h(n)$ se escoge de modo que

$$\lim_{n\to\infty} h(n) = 0,$$

entonces

$$\text{E}[\hat{\lambda}_n^*(y)] \to \lambda(s)$$

conforme $n \to \infty$, es decir, es puntualmente asintóticamente insesgado.

Si además se tiene que

$$\sup_z |\kappa(z)| < \infty, \int |\kappa(z)|dz < \infty, \lim_{|z|\to\infty} |z\kappa(z)| = 0$$

se tiene que

$$nh_n \mathrm{Var}[\hat{\lambda}_n^*(s)] \to \lambda(s)\rho_\kappa$$

conforme $n \to \infty$ donde $\rho_\kappa = \int \kappa^2(z)dz$.

De estos dos resultados se sigue que

$$\mathrm{E}|\lambda_n(s) - \lambda(s)|^2 \to 0$$

conforme $n \to \infty$, es decir, el estimador es puntualmente consistente.

Usualmente no se tienen formas cerradas para evaluar el MISE, por lo que se recurren a aproximaciones, y tampoco es usual dar con la forma cerrada de un h óptimo. En nuestro caso, como se trata dos dimensional (dos dimensiones espaciales), podemos flexibilizar un poco más la familia de Kerneles para aproximar la función de intensidad con

$$\hat{\lambda}_H(s) \equiv \frac{1}{p_H(s)} \left\{ \sum_{i=1}^{n} \phi(H^{-1/2}(s - s_i)) \right\}, s \in A.$$

La regla de Silverman consiste en usar

$$\sqrt{\mathbf{H}_{ii}} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{-1}{d+4}} \sigma_i$$

donde σ_i es la desviación estándar de la i-ésima variable y $\mathbf{H}_{ij} = 0, i \neq j$. Notemos que la regla de Silverman tiene como caso particular el $\tilde{h}$ óptimo respecto al MISE del caso univariado. En nuestro caso

$$\mathbf{H}^{-1/2} = \begin{bmatrix} n^{\frac{1}{6}}/\hat{\sigma}_1 & 0 \\ 0 & n^{\frac{1}{6}}/\hat{\sigma}_2 \end{bmatrix}.$$

1.3. Regresión kernel (Estimador de Nadaraya-Watson)

Cuando la covarianza entre puntos espaciales no parece poder describirse con un semivariograma paramétrico (ver Sección 1.5.1), una alternativa es usar regresión múltiple no paramétrica. La ventaja es que esta técnica permite analizar datos con la misma estructura que la necesaria en Kriging.

La regresión kernel es una técnica estadística no paramétrica usada para estimar la esperanza condicional de una variable aleatoria. El objetivo es encontrar una relación no lineal

entre un par de variables aleatorias X y Y. La esperanza condicional puede expresarse

$$\mathrm{E}(Y|X) = m(X)$$

donde m es una función desconocida.

Nadaraya y Watson propusieron, en 1964, un estimador de m mediante un promedio localmente ponderado, usando una función kernel como función ponderadora. Así, el estimador es

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{j=1}^n K_h(x - x_j)}$$

donde K_h es el kernel con ancho de banda o ventana h. El denominador es un término de ponderación con suma 1.

1.3.1. Derivación del estimador de Nadaraya-Watson

Notemos que

$$\mathrm{E}(Y|X = x) = \int y f(y|x) dy = \int y \frac{f(y, x)}{f(x)} dy$$

Usando un estimador de tipo kernel (ver 1.2.3) para la distribución conjunta con densidad $f(x, y)$ y la densidad $f(x)$ con un mismo tipo de kernel K, tenemos

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) K_h(y - y_i) \text{ y}$$

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i)$$

tenemos que, por la definición de densidad condicional,

$$\begin{aligned}
\hat{\mathrm{E}}(Y|X = x) &= \int \frac{y \hat{f}(x, y)}{\hat{f}(x)} dy \\
&= \int \frac{y \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) K_h(y - y_i)}{\frac{1}{n} \sum_{i=1}^n K_h(x - x_i)} dy \\
&= \frac{\sum_{i=1}^n K_h(x - x_i) \int y K_h(y - y_i) dy}{\sum_{i=1}^n K_h(x - x_i)} \\
&= \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)},
\end{aligned}$$

que es el estimador de Nadaraya-Watson. Para más detalles puede consultarse Watson, 1964.

1.3.2. Regresión múltiple no paramétrica

Podemos combinar las ideas de regresión del estimador de Nadaraya-Watson con el uso de kernels para estimar densidades multivariadas. Debido a que la justificación del estimador de Nadaraya-Watson se mantiene incluso en el caso donde X es un vector aleatorio podemos extender la forma del estimador de regresión al caso donde X es multivariado.

Así, el estimador es

$$\hat{m}_H(x) = \frac{\sum_{i=1}^{n} K_H(x - x_i) y_i}{\sum_{j=1}^{n} K_H(x - x_j)}$$

En general, es laborioso encontrar una matriz de ancho de banda H óptimo para cada caso. Para nuestra aplicación parece razonable emplear la regla de pulgar de Silverman, $\sqrt{H_{ii}} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{'_1}{d+4}} \sigma_i$ y $H_{ij} = 0, i \neq j$, para el caso multivariado ya que, como vimos en la derivación del estimador de Nadaraya-Watson, este emplea fundamentalmente dos estimadores kernel de densidades para estimar la densidad condicional.

1.4. Campos Aleatorios

Un campo aleatorio o estocástico se define como una colección de variables aleatorias indexadas por un conjunto S cualquiera, por lo que cada variable aleatoria del proceso estocástico está únicamente asociado a un elemento del conjunto. A este conjunto S se le llama conjunto de índices. Dado un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, un campo aleatorio con valores en X es una colección de variables aleatorias que toman valores en X indexados por elementos en el conjunto S. Es decir, un campo aleatorio F es una colección $\{F_s : s \in S\}$ donde cada F_s es una variable aleatoria que toma valores en X. Si S es un intervalo en $\mathbb{R}$ también se le conoce como proceso aleatorio y al conjunto de índices se le suele denotar con T pues es interpretado como el tiempo.

En principio, no podemos asegurar que las variables aleatorias X_s definan un proceso o campo aleatorio en el sentido de éste ser una variable aleatoria en un espacio de probabilidad. Para ello existe el Teorema de extensión de Kolmogorov que garantiza que una colección "consistente" de distribuciones finito-dimensionales define un proceso estocástico. Por simplicidad consideraremos el caso donde S es un intervalo en $\mathbb{R}$ y lo llamaremos T.

Teorema 1.4.1. *Teorema de extensión de Kolmogorov*

Sea T un intervalo, y sea $n \in \mathbb{N}$. Para cada $k \in \mathbb{N}$ y una sucesión finita de tiempos distintos $t_1, \ldots, t_k \in T$, sea $\nu_{t_1 \ldots t_k}$ una medida de probabilidad sobre $(\mathbb{R}^n)^k$. Supongamos que estas medidas satisfacen las siguientes dos condiciones de consistencia:

1. *Para todas las permutaciones π de $\{1, \ldots, k\}$ y conjuntos medibles $F_i \subseteq \mathbb{R}^n$,*

$$\nu_{t_{\pi(1)} \ldots t_{\pi(k)}} \left(F_{\pi(1)} \times \cdots \times F_{\pi(k)}\right) = \nu_{t_1 \ldots t_k} \left(F_1 \times \cdots \times F_k\right);$$

2. *Para todos los conjuntos medibles* $F_i \subseteq \mathbb{R}^n$,$m \in \mathbb{N}$ *:*

$$\nu_{t_1 \dots t_k}(F_1 \times \cdots \times F_k) = \nu_{t_1 \dots t_k, t_{k+1}, \dots, t_{k+m}}\left(F_1 \times \cdots \times F_k \times \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{m}\right).$$

Entonces existe un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ *y un proceso estocástico* $X : T \times \Omega \to \mathbb{R}^n$ *tal que* $\nu_{t_1 \dots t_k}(F_1 \times \cdots \times F_k) = \mathbb{P}(X_{t_1} \in F_1, \dots, X_{t_k} \in F_k)$ *para todo* $t_i \in T$, $k \in \mathbb{N}$ *y conjuntos medibles* $F_i \subseteq \mathbb{R}^n$*, es decir,* X *tiene a* $\nu_{t_1 \dots t_k}$ *como sus distribuciones finito-dimensionales asociadas a los tiempos* $t_1 \dots t_k$*.*

De hecho, siempre es posible tomar el espacio de probabilidad subyacente como $\Omega = (\mathbb{R}^n)^T$ y tomar por X el proceso canónico $X\colon (t, Y) \mapsto Y_t$. Por lo tanto, una forma alternativa de enunciar el teorema de extensión de Kolmogorov es que, suponiendo que las condiciones de consistencia se satisfacen, existe una (única) medida ν sobre $(\mathbb{R}^n)^T$ con marginales $\nu_{t_1 \dots t_k}$ para cualquier colección finita de tiempos $t_1 \dots t_k$. El teorema de extensión de Kolmogorov también aplica cuando T es no numerable, pero el precio a pagar para este grado de generalidad es que la medida ν sólo está definida sobre el producto σ-álgebra de $(\mathbb{R}^n)^T$.

Procesos estacionarios en sentido estricto y amplio

Una propiedad importante que pueden tener los procesos aleatorios es la estacionariedad. Intuitivamente un proceso aleatorio $\{X(t), t \in J\}$ es estacionario si sus propiedades estadísticas no cambian por desplazamientos en el tiempo. Usualmente se habla de dos nociones de estacionariedad, estacionariedad en el sentido estricto y estacionariedad en el sentido amplio.

Hablamos de estacionariedad en el sentido estricto cuando para un proceso aleatorio X_t y $X_{t+\Delta}$ tienen la misma distribución de probabilidad. Este sentido de estacionariedad no es tan usado debido a que no es sencillo probar que un proceso es estacionario en el sentido estricto.

Definición 1. *Un proceso aleatorio continuo en el tiempo* $\{X_t, t \in \mathbb{R}\}$ *es estacionario en el sentido estricto si para todo* $t_1, t_2, \dots, t_r \in \mathbb{R}$ *y para todo* $\Delta \in \mathbb{R}$*, la función de distribución acumulada conjunta de* $X_{t_1}, X_{t_2}, \dots, X_{t_r}$ *es la misma función de distribución acumulada conjunta de* $X_{t_1+\Delta}, X_{t_2+\Delta}, \dots, X_{t_r+\Delta}$*.*

En contraste, una de las formas de estacionariedad más usadas en la práctica es la llamada estacionariedad en sentido amplio. Un proceso aleatorio de dice estacionario en el sentido amplio si su función de media y su función de correlación no cambian por desplazamientos en el tiempo. Más precisamente X_t es estacionario en el sentido amplio si para todo $t_1, t_2 \in \mathbb{R}$ y para todo $\Delta \in \mathbb{R}$,

$$\mathrm{E}[X_{t_1}] = \mathrm{E}[X_{t_2}],$$

$$\mathrm{E}[X_{t_1} X_{t_2}] = \mathrm{E}[X_{t_1+\Delta} X_{t_2+\Delta}].$$

Esta definición es equivalente a:

Definición 2. *Un proceso aleatorio continuo en el tiempo $\{X_t, t \in \mathbb{R}\}$ es estacionario en el sentido amplio si*

1. $\mu_{X_t} = \mu_X$ *para todo* $t \in \mathbb{R}$
2. $C_X(t_1, t_2) = C_X(t_1 - t_2)$ *para todo* $t_2, t_2 \in \mathbb{R}$*, es decir, la función de covarianza sólo depende de la distancia.*

1.4.1. Campos Aleatorios Markovianos

Si bien la propiedad Markoviana puede extenderse a procesos aleatorios continuos para efectos de este trabajo os cencentraremos en la propiedad de Markov para procesos con un conjunto de índices S numerables. En particular si S es equiespaciado $S = \mathbb{Z}^d$. Más aún, usualmente tendremos observaciones para todo $s \in S$, como es el caso de modelos en imágenes y regiones geográficas. Por lo tanto, ni siquiera hace falta pasar por el Teorema 1.4.1 de extensión de Kolmogorov, que sería necesario si el conjunto S indexará una sucesión numerable o un conjunto no numerable de variables aleatorias. Pero al tratarse de una sucesión finita de variables aleatorias se puede tratar simplemente como un vector aleatorio. Por esto mismo, a diferencia de los modelos geoestadísticos y los de procesos puntuales, no interesa solamente características globales del proceso (como la función de media y la función de covarianza) si no que podemos encontrar la distribución conjunta, que llamaremos p, que deberíamos poder construir usando sus distribuciones condicionales (ver Teorema 3.2.3).

Otra diferencia importante respecto a los modelos geoestadísticos y los de procesos puntuales, es que ahora E no se tiene que ser $\mathbb{R}^p$, sino que puede ser $\mathbb{R}^+$ (como un modelo gamma), $\mathbb{N}$ (como un modelo de conteos), categórica o binaria.

Ahora, conviene enunciar una notación muy útil para estos modelos.

$$x_{-A} \equiv \{x_s : s \in S\} \setminus \{x_l : l \in A\}.$$

En particular, respecto cualquier elemento x_i, podemos llamar resto de variables

$$x_{-\{i\}} \equiv x_{-i} \equiv \{x_s : s \in S\} \setminus \{x_i\}.$$

En el caso de datos en una celosía es posible que S este indexado por dos índices i y j, por ejemplo donde la dimensión asociada a i es el espacio y la j esta asociada el tiempo, y podemos emplear la notación

$$x_{-A} \equiv \{x_s : s \in S\} \setminus \{x_{i,j} : (i,j) \in A\}.$$

De modo que, respecto cualquier elemento x_i, podemos llamar resto de variables

$$x_{\{-i,j\}} \equiv x_{-i,j} \equiv \{x_s : s \in S\} \setminus \{x_{i,j}\}.$$

Podríamos decir que la pregunta que entonces surge es, dada una familia $\{\nu_i(\cdot|x_{-i}), i \in S\}$ de distribuciones en E determinadas por las observaciones en x^i, ¿bajo qué condiciones estas distribuciones representan distribuciones condicionales de una distribución conjunta p?

Responder a esta pregunta implica poder especificar completa o parcialmente la distribución conjunta p usando sólo las condicionales.

Notemos que, si $\nu_i(\cdot|x_{-i})$ es sólo localmente dependiente (sólo depende de sus vecinos), la complejidad del modelo se reduciría significativamente. Cuando esto es así, estamos hablando de un campo aleatorio Markoviano. Una pregunta que surge es si dada una estructura condicional existe la distribución conjunta. En las cadenas de Markov no es de mucho interés conocer explíctamente la distribución conjunta, pero fácilmente se puede identificar cuando la conjunta existe. Esto se debe a que los índices del tiempo están dentro de la recta real y dado este orden es fácil ver que son coherentes. Se tiene una variable aleatoria inicial con distribución no condicionada y basta multiplicar su función de densidad (o probabilidad) por las funciones de densidad o probabilidad condicionadas de las variables que siguen.

En este contexto, la propiedad de Markov ya no se refiere a que el estado presente depende del pasado el estado siguiente. Ahora, la propiedad de Markov consiste en que dado los vecinos de una observaciones, esta ya no depende del resto de los estados.

Definición $X_1, X_2, ..., X_n$ es un campo aleatorio Markoviano si existe un grafo $G = (V, E)$ tal que $V = \{X_1, ..., X_n\}$ y $(X_i, X_j) \in E$ si y sólo si

$$p(x_i|\{X_1, ..., X_n\} \setminus \{X_i\}) \neq p(x_i|\{X_1, ..., X_n\} \setminus \{X_i, X_j\})$$

o bien

$$p(x_i|X_s \text{ con } s \neq i) = p(x_i|X_s \text{ tal que } (X_i, X_s) \in E).$$

Definiremos $\mathcal{N}(X_i)$ los vecinos de X_i. Esto es, $\mathcal{N}(X_i) = \{X_j \in V : (X_i, X_j) \in E\}$.

Los vecinos $\mathcal{N}(x_i)$ de un estado puede variar de acuerdo al modelo que empleemos. Si se considera el espacio dividido en una celosía con elementos cuadrados las especificaciones más comunes (en dos dimensiones) son a 4 o 8 vecinos. Entonces

$$\mathcal{N}_4(x_{i,j}) = \{x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1}\}$$

y

$$\mathcal{N}_8(x_{i,j}) = \{x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1}, x_{i-1,j-1}, x_{i+1,j-1}, x_{i+1,j-1}, x_{i+1,j+1}\}$$

En el contexto espacial, podemos hablar de tres propiedades de Markov que, a diferencia del caso de cadenas de Markov, las tres no son equivalentes. La propiedad global es más fuerte que la propiedad local, y a su vez la propiedad local es mas fuerte que la propiedad por pares.

Definición Dado un grafo no dirigido $G = (V, E)$, el conjunto de variables aleatorias $X = (X_v)_{v \in V}$ indexados por V forman un *campo aleatorio de Markov* con respecto a G si satisfacen las propiedades de Markov locales:

- **Propiedad por pares**: Cualesquiera dos variables no adyacentes son condicionalmente independientes dados sus vecinos

$$X_u \perp\!\!\!\perp X_v | X_{\mathcal{N}(u) \cup \mathcal{N}(v)}$$

- **Propiedad local**: Una variable es condicionalmente independiente del resto de variables dados sus vecinos

$$X_u \perp\!\!\!\perp X_{V \setminus \mathcal{N}[v]} | X_{\mathcal{N}(v)}$$

  donde $\mathcal{N}[v] = v \cup \mathcal{N}(v)$ se llama la vecindad cerrada de v.

- **Propiedad global**: Cualesquiera dos subconjuntos de variables son condicionalemnte independientes dado un subconjunto separante (se les llama corte en teoría de grafos)

$$X_A \perp\!\!\!\perp X_B | X_S$$

  donde cada camino de un nodo en A a un nodo en B pasa por S.

Vale la pena destacar que pese a que en los modelos Markovianos suponemos que hay un grafo detrás (siendo este de adyacencias de regiones geográficas o vecindades en una celosía), estos modelos no son tan restrictivos. Por ejemplo, es posible tratar un problema de procesos puntuales construyendo una grafo que enlace dos observaciones si la distancia entre ellas es menor a cierto r fijo. Por supuesto, el grafo puede cambiar drásticamente dependiendo de la r que se elija, pero aun así es una posible aproximación al problema y puede hacerse análisis de sensibilidad a la elección de r (por ejemplo mediante validación cruzada, que es común usarla en parámetros de este estilo).

1.4.2. Campos Aleatorios Gaussianos

Un campo o proceso estocástico se dice Gaussiano si y solo si para todo conjunto finito de $s_1, \ldots, s_n$ en S el vector

$$\mathbf{X}_{s_1,\ldots,s_n} = (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_n})$$

tiene distribución normal multivariada. Es lo mismo decir que toda combinación lineal de $(\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_n})$ tiene distribución normal univariada. Usando la función característica, la propiedad de Gaussianidad puede establecerse como sigue: $\{X_s; s \in S\}$ es Gaussiana si y solo si, para cada conjunto finito de índices $s_1, \ldots, s_n$, hay valores en los reales $\{\sigma_{\ell j}\}$, $\{\mu_\ell\}$ donde $j, \ell = 1, \ldots, n$ con $\sigma_{jj} > 0$ tales que la siguiente igualdad se cumple para todo $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \in \mathbb{R}$

$$\mathrm{E}\left(\exp\left(i \sum_{\ell=1}^{n} \varepsilon_\ell \, \mathbf{X}_{\varepsilon_\ell}\right)\right) = \exp\left(-\frac{1}{2} \sum_{\ell,j} \sigma_{\ell j} \varepsilon_\ell \varepsilon_j + i \sum_{\ell} \mu_\ell \varepsilon_\ell\right)$$

donde i es la unidad imaginaria $\sqrt{-1}$.

Los números $\sigma_{\ell j}$ y μ_ℓ son la covarianza y la media de las variables del proceso. La conveniencia de usar la función caracteristica es que, a diferencia de la función densidad, siempre existe, incluso cuando la matriz de covarianzas es singular pues no aparece en la exponencial la matriz de precisión si no solo los elementos de la matriz de covarianza.

1.4.3. Campos Aleatorios Markovianos Gaussianos

Diremos que un campo aleatorio es Markoviano Gaussiano (GMRF) si cumple ser un campo Markoviano y también ser un campo Gaussiano, es decir, si para todo conjunto finito de $s_1, \ldots, s_n$ en S el vector

$$\mathbf{X}_{s_1,\ldots,s_n} = (\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_n})$$

tiene distribución normal multivariada y a su vez este vector $\mathbf{X}_{s_1,\ldots,s_n}$ tiene asociado un grafo no dirigido $G_{\mathbf{X}_{s_1,\ldots,s_n}} = (V_{\mathbf{X}_{s_1,\ldots,s_n}}, E_{\mathbf{X}_{s_1,\ldots,s_n}})$, cumple alguna de las propiedades de Markov (independencia condicional dados las variables adyacentes si y sólo si las variables no son adyacentes).

Definición 3. *Sea $\boldsymbol{X} = X_1, \ldots, X_n$ un vector aleatorio con distribución Normal(μ, Σ). Sea $G = (\mathcal{V}, E)$ un grafo etiquetado donde el conjunto de vértices ($\mathcal{V} = 1, \ldots, n$ y el conjunto de arcos $\mathcal{E}$ es tal que no hay arco entre i y j si y sólo si $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{-ij}$. Entonces se dice que $\boldsymbol{X}$ es un GMRF con respecto a G.*

En lo subsecuente, para $i, j \in \mathcal{V}$ denotaremos $i \sim j$ si los vértices i y j son vecinos dentro del grafo G. Como la relación es simétrica los grafos no son dirigidos.

Debido al Teorema 3.2.1 sabemos que si $\boldsymbol{X}$ se distribuye Gaussiana $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{-ij}$ si y sólo si la entrada ij en la matriz de precisión $\boldsymbol{Q}$ es cero. Por lo que un campo es Markoviano Gaussiano si para todo conjunto finito de $s_1, \ldots, s_n$ en S el vector

$$\mathbf{X}_{s_1,\ldots,s_n} = (X_{s_1}, \ldots, X_{s_n})$$

con ceros en $\boldsymbol{Q}$ si y solo si hay ceros en la matriz de adyacencia asociada a $G_{\mathbf{X}_{s_1,\ldots,s_n}}$.

En el caso donde S es finito es aun más simple pues básicamente pedimos que el vector X_S tenga ditribución Normal con ceros en $\boldsymbol{Q}$ si y solo si hay ceros en la matriz de adyacencia asociada a $G_{\mathbf{X}_S}$. Este es el caso que trataremos al emplear la metodologia de INLA.

Adicionalmente, podemos definir lo que se conoce como campo aleatorio Markoviano Gaussiano Latente u Oculto. Se trata de un campo aleatorio Markoviano Gaussiano $\boldsymbol{X}$ que no es observable, pero ciertas variables $\boldsymbol{Y}$ que dependen estocásticamente de él si son observables. Se dice que las variables X_s son variables latentes o bien que el campo $\boldsymbol{X}$ es un campo latente. Por ejemplo, si $\boldsymbol{Y}|\boldsymbol{X}$ tiene cierta distribución y $\boldsymbol{X}$ es un campo Markoviano Gaussiano pero sólo observamos $\boldsymbol{y}$.

Más detalles sobre los procesos Markovianos y los procesos Gaussianos pueden verse en la Parte II de Cressie, 1992 y en Rue & Martino, 2007.

1.5. Kriging

La idea básica de Kriging es predecir el valor de una función al calcular el promedio ponderado de todos los valores conocidos de la función en una vecindad del punto de interés. El método guarda cierto parecido al análisis de regresión pues ambos pueden derivar un Mejor Estimador Lineal Insesgado (MELI o BLUE), basado en supuestos sobre la covarianza. Vale la pena destacar que en el caso de Kriging hablamos de predecir una variable aleatoria no observada, por lo que se trata del Mejor Predictor Lineal Insesgado (MPLI o BLUP). Ambos hacen uso del teorema de Gauss Markov para probar la independencia del predictor y su error. Sin embargo, los enfoques son muy distintos. Mientras que los modelos de regresión se basan en múltiples observaciones de un conjunto multivariado de datos, Kriging es usado para estimar una sola realización de un campo aleatorio.

El Kriging también puede interpretarse como un spline en un espacio de Hilbert con núcleo reproductor. Aquí, el núcleo reproductor está dado por la función de covarianza. La diferencia con el enfoque clásico de kriging está en la interpretación. Mientras que la aproximación por splines está motivado por la interpolación de mínima norma en un espacio de Hilbert, el Kriging puede verse bajo el enfoque de inferencia Bayesiana. Se propone una distribución inicial sobre el espacio de funciones que es un proceso gaussiano (ver Apéndice A) con función de covarianza (o kernel) la covarianza que elegimos. Se recoge un conjunto de observaciones, cada uno con su lugar en A. Podemos predecir un nuevo valor en cualquier posición en A al combinar la distribución previa de Proceso Gaussiano con la función de verosimilitud Gaussiana de cada una de las observaciones. La distribución posterior es también es proceso Gaussiano (se trata de un modelo conjugado para la media) con función de media y de covarianza calculable.

En los modelos geoestadísticos, la muestra es interpretada como el resultado de un proceso aleatorio. El hecho de que estos modelos incorpore incertidumbre en su conceptualización no significa que el fenómeno (un bosque un acuífero o un deposito de minerales), haya resultado de un proceso realmente aleatorio , sino que nos permite construir una base metodológica para nuestra inferencia espacial de cantidades en sitios no observados, así como cuantificar la incertidumbre asociada este predictor. Un proceso estocástico en este contexto es sólo una forma de tratar con la muestra. Buscamos el proceso estocástico que mejor describa lo observado.

El valor observado asociado al lugar s_i (usualmente una coordenada geográfica) se interpreta como una realización de la variable aleatoria X_{s_i}. En el espacio S, donde la muestra está dispersa, hay n realizaciones de las variables aleatorias $X_{s_1}, X_{s_2}, \ldots, X_{s_n}$, correlacionadas entre ellas.

El conjunto de variables aleatorias constituye una función aleatoria de la que solo conocemos el conjunto $\{X_{s_i}\}$. Con sólo una realización de cada variable aleatoria no es posible hacer inferencia de los parámetros de cada una de las variables aleatorias, o de la función aleatoria.

La solución propuesta, en el enfoque geoestádistico, consiste en asumir varios grados de estacionariedad en la función aleatoria. Esto permite hacer inferencia de algunas cantidades. Asumiendo, por ejemplo, la hipótesis de que la media de todas las variables es

la misma, se asume también que es posible estimarla mediante la media aritmética de los valores observados.

Para hacer predicción de X_s para cualquier $s \in S \subset \mathbb{R}^2$ asumiremos que X es un campo aleatorio Gaussiano (ver Sección 1.4.2). Buscamos que estos predictores dependan linealmente de los x_s observados.

1.5.1. Variograma, semivariograma y función de covarianza

El variograma se define como la varianza de la diferencia entre los valores (en el campo) en dos lugares

$$2\gamma(s_1, s_2) \equiv \text{Var}\left(X_{s_1} - X_{s_2}\right) = E\left[\left(\left(X_{s_1} - \mu(s_1)\right) - \left(X_{s_2} - \mu(s_2)\right)\right)^2\right].$$

Adicionalmente, a la función $\gamma(s_1, s_2)$ la llamamos semivariograma. Mientras que la función de covarianza se define como la covarianza entre los valores (en el campo) en dos lugares

$$C(s_1, s_2) \equiv \text{Cov}(X_{s_1}, X_{s_2}).$$

La hipótesis de homogeneidad en la varianza se fija al suponer que la correlación entre cualesquiera dos variables aleatorias solamente depende de la distancia espacial entre ellas y es independiente de su localización u orientación (proceso isotrópico). Si $\mathbf{h} = s_1 - s_2$ y $|\mathbf{h}| = h$ la función semivariograma sólo depende de la distancia h

$$\gamma(h) \equiv \gamma(s_1, s_1 + \mathbf{h}).$$

así como para la función de covarianza

$$C(h) \equiv C(s_1, s_1 + \mathbf{h})$$

Si la función de covarianza asociada a un proceso estacionario en sentido amplio (media y varianza) existe se relaciona con el variograma de esta forma

$$2\gamma(s_1, s_2) = C(s_1, s_1) + C(s_2, s_2) - 2C(s_1, s_2).$$

Esta hipótesis nos permite inferir el semivariograma y la función de covarianza mediante los llamados semivariograma empírico y función de covarianza empírica, respectivamente,

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \left(X_{s_i} - X_{s_j}\right)^2 \tag{1.1}$$

$$\hat{C}(h) = \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} \left(X_{s_i} - m(h)\right)\left(X_{s_j} - m(h)\right) \tag{1.2}$$

donde

$$m(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \left[X_{s_i} + X_{s_j}\right].$$

Aquí $N(h)$ denota el conjunto de pares de observaciones i, j tales que $|s_i - s_j| = h$, y $|N(h)|$ es el número de pares en el conjunto. En este conjunto, (i, j) y (j, i) denota al mismo elemento. Generalmente una "distancia aproximada" h se usa, empleando cierta tolerancia.

El variograma empírico no puede calcularse en toda distancia h y debido a la variación en la estimación, no es seguro que se obtenga un variograma válido. Sin embargo, algunos métodos geoestadísticos requieren semivariogramas válidos. Por esto, en las aplicaciones los variogramas empíricos con aproximados por una función modelo.

El variograma es una herramienta útil no solo para definir el modelo espacial a ajustar a los datos sino que puede visualizarse para entender mejor la relación espacial del campo aleatorio. Además, existen 3 cantidades de interés asociadas al semivariograma $\gamma(h)$ que tienen interpretaciones muy claras e importantes.

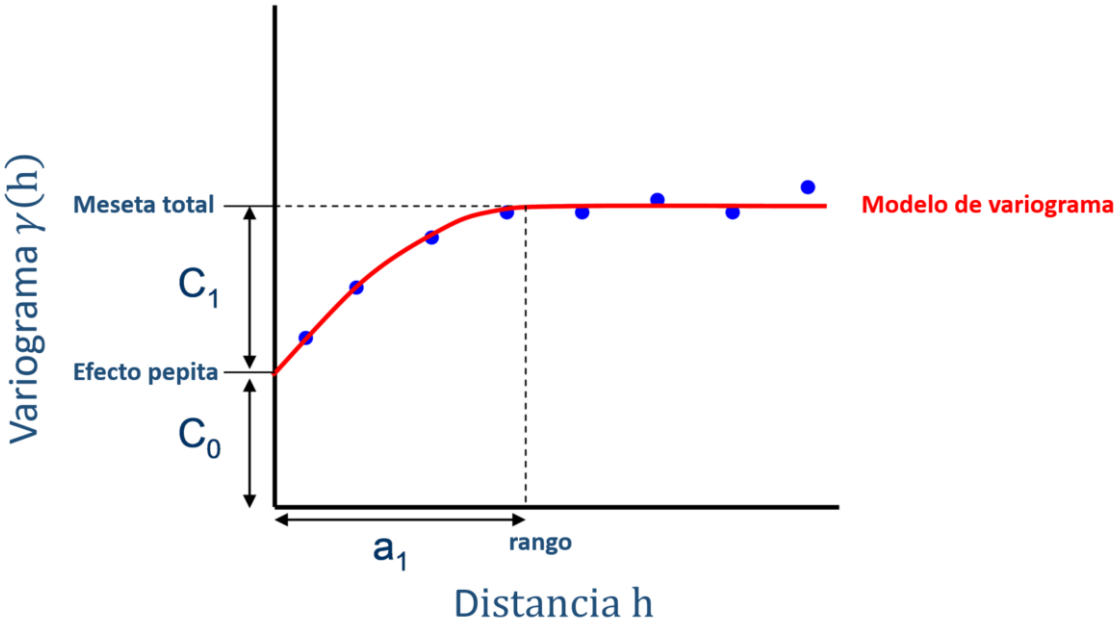


Figura 1.1: Esquema de elementos del semivariograma

- **Efecto pepita (Nugget)**

  Representa la variabilidad a micro escala. Se define $\lim_{h\to 0} \gamma(h)$.

- **Meseta o silo (Sill)**

  Se define $\lim_{h\to \infty} \gamma(h)$, y representa la varianza del campo aleatorio.

- **Rango (Range)**

  Se define como la distancia en la que el variograma alcanza el silo. En el caso de que el variograma se aproxime asintóticamete al silo, convencionalmente se considera la distancia en la que el variograma llega al $95\,\%$ del silo.

Estas cantidades son tan importantes que es usual proponer modelos de semivariograma que tienen por parámetros estas cantidades. Algunos de los modelos más usados en la práctica son:

- Esférica
$$\gamma(h) = \begin{cases} c_0 + c[1.5(\frac{h}{a}) - 0.5(\frac{h}{a})^3], & \text{si } h < a \\ c_0 + c, & \text{si } h > a \end{cases}$$

- Exponencial
$$\gamma(h) = c_0 + c(1 - e^{\frac{-h}{b}})$$

- Gaussiana
$$\gamma(h) = \begin{cases} c_0 + c(1 - e^{\frac{-h^2}{a^2}}), & \text{si } h < a \\ c_0 + c, & \text{si } h > a \end{cases}$$

- Lineal
$$\gamma(h) = \begin{cases} c_0 + c\left(\frac{h}{a}\right), & \text{si } h < a \\ c_0 + c, & \text{si } h > a \end{cases}$$

- Circular
$$\gamma(h) = \begin{cases} c_0 + c\left(1 - \frac{2}{\pi}\cos^{-1}\left(\frac{h}{a}\right) + \sqrt{1 - \frac{h^2}{a^2}}\right), & \text{si } h < a \\ c_0 + c, & \text{si } h > a \end{cases}$$

  Donde c_0 es el efecto "nugget", a es el rango, $c_0 + c$ es la semivarianza asintótica o meseta y h es la distancia de separación. En el caso exponencial b es un parámetro de decaimiento.

Un ejemplo interesante en $\mathbb{R}$ es cuando $C(h) = c\exp(-\lambda|h|)$. Ya que este modelo se acerca a la meceta de forma asintótica, el rango práctico usualmente se toma como la distancia en la que la semivarianza alcanza el 95 % de la meceta. Por la ecuación que relaciona función de covarianza con semivariograma tenemos

$$\begin{aligned} \gamma(s_1, s_2) &= \tfrac{1}{2}[C(s_1, s_1) + C(s_2, s_2) - 2C(s_1, s_2)] \\ &= \tfrac{1}{2}\left[c + c - 2c\exp(-\lambda|h|)\right] \\ &= c(1 - \exp(-\lambda|h|)). \end{aligned}$$

Esto coincide con el caso exponencial con $c_0 = 0$. Un proceso Gaussiano aleatorio con este covariograma exponencial se le llama Proceso Gaussiano Markoviano. Supongamos

que las observaciones están en $s_i = i, i = 1, \ldots, n$. Siendo $\rho = \exp(-\lambda)$ puede mostrarse que la matriz de covarianza coincide con la matriz de varianza de una caminata aleatoria de orden 1, es decir, con

$$\boldsymbol{\Sigma} = c \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{t-1} \\ \rho & 1 & \rho & \ldots & \rho^{t-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{t-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \ldots & 1 \end{pmatrix}$$

y

$$\boldsymbol{Q} \equiv \boldsymbol{\Sigma}^{-1} = \frac{1}{c(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & \ldots & 0 \\ -\rho & 1+\rho^2 & -\rho & \ldots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \ldots & -\rho & 1+\rho^2 & -\rho \\ 0 & \ldots & 0 & -\rho & 1 \end{pmatrix}$$

1.5.2. Inferencia de las ponderaciones

La predicción de la cantidad X_{s_0}, en un lugar s_0 donde no se tiene valor observado x_{s_0}, se calcula mediante una combinación lineal de los valores en efecto observados $\{X_{s_i} = x_i\}_{i+1}^{N}$ y ponderación $w_i(x_0),\ i = 1, \ldots, N$ tal que

$$\hat{X}_{s_0} = \begin{bmatrix} w_1 & w_2 & \cdots & w_N \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \sum_{i=1}^{N} w_i(x_0) \times X_{s_i}$$

La ponderación $\{w_i\}$ debe cumplir dos funciones importantes para la inferencia espacial

- Refleja la proximidad de las cantidades observadas al lugar s_0.
- Por otro lado, debe tener un efecto desagregativo para evitar el sesgo debido a un posible conglomerado de observaciones.

Al calcular los ponderadores $\{w_i\}$, se tienen dos objetivos, insesgamiento y mínima varianza de la predicción.

Si graficamos en un diagrama de dispersión los valores reales X_{s_0} contra los valores estimados $\hat{X}_{s_0}$, el criterio de insesgadez global ("estacionariedad intrínseca" o estacionariedad del proceso del campo) implica que la media de las estimaciones debe ser igual a la media de los valores reales.

El segundo criterio indica que la media del cuadrado de las desviaciones $(\hat{X}(s) - X(s))$ debe ser mínima, lo que significa que la nube de valores estimados graficados contra la nube de valores reales es más dispersa, el predictor es más impreciso.

Dependiendo de las propiedades estocásticas del campo aleatorio y los varios grados de estacionariedad asumida, distintos métodos de calculo de las ponderaciones pueden ser deducidos, es decir, distintos tipos de Kriging aplican.

En este trabajo, veremos dos de los métodos clásicos. Además, el Kriging ordinario es el que usaremos en nuestra aplicación con datos reales.

- Kriging ordinario supone una media constante desconocida solo sobre una vecindad de s_0.

- Kriging simple asume estacionariedad de la media sobre todo A con una media conocida $\mathrm{E}\{X(s)\} = \mathrm{E}\{X_{s_0}\} = m$, donde m es la media conocida.

1.5.3. Kriging ordinario

El valor desconocido X_{s_0} es interpretado como una variable aleatoria localizada en s_0, así como los valores de las observaciones vecinas $X_{s_i}, i = 1, \ldots, n$. El predictor $\hat{X}_{s_0}$ se interpreta también como una variable aleatoria localizada en s_0, ya que es combinación lineal de variables aleatorias.

Para deducir el sistema de Kriging para estos supuestos, el error al que se incurre al estimar X_s en s_0 se fija

$$\begin{aligned}
\varepsilon(s_0) &= \hat{X}_{s_0} - X_{s_0} \\
&= \left[\sum_{i=1}^{N} w_i(x_0) \times X_{s_i}\right] - X_{s_0} \\
&= \begin{bmatrix} W^T & -1 \end{bmatrix} \cdot \begin{bmatrix} X_{s_1} & \cdots & X_{s_n} & X_{s_0} \end{bmatrix}^T .
\end{aligned}$$

Los dos criterios de calidad pueden expresarse en términos de la media y la varianza de la nueva variable aleatoria $\varepsilon(s_0)$:

Insesgadez

Ya que la función aleatoria es estacionaria, $\mathrm{E}(X_{s_i}) = \mathrm{E}(X_{s_0}) \equiv m$, las siguientes restricciones se siguen

$$\begin{aligned}
\mathrm{E}\left(\varepsilon(x_0)\right) = 0 \Leftrightarrow & \left[\sum_{i=1}^{n} w_i(x_0) \times \mathrm{E}(X_{s_i})\right] - \mathrm{E}(X_{s_0}) = 0 \\
\Leftrightarrow & \, m \sum_{i=1}^{n} w_i(x_0) - m = 0 \\
\Leftrightarrow & \sum_{i=1}^{n} w_i(x_0) = 1 \\
\Leftrightarrow & \, \mathbf{1}^T \cdot W = 1
\end{aligned}$$

Por lo tanto, para asegurar que el modelo es insesgado, los ponderadores deben sumar 1.

Varianza mínima

Dos predictores pueden tener $\mathrm{E}\left[\varepsilon(s_0)\right] = 0$, pero la dispersión alrededor de la media determina la diferencia entre la calidad de los predictores. Para encontrar un predictor de mínima varianza, se requiere minimizar $\mathrm{E}\left(\varepsilon(s_0)^2\right)$.

$$\begin{aligned}\mathrm{Var}(\varepsilon(s_0)) &= \mathrm{Var}\left(\begin{bmatrix} W^T & -1 \end{bmatrix} \cdot \begin{bmatrix} X_{s_1} & \cdots & X_{s_n} & X_{s_0} \end{bmatrix}^T\right) \\ &= \begin{bmatrix} W^T & -1 \end{bmatrix} \cdot \mathrm{Var}\left(\begin{bmatrix} X_{s_1} & \cdots & X_{s_n} & X_{s_0} \end{bmatrix}^T\right) \cdot \begin{bmatrix} W \\ -1 \end{bmatrix}\end{aligned}$$

$$\mathrm{Var}(\varepsilon(s_0)) = \begin{bmatrix} W^T & -1 \end{bmatrix} \cdot \begin{bmatrix} \mathrm{Var}_{s_i} & \mathrm{Cov}_{s_i s_0} \\ \mathrm{Cov}^T_{s_i s_0} & \mathrm{Var}_{s_0} \end{bmatrix} \cdot \begin{bmatrix} W \\ -1 \end{bmatrix}$$

donde $\mathrm{Var}_{s_0} = \mathrm{Var}(X_{s_0})$, $\mathrm{Cov}_{s_i s_0} = \mathrm{Cov}\left(\begin{bmatrix} X_{s_1} & \cdots & X_{s_n} \end{bmatrix}^T, X_{s_0}\right)$

y $\mathrm{Var}_{s_i} = \mathrm{Var}\left(\begin{bmatrix} X_{s_1} & \cdots & X_{s_n} \end{bmatrix}^T\right)$.

Una vez definido el modelo de covarianza o variograma, $C(\mathbf{h})$ o $\gamma(\mathbf{h})$, valido en todo el campo de análisis de $X(s)$, podemos escribir una expresión para la predicción de la varianza de cualquier predictor en función de la covarianza entre observaciones y las covarianzas entre observaciones y el punto a estimar

$$\begin{aligned}\mathrm{Var}(\varepsilon(s_0)) &= W^T \cdot \mathrm{Var}_{s_i} \cdot W - \mathrm{Cov}^T_{s_i s_0} \cdot W - W^T \cdot \mathrm{Cov}_{s_i s_0} + \mathrm{Var}_{s_0} \\ &= \mathrm{Cov}(0) + \sum_i \sum_j w_i w_j C(s_i, s_j) - 2 \sum_i w_i C(s_i, s_0)\end{aligned}$$

Algunas conclusiones pueden ser afirmadas de esta expresión. La varianza de la predicción

- crece al crecer la covarianza entre observaciones y decrece al aumentar la covarianza entre las observaciones y el punto a predecir. Esto significa que cuando las observaciones están lejos de s_0 la predicción es peor.
- crece junto con la varianza inicial $C(0)$ de la variable X_{s_0}. Cuando la variable es menos dispersa, la varianza es menor para cualquier punto de A.
- no depende de los valores de las observaciones, sólo a través del modelo de covarianza o del variograma.

Esto significa que la misma configuración espacial (con las mismas relaciones geométricas entre observaciones y el punto a estimar) siempre produce la misma varianza de predicción en cualquier parte del área S. Así, la varianza no cuantifica la incertidumbre de predicción producida por la variables locales observadas, sino la incertidumbre considerando las propiedades del proceso aleatorio.

Para obtener el predictor insesgado de mínima varianza debemos resover el siguiente problema de optimización con restricción:

$$\begin{aligned} \min_{W} \quad & \left\{W^T \cdot \mathrm{Var}_{s_i} \cdot W - \mathrm{Cov}_{s_i s_0}^T \cdot W - W^T \cdot \mathrm{Cov}_{s_i s_0} + \mathrm{Var}_{s_0}\right\} \\ \text{sujeto a} \quad & \mathbf{1}^T \cdot W = 1 \end{aligned}$$

Al resolver este problema de optimización obtenemos el sistema de kriging

$$\begin{bmatrix} \hat{W} \\ \mu \end{bmatrix} = \begin{bmatrix} \mathrm{Var}_{s_i} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathrm{Cov}_{s_i s_0} \\ 1 \end{bmatrix} = \begin{bmatrix} \gamma(s_1, s_1) & \cdots & \gamma(s_1, s_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n, s_1) & \cdots & \gamma(s_n, s_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(s_1, s^*) \\ \vdots \\ \gamma(s_n, s^*) \\ 1 \end{bmatrix}.$$

El parámetro adicional μ es un multiplicador de Lagrange que se usa en la minimización del error de Kriging $\sigma_k^2(x)$ para satisfacer la condición de insesgadez.

1.5.4. Kriging simple

Kriging simple es el más simple, matemáticamente, y también el menos general. Asume que el valor esperado del campo aleatorio es conocido, y depende de una función de covarianza. Sin embargo, en la mayoría de las aplicaciones ni la esperanza ni la función de covarianza son conocidas de antemano.

Los supuestos prácticos de Kriging simple son

- Estacionariedad en el sentido amplio de proceso del campo (Igualdad de media y varianza en el proceso).
- La esperanza es 0 en todo S: $\mu(s) = 0$.
- Función de varianza conocida $C(s_1, s_2) = \mathrm{Cov}(X_{s_1}, X_{s_2})$

Las ponderaciones de Kriging simple no tienen condición de insesgadez y solo deben cumplir el siguiente sistema de ecuaciones

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} C(s_1, s_1) & \cdots & C(s_1, s_n) \\ \vdots & \ddots & \vdots \\ C(s_n, s_1) & \cdots & C(s_n, s_n) \end{pmatrix}^{-1} \begin{pmatrix} C(s_1, s_0) \\ \vdots \\ C(s_n, s_0) \end{pmatrix}$$

Esto es análogo a hacer regresión lineal sobre X_{s_0} respecto a $x_1, \ldots, x_n$.

La interpolación vía Kriging simple esta dada por

$$\hat{X}_{s_0} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}^T \begin{pmatrix} C(s_1,s_1) & \cdots & C(s_1,s_n) \\ \vdots & \ddots & \vdots \\ C(s_n,s_1) & \cdots & C(s_n,s_n) \end{pmatrix}^{-1} \begin{pmatrix} C(s_1,x_0) \\ \vdots \\ C(s_n,x_0) \end{pmatrix}$$

El error esta dado por

$$\text{Var}\left(\hat{X}_{s_0} - X_{s_0}\right) = \underbrace{C(s_0,s_0)}_{\text{Var}(X_{s_0})} - \underbrace{\begin{pmatrix} C(s_1,s_0) \\ \vdots \\ C(s_n,s_0) \end{pmatrix}^T \begin{pmatrix} C(s_1,s_1) & \cdots & C(s_1,s_n) \\ \vdots & \ddots & \vdots \\ C(s_n,s_1) & \cdots & C(s_n,s_n) \end{pmatrix}^{-1} \begin{pmatrix} C(s_1,s_0) \\ \vdots \\ C(s_n,s_0) \end{pmatrix}}_{\text{Var}(\hat{X}_{s_0})}$$

lo que conduce a la versión de mínimos cuadrados generalizada del teorema de Gauss-Markov

$$\text{Var}(X_{s_0}) = \text{Var}(\hat{X}_{s_0}) + \text{Var}\left(\hat{X}_{s_0} - X_{s_0}\right).$$

1.5.5. Inferencia de los parámetros de la función de covarianza o semivariograma

El método más común para hacer la inferencia de los parámetros del semivariograma teórico o de la función de covarianza consiste en un ajuste por mínimos cuadrados, donde se minimiza el error o distancia entre el semivariograma teórico $\gamma(h;\theta)$ contra el semivariograma empírico $\hat{\gamma}(h)$ o la función de covarianza teórica $\gamma(h;\theta)$ contra la función de covarianza empírica $\hat{C}(h)$, respectivamente. Recordando la forma de $\hat{\gamma}(h)$ en (1.1) y $\hat{C}(h)$ en (1.2) en los estimadores de los parámetros θ son

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1,\ldots,m} (\gamma(h_i;\theta) - \hat{\gamma}(h_i))^2$$

o bien

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1,\ldots,m} (C(h_i;\theta) - \hat{C}(h_i))^2$$

donde h_i, $i = 1, ..., m$ son los nodos de h para calcular los estimadores empíricos del semivariograma o función de covarianza.

1.5.6. Propiedades de Kriging

- La predicción por Kriging es insesgada

$$E[\hat{X}_{s_i}] = E[X_{s_i}]$$

- Esta versión de la predicción por Kriging confía plenamente en la exactitud de las observaciones $\hat{X}_{s_i} = x_{s_i}$ (no se asume error de medición), por eso se le considera como un caso de interpolación.
- La predicción por Kriging $\hat{X}_s$ es el MPLI de X_s si los supuestos se cumplen. Sin embargo
  - Como cualquier método, puede ser muy malo si los supuestos no se cumplen.
  - Predictores no lineales o no insesgados pueden ser mejores.
  - No se garantizan las propiedades si se usa un variograma incorrecto. Sin embargo, usualmente aun así se produce una buena interpolación.
  - Que sea el MPLI, no significa necesariamente que sea siquiera bueno, por ejemplo, si no hay dependencia espacial la interpolación por kriging es tan buena como la media aritmética.
  - Kriging da σ_k^2 como una medida de la precisión. Pero, esta medida depende de que el variograma usado sea el correcto.
- Existen resultados asintóticos para estos predictores lineales. Stein distingue dos sentidos de resultado asintótico
  - Cuando la región de observación crece con el número de observaciones de tal modo que la distancia entre observaciones vecinas se mantenga constante (de forma aproximada).
  - Cuando se incrementan las observaciones en una región fija y acotada.

  El segundo enfoque, que es el que Stein sigue, encuentra que existen diferencias entre el comportamiento de las interpolaciones (predicciones en lugares rodeados"por observaciones) y extrapolaciones (predicciones en lugares que salen del rango de observaciones).

Más detalles sobre variogramas, Kriging ordinario y Kriging simple en la Parte I de Cressie, 1992.

CAPÍTULO 2

Preeliminares de INLA

2.1. Modelos jerárquicos

Los modelos jerárquicos son modelos estadísticos escritos de forma condicional en múltiples niveles (forma jerárquica). Los parámetros en niveles inferiores son realizaciones no observadas de variables aleatorias cuya distribución es especificada por los parámetros de los niveles superiores.

Un modelo jerárquico (de tres niveles) usualmente tiene la siguiente estructura

Nivel I. Observaciones

$$p(x|\theta) = p(x_1, ..., x_k|\theta_1, ..., \theta_k) = \prod_{i=1}^{k} p(x_i|\theta_i).$$

Nivel II. Parámetros/ Variables latentes u observaciones perdidas

$$p(\theta; \phi) = p(\theta_1, ..., \theta_k; \phi).$$

Nivel III. Hiperparámetros

$$\phi$$

Es usual emplear esquemas de este tipo de modelos mediante grafos dirigidos. Aquí los vértices son los parámetros y variables aleatorias donde las aristas dirigidas representan la distribución condicional de la variable aleatoria en el extremo final dado el parámetro en el extremo inicial (Ver Figura 2.1).

En este caso ϕ juega el papel de hiperparámetro, las θ_i son variables aleatorias y también son parámetros de la distribución de las x_i que son las observaciones.

Una de las cualidades de estos modelos es que son muy flexibles y permiten incorporar componentes (por ejemplo, espacial y temporal) de forma sencilla. Además, esta formu-

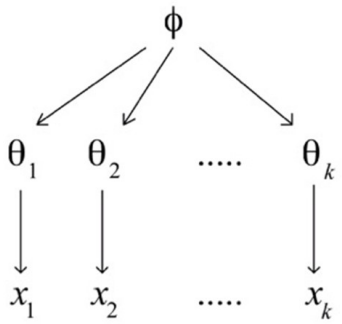


Figura 2.1: Esquema básico de modelos jerárquicos

lación jerárquica permite juntar la información de distintas unidades (estados, municipios, etc.) en un mismo modelo sin dejar de considerar que los datos vienen de unidades diferentes. Las observaciones individuales están asociadas a alguna unidad y a su vez todas las unidades se rigen por una ley probabilística común.

A continuación expondremos una versión de modelo jerárquico de tres niveles donde la asociación de variables latentes e hiperperámetros es lineal, todos con distribución Normal. La utilidad de esta formulación es que tanto la distribución marginal posterior $\alpha|y$ como la condicional posterior $\beta|\alpha, y$ se conocen explícitamente (ver Gutiérrez-Peña, 1998). Para esto asume que las matrices de covarianza Σ_y y Σ_β son conocidas, pero puede dejarse unos factores multiplicativos σ_y^2 y σ_β^2 libres de modo que se hace la inferencia a través de un Muestreador de Gibbs.

Ejemplo 2.1.1. *Versión general de modelo jerárquico lineal Gaussiano*

$$\begin{aligned} Y &= X\beta + \varepsilon, \quad \varepsilon \sim \textit{Normal}_n(0, \Sigma_y); \\ \beta &= H\alpha + \omega, \quad \omega \sim \textit{Normal}_p(0, \Sigma_\beta); \\ \alpha &\sim \textit{Normal}_q(\alpha_0, \Sigma_\alpha). \end{aligned}$$

En este caso X y H son covariables; X se usa para modelar Y, mientras que H se usa para modelar β. Es común usar X para codificar las observaciones por individuos, tiempo o estratos así como sus respectivas características; mientras que H se usa para codificar y agregar las características de las unidades en el nivel superior de agregación (escuelas, hospitales, distritos, etc.). Para mostrar que este tipo de modelos puede abarcar muchos casos particulares destacamos que, por ejemplo, este modelo jerárquico lineal bajo ciertas especificaciones corresponde a un modelo de efectos aleatorios simple.

Si suponemos $p = n = k$ y $q = 1$, y además

Nivel I Sean $\boldsymbol{y} = (y_1, ..., y_k)^T$, $\boldsymbol{X} = \mathbb{I}_k$ y $\Sigma_y = \sigma_y^2 \mathbb{I}_k$.

Nivel II Sean $\boldsymbol{H} = \mathbb{1}_k = (1, ..., 1)^T$ y $\Sigma_\beta = \sigma_\beta^2 \mathbb{I}_k$ y $\alpha \in \mathbb{R}$.

Nivel III Sean $\alpha_0 \in \mathbb{R}$ y $\Sigma_\alpha = \sigma_\alpha^2$.

El modelo general se corresponde a un modelo de efectos aleatorios donde la varianza de las observaciones σ_y^2 y de los efectos aleatorios σ_β^2 se suponen conocidas, con una distri-

bución previa conjugada Normal$(\alpha_0, \sigma_\alpha^2)$ para la media del efecto aleatorio α (que en este caso párticular es equivalente a un intercepto aleatorio), es decir,

$$Y_i|\boldsymbol{X}, \beta \overset{\text{i.i.d.}}{\sim} \text{Normal}(\beta_i, \sigma_y^2) \text{ con } i = 1, ..., k;$$

$$\beta_i|\boldsymbol{H}, \alpha \overset{\text{i.i.d.}}{\sim} \text{Normal}(\alpha, \sigma_\beta^2) \text{ con } i = 1, ..., k;$$

$$\alpha \sim \text{Normal}(\alpha_0, \sigma_\alpha^2).$$

Notemos que la formulación de modelos de forma jerárquica no está comprometida con un enfoque de inferencia en particular. En nuestro ejemplo anterior, lo único Bayesiano es la asignación de una previa a α. Desde el enfoque frecuentista, pueden encontrarse los estimadores máximo verosímil a través del algoritmo EM. Los modelos jerárquicos Bayesianos además asignan distribuciones iniciales a los parámetros del modelo. Debido a la estructura condicional del modelo, bajo el enfoque Bayesiano resulta muy conveniente el uso del muestreador de Gibbs (MCMC) para hacer la inferencia de los parámetros (aproximar las distribuciones posteriores) pues la formulación de los pasos iterativos aprovecha que no todos los parámetros dependen directamente entre sí (esto es, cualquier parámetro condicionado a otros ciertos parámetros es independiente a los restantes).

Sin embargo, debido a la enorme cantidad de parámetros por estimar (recordemos que cada cantidad desconocida implica una simulación por elemento de la cadena en el muestreador de Gibbs) la generación de cadenas de elementos de la muestra puede ser extremadamente costosa en términos de poder computacional.

Implementar el algoritmo de muestreador de Gibbs involucra simular cada variable latente o parámetro desconocido. Esto se puede hacer calculando la densidad condicional completa de cada variable y simularlas una por una de forma iterativa. También, esto puede hacerse simulando por bloques, por ejemplo, si algunas variables tienen una distribución condicional Gaussiana multivariada. El problema de emplear MCMC, en especial el muestreador de Gibbs, es que obtener una sola observación de todas las variables puede involucrar muchas simulaciones y, al estar posiblemente muy correlacionadas las variables, la muestra efectiva resultante (que debe ser suficientemente pseudo-independiente) puede ser muy reducida, cuya corrección requeriría cadenas más largas.

Afortunadamente, la independencia condicional también puede aprovecharse al hacer una aproximación de Laplace de las distribuciones marginales posteriores. Mientras que la independencia condicional de las distribuciones posteriores conduce a distribuciones condicionales completas que sólo dependen de algunos parámetros, las matrices de precisión posteriores resultan tener ceros cuando dos variables (parámetros) son condicionalmente independientes. La llamada aproximación de Laplace puede obtenerse aplicando de forma específica una aproximación Gaussiana, que no es más que emplear sólo la información del vector de medias y matriz de precisión (o covarianzas) del vector aleatorio cuya densidad se desea aproximar.

El uso de aproximaciones analíticas, como la aproximación de Laplace y la aproximación Gaussiana, no requiere un gran número de iteraciones para asegurar que se tiene una buena aproximación de la densidad. Sin embargo, el error de aproximación estará determi-

nado por el tamaño de muestra y no puede mejorarse aumentando el tamaño de la cadena. Este hecho implica tiempos de ejecución considerablemente menores usando aproximaciones analíticas respecto a los tiempos de ejecución de las aproximaciones vía MCMC, si bien estas, con suficiente poder y tiempo computacional, tendrían errores de aproximación menores.

Para ver más casos de modelos jerárquicos en diversas aplicaciones estadísticas se recomienda ver Skrondal & Rabe-Hesketh (2004).

2.1.1. Modelos jerárquicos con componente espacial y temporal

Primero, definimos $\boldsymbol{y} = (y_1, ..., y_n)$ el vector de variables de respuesta observadas, cuya distribución es miembro de la familia exponencial (generalmente), y la media μ_i (para la observación y_i) está convenientemente ligado a un predictor lineal η_i usando una función liga apropiada (también es posible ligar el predictor a un cuantil). Este predictor lineal puede incluir coeficientes en covariables (efectos fijos) y distintos tipos de efectos aleatorios

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(j)}(u_{ki}) + \varepsilon_i \tag{2.1}$$

donde α es el intercepto, $\beta_j, j = 1, ..., n_\beta$ los coeficientes para algunas covariables $\boldsymbol{z}$, las funciones $f^{(j)}$ definen n_f efectos aleatorios sobre covariables $\boldsymbol{u}$. Por último, cada ε_i es un término de error.

El vector de todos los efectos latentes se denota con el vector $\boldsymbol{x}$ e incluye el predictor lineal, coeficientes asociados a las covariables, es decir,

$$\boldsymbol{x} = (\alpha, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\eta}).$$

Además de depender de $\boldsymbol{x}$, la distribución de $\boldsymbol{y}$ dependerá de un vector de hiperparámetros $\boldsymbol{\theta}_1$. Cuáles son estos $\boldsymbol{\theta}_1$ depende del modelo que eligamos para las variables de respuesta $\boldsymbol{y}|\boldsymbol{x}$.

La distribución de los efectos latentes $\boldsymbol{x}$ es un Campo Aleatorio Gaussiano Markoviano (GMRF) (ver Sección 1.4.3) con media cero y matriz de precisión $Q(\boldsymbol{\theta}_2)$, donde $\boldsymbol{\theta}_2$ es un vector de hiperparámetros. Denotamos el vector de todos los hiperparámetros del modelo como $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Mencionamos que es posible ligar el predictor η_i a cuantiles, pero es más común ligarlo a la media. Esto se hace a través de una función liga de modo que si $\mu_i = \mathrm{E}(y_i|x_i)$, entonces $\eta_i = g(\mu_i)$ con $g(\cdot)$ es la llamada función liga.

En este esquema de efectos aleatorios es fácil introducir el efecto espacial, el efecto temporal así como un efecto que considere una interacción entre espacio y tiempo. Una especificación posible es

$$\eta_{it} = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \upsilon_i + \nu_i + \gamma_t + \phi_t + \delta_{it},$$

donde υ_i es el efecto espacial local asociado a los vecinos, ν_i el efecto espacial asociado al estado propio, γ_t es el efecto temporal asociado a los tiempos vecinos (o pasados), ϕ_t es el efecto temporal asociado al tiempo propio y δ_{it} el efecto de interacción espacial-temporal. En este caso, respecto a la Ecuación (2.1), $\boldsymbol{u} = (\boldsymbol{\upsilon}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\delta})$ y $f^{(j)}(\cdot) = I(\cdot)$ es la función identidad para todo j. Respecto al efecto espacial, podemos proponer una especificación Besag-York-Mollie (BYM) para υ_i que es el residual estructurado espacialmente modelado con estructura autorregresiva condicional intrínseca (iCAR)

$$\upsilon_i | \upsilon_{j \neq i} \sim \text{Normal}(m_i, s_i^2),$$

donde

$$m_i = \frac{\sum_{j \in \mathcal{N}(i)} \upsilon_j}{\#\mathcal{N}(i)}, \; s_i^2 = \frac{\sigma_v^2}{\#\mathcal{N}(i)},$$

$\mathcal{N}(i)$ son las áreas vecinas del área i y $\#\mathcal{N}(i)$ es el número de vecinos del área i.

El efecto espacial propio del área i, ν_i, se modela con un vector $\boldsymbol{\nu}$, Normal intercambiable en el sentido finito (ver Definición 4). Por ejemplo, pueden definirse como condicionalmente independientes dado τ_ν,

$$\nu_i \sim \text{Normal}(0, \tau_\nu).$$

El término γ_t en (2.1) representa el efecto temporal estructurado, modelado de forma dinámica en una estructura de vecinos. Por ejemplo, podemos proponer un modelo de caminata aleatoria donde las distribuciones de los efectos son

$$\begin{aligned} \gamma_t | \gamma_{-t} &\sim \text{Normal}(\gamma_{t+1}, \tau_\gamma) \text{ para } t = 1 \\ \gamma_t | \gamma_{-t} &\sim \text{Normal}\left(\frac{\gamma_{t+1} + \gamma_{t-1}}{2}, \frac{\tau_\gamma}{2}\right) \text{ para } t = 2, ..., T-1 \\ \gamma_t | \gamma_{-t} &\sim \text{Normal}(\gamma_{t-1}, \tau_\gamma) \text{ para } t = T \end{aligned}$$

Mientras que el vector $\boldsymbol{\phi} = (\phi_1, ..., \phi_T)$ se puede modelar con una Normal intercambiable tal que

$$\phi_t \sim \text{Normal}(0, \tau_\phi) \text{ para } t = 1, .., T.$$

Existen varias formas de definir el término de interacción δ_{it}. Se puede asumir que los dos efectos no estructurados ν_i y ϕ_t interactúan a través de modificar la matriz de precisión de $(\boldsymbol{\nu}, \boldsymbol{\phi})$ como el producto escalar τ_ν (o τ_ϕ) y la llamada matriz de estructura F_ν (F_ϕ), que identifica la estructura vecindante; aquí la matriz de estructura F_δ puede factorizarse como el producto de Kronecker de las matrices de estructura para $\boldsymbol{\nu}$ y $\boldsymbol{\phi}$: $F_\delta = F_\nu \otimes F_\phi = I \otimes I = I$ (ya que ambos $\boldsymbol{\nu}$ y $\boldsymbol{\phi}$ son no estructurados respecto a las covarianzas, es decir, tienen matrices diagonales por matriz de covarianza). Consecuentemente en este caso, no asumimos estructura ni espacial ni temporal en la interacción y por lo tanto $\delta_{it} \sim \text{Normal}(0, \tau_\delta)$.

Este modelo puede considerarse como un caso particular del modelo especificado en la Tabla 2.1 donde $\boldsymbol{x} = \{\alpha, \boldsymbol{\beta}, \boldsymbol{\upsilon}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\eta}\}$ es el campo Gaussiano Markoviano y $\boldsymbol{\theta} =$

$\{\tau_\upsilon, \tau_\nu, \tau_\gamma, \tau_\phi, \tau_\delta\}$ son los hiperparámetros .

Recordemos que la especificación del efecto espacial y temporal (mientras sea en términos de distribuciones normales) simplemente cambia la forma y parámetros de la matriz de varianza (o de precisión). En general podemos proponer ver modelos con efectos fijos y aleatorios (espacio-temporal) como un modelo con estructura jerárquica siguiendo el esquema en la Tabla 2.1.

Tabla 2.1: Esquema general de modelos jerárquicos con efectos mixtos lineales

Nivel	Nombre	FGDP condicional	Especificación
I	Observaciones	$p(\boldsymbol{y}\|\boldsymbol{x},\theta)$	$p(\boldsymbol{y}\|\boldsymbol{x},\theta) = \prod_{i=1}^{n} p(y_i\|\theta)$
II	Campo aleatorio latente	$p(\boldsymbol{x};\theta)$	$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{w} + \boldsymbol{B}\beta, \boldsymbol{w} \sim \mathrm{N}\left(\boldsymbol{0}, Q^{-1}(\theta)\right)$
III	Hiperparámetros	$p(\theta)$	Definirse según conocimiento previo

Otros ejemplos de modelos jerárquicos que incorporan efectos aleatorios espaciales se pueden consultar en Banerjee, Carlin & Gelfand (2014).

2.1.2. Modelos para datos con excesos de ceros

Son diversas las razones por la que los datos tienden a reportar más ceros que los que que se pueden modelar con un proceso puntual simple. Entre estas razones están la naturaleza de los datos y la deficiencia en el sistema de captura o vigilancia. Debido a ésto, es conveniente poder incorporar modelos que capturen este excedente de ceros en los datos. A continuación describimos dos de ellos.

Usualmente se usan uno de dos enfoques para tratar datos con exceso de ceros. El modelo con obstáculo (*Hurdle*) y el modelo inflado en cero. Ambos modelos pueden verse como modelos de mezcla, y por tanto también como modelos jerárquicos. Ambos enfoques son similares, pero existen diferencias sutiles. A continuación se describen cada uno.

Modelos con obstáculo

Se trata de un modelo de mezclas con dos componentes. El primero genera valores distintos de ceros y el segundo es un punto de masa en cero. En general, el obstáculo podría estar en cualquier valor, no necesariamente cero. Notamos que bajo este modelo todos los ceros son generados por un mismo proceso, es decir que los ceros son ceros estructurales. En particular, consideramos la mezcla entre un punto de masa en el cero con probabilidad π_i y una distribución Poisson truncada en cero con probabilidad $(1 - \pi_i)$. Esto es,

$$P(Y_i = 0) = \pi_i, \qquad 0 \leq \pi_i \leq 1$$

$$P(Y_i = k) = (1 - \pi_i)\frac{\lambda_i^k e^{-\lambda_i}}{k!(1 - e^{-\lambda_i})}, \qquad \lambda_i > 0, k = 1, 2, ...$$

donde Y_i es la i-ésima observación. Esta definición puede extenderse de tal modo que consideremos un modelo de regresión log-lineal (Modelos de Regresión Lineal Generalizados Poisson) sobre los parámetros λ_i. Análogamente, se puede considerar una regresión logística sobre las π_i.

Modelos inflados

En contraste con el modelo anterior, la probabilidad de obtener un valor particular, como el cero, no sólo viene dada por la probabilidad de obtener el valor estructural. En el caso Poisson con valor inflado cero, la función de probabilidad del llamado modelo Poisson inflado en cero (ZIP) es

$$P(Y_i = 0) = \pi_i + (1 - \pi_i)e^{-\lambda_i}, \qquad \lambda_i > 0, 0 \leq \pi_i \leq 1$$

$$P(Y_i = k) = (1 - \pi_i)\frac{\lambda_i^k e^{-\lambda_i}}{k!}, \qquad \lambda_i > 0, k = 1, 2, ...$$

donde Y_i es la i-ésima observación. Podemos ver este modelo como un modelo de mezcla entre una v.a. concentrada en cero y un modelo Poisson. Al hacerlo de esta forma, estaríamos concibiendo una variable latente (x_i) que representa si la observación i-ésima es un cero estructural (viene de la distribución degenerada en cero) o no. Notemos que en este caso, es un poco más complicado proponer un modelo de regresión sobre π_i, pues a diferencia del modelo con obstáculo no todos los ceros son estructurales y no pueden tratarse como procesos de inferencia separados.

Si suponemos $\pi_i = \pi$ y $\lambda_i = \lambda$ para todo $i = 1, ..., n$, es fácil construir este modelo como un modelo jerárquico de tres niveles.

Nivel I. Observaciones

$$p(y_i|x_i, \lambda, \pi) = \begin{cases} \mathbb{1}[y_i = 0] & \text{si } x_i = 1 \\ \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} & \text{si } x_i = 0 \end{cases}.$$

Nivel II. Variables latentes

$$X_i|\pi, \lambda \sim \text{Bernoulli}(\pi).$$

Nivel III. Hiperparámetros

$$\lambda, \pi$$

Proponiendo un modelo semiconjugado para λ y π, $\lambda \sim \text{Gamma}(\alpha, \beta)$ y $\pi \sim \text{Beta}(a, b)$ independientes y escribiendo las densidades de forma conveniente

Nivel I. Observaciones

$$p(y_i|x_i, \lambda, \pi) = (\mathbb{1}[y_i = 0])^{x_i} \left(\frac{\lambda^{y_i} e^{-\lambda}}{y_i!}\right)^{1-x_i}.$$

Nivel II. Variables latentes

$$p(x_i|\pi,\lambda) \propto (\pi)^{x_i}(1-\pi)^{1-x_i}.$$

Nivel III. Hiperparámetros

$$p(\lambda,\pi) \propto \lambda^{\alpha-1}e^{-\beta\lambda}(\pi)^{a-1}(1-\pi)^{b-1}$$

La densidad generalizada conjunta posterior del modelo Poisson inflado en cero es

$$p(\boldsymbol{x},\lambda,\pi|\boldsymbol{y}) \propto \lambda^{\alpha-1}e^{-\beta\lambda}(\pi)^{a-1}(1-\pi)^{b-1}\prod_{i=1}^{n}\left[(\pi\mathbb{1}[y_i=0])^{x_i}\left((1-\pi)\frac{\lambda^{y_i}e^{-\lambda}}{y_i!}\right)^{1-x_i}\right]$$

Es interesante observar lo sencillo que es obtener el muestreador de Gibbs para este modelo, pues

$$X_i|X_{-i},\lambda,\pi \sim \text{Bernoulli}\left(\frac{\pi\mathbb{1}[y_i=0]}{\pi\mathbb{1}[y_i=0]+(1-\pi)\frac{\lambda^{y_i}e^{-\lambda}}{y_i!}}\right)$$

$$\pi|\boldsymbol{X}=\boldsymbol{x},\lambda \sim \text{Beta}\left(a+\sum_{i=1}^{n}x_i, b+n-\sum_{i=1}^{n}x_i\right)$$

$$\lambda|\boldsymbol{X}=\boldsymbol{x},\pi \sim \text{Gamma}\left(\alpha+\sum_{i=1}^{n}[(1-x_i)\times y_i], \frac{\beta}{1+(n-\sum_{i=1}^{n}x_i)\beta}\right)$$

Más detalles sobre los modelos para datos con exceso de ceros, tales como la aplicación de regresión en los parámetros de la mezcla, pueden encontrarse en Arab (2015).

2.2. Diferencia entre estimación y predicción

En el campo de la estadística existen una gran cantidad de tipos de problemas estadísticos, por mencionar algunos, estimación puntual, predicción por intervalos, pruebas de hipótesis, análisis de discriminante, análisis de grupos, etc. Todos ellos se diferencian según el tipo de variable (categórica nominal, ordinal, continua, discreta) de variable de respuesta y de las variables explicativas, así como el espacio en el que vive la solución que buscamos.

Una clasificación particularmente interesante es la dicotomía entre problemas de estimación y problemas de predicción. Tradicionalmente, se suelen caracterizar de forma informal estos dos problemas siendo la estimación lo que podemos decir de cantidades fijas que indexan o caracterizan nuestro modelo, y la predicción queda reservada para el conocimiento de cantidades aleatorias no observadas generadas por nuestro modelo estocástico.

Esta distinción se hace menos clara al tratar de definir que tipo de problema de inferencia es el que se hace sobre, por ejemplo, variables latentes. Según nuestro modelo, no son cantidades fijas ya que tienen una distribución de probabilidad, pero ya han ocurrido

(pues afectan las variables observadas) pero no fueron observadas y, como una observación perdida, ya no tenemos esperanzas de eventualmente observarla.

Una observación que podría ayudarnos a dar una distinción, al menos pragmática, de estos dos casos es que en la estimación las medidas de incertidumbre (como la varianza en la estimación puntual o el tamaño de los intervalos en estimación por intervalos) se reduce al aumentar el tamaño de muestra. Sin embargo, esto no ocurre con, por ejemplo, los intervalos de predicción, esta reducción esta limitada por lo que nuestro modelo considera ruido, error, o variabilidad intrínseca. En este sentido, la precisión de la predicción del modelo está comúnmente limitada por lo que el modelo considera ruido. Adicionalmente, en el caso de estimación puntual, podemos mencionar que la estimación tiene la noción de consistencia, mientras que para la predicción solo lo tendría respecto a la media.

Para ejemplificar esto consideremos el siguiente un modelo clásico Normal (μ, σ^2) y veamos el comportamiento de los intervalos de estimación y predicción.

Para una muestra $X_1, ..., X_n \sim \text{Normal}(\mu, \sigma^2)$ independiente con μ y σ^2 desconocidas,

$$\frac{\bar{X}}{S\sqrt{1/n}} \sim \text{t-Student}_{n-1}.$$

Por lo que los intervalos bilaterales de confianza $1-\alpha$ para este caso de estimación son de la forma

$$\bar{X} \pm t_{n-1}^{1-\alpha/2} S\sqrt{\frac{1}{n}}$$

Siendo $X_1, ..., X_n \sim \text{Normal}(\mu, \sigma^2)$, con μ,σ^2 y X_{n+1} desconocidas tenemos que

$$\frac{X_{n+1} - \bar{X}}{S\sqrt{1 + (1/n)}} \sim \text{t-Student}_{n-1}.$$

Por lo que los intervalos bilaterales de confianza $1-\alpha$ para este caso son de la forma

$$\bar{X} \pm t_{n-1}^{1-\alpha/2} S\sqrt{1 + \frac{1}{n}}.$$

Este ejemplo se trata del método de cantidades pivotales a un problema de predicción.

En relación al intervalo de predicción, consideremos $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$, tenemos que $\bar{X} \sim \text{Normal}\left(\mu, \frac{1}{n}\right)$ que es independiente a X_{n+1}. Entonces

$$\frac{X_{n+1} - \overline{x}}{\sqrt{1 + (1/n)}} \sim \text{Normal}(0, 1)$$

Además, si $X_1, ..., X_n \sim \text{Normal}(0, \sigma^2)$ entonces

$$\frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

De nuevo, por independencia de s y X_{n+1}

$$\frac{X_{n+1}}{s} \sim \text{t-Student}_{n-1}.$$

Finalmente, como $\bar{x}_n$ y s son independientes, podemos combinarlas en una cantidad pivotal

$$\frac{X_{n+1} - \bar{x}_n}{s_n\sqrt{1+1/n}} \sim \text{t-Student}_{n-1}.$$

2.3. Aproximación de Laplace clásica

La aproximación de Laplace es una aproximación a una integral con cierta forma. A pesar de que podría parecer restrictiva, en realidad esta forma es muy general. Se desea calcular la integral de la forma

$$I = \int q(\boldsymbol{\theta}) \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}$$

donde $q : \mathbb{R}^d \to \mathbb{R}$ y $h : \mathbb{R}^d \to \mathbb{R}$ son funciones suaves de $\boldsymbol{\theta}$. Supongamos que $h(\cdot)$ tiene un mínimo en $\hat{\boldsymbol{\theta}}$ y es dos veces diferenciable. El método de Laplace aproxima I a través de

$$\hat{I} = q(\hat{\boldsymbol{\theta}})(2\pi/n)^{d/2}|\Sigma(\hat{\boldsymbol{\theta}})|^{1/2} \exp\{-nh(\hat{\boldsymbol{\theta}})\}, \tag{2.2}$$

donde

$$\Sigma(\hat{\boldsymbol{\theta}}) = \left\{ \frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}} \right\}^{-1}.$$

Proposición 1. *Si* $(\theta - \hat{\theta}) = \mathcal{O}(n^{-1/2})$*, podemos decir que el error relativo de* $\hat{I}$ *es* $\mathcal{O}(1/n)$*. Es decir,*

$$\hat{I} = I\{1 + O(n^{-1})\}.$$

Esta demostración de la aproximación de Laplace se basa en la expansión de Taylor. Se reproduce el caso univariado de Gutiérrez-Peña (1997) ya que el caso multivariado es análogo.

Demostración. Consideramos la expansión en serie de Taylor tanto de $h(\cdot)$ como de $q(\cdot)$ alrededor de $\hat{\boldsymbol{\theta}}$. Supongamos que $(\theta - \hat{\theta}) = \mathcal{O}(n^{-1/2})$.

Desarrollando $h(\theta)$ alrededor de $\hat{\theta}$ tenemos

$$h(\theta) = h(\hat{\theta}) + \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k h(\hat{\theta})}{\partial \theta^k} (\theta - \hat{\theta})^k. \tag{2.3}$$

Notemos que,

$$\frac{1}{k!} \frac{\delta^k h(\hat{\theta})}{\partial \theta^k} = \mathcal{O}(1) \text{ para toda } k = 1, 2, ...$$

por lo que al multiplicar por $n(\theta - \hat{\theta})$,

$$\frac{n}{k!}\frac{\delta^k h(\hat{\theta})}{\partial\theta^k}(\theta - \hat{\theta}) = \mathcal{O}(n^{1-k/2}) \text{ para toda } k = 1, 2, ...$$

Considerando (2.3) hasta su tercer grado, multiplicando por n y substituyendo la expresión anterior se tiene

$$nh(\theta) = nh(\hat{\theta}) + \frac{n}{2\Sigma(\hat{\theta})}(\theta - \hat{\theta})^2 + \frac{n\tau(\theta)}{6} + \mathcal{O}(n^{-1}),$$

donde

$$\Sigma(\theta) = \left\{\frac{\partial^2 h(\theta)}{\partial\theta^2}\right\}^{-1} \text{ y } \tau(\theta) = \frac{\partial^3 h(\hat{\theta})}{\partial\theta^3}(\theta - \hat{\theta})^3.$$

Es decir,

$$\exp\{-nh(\theta)\} = \exp\{-nh(\hat{\theta})\}\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta - \hat{\theta})^2\right\}\left\{1 - \frac{n\tau(\theta)}{6} + \mathcal{O}(n^{-1})\right\}\{1+\mathcal{O}(n^{-1})\}.$$

De manera similar, al desarrollar Taylor de $q(\theta)$ alrededor de $\hat{\theta}$ tenemos

$$q(\theta) = q(\hat{\theta}) + \frac{\partial q(\hat{\theta})}{q'(\hat{\theta})\theta}(\theta - \hat{\theta}) + \mathcal{O}(n^{-1}).$$

Así,

$$\begin{aligned} q(\theta)\exp\{-nh(\theta)\} = &\left\{q(\hat{\theta}) + q'(\hat{\theta})(\theta - \hat{\theta}) - \frac{nq(\hat{\theta})\tau(\theta)}{6} - \frac{nq'(\hat{\theta})\tau(\theta)}{6}(\theta - \hat{\theta}) + \mathcal{O}(n^{-1})\right\} \\ &\times \exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta - \hat{\theta})^2\right\}\exp\{-nh(\hat{\theta})\}\{1 + \mathcal{O}(n^{-1})\}. \end{aligned} \tag{2.4}$$

Nos interesa integrar (2.4) y notemos que

$$\begin{aligned} \int \exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta - \hat{\theta})^2\right\} d\theta &= (2\pi/n)^{1/2}\Sigma(\hat{\theta})^{1/2}, \\ \int (\theta - \hat{\theta})\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta - \hat{\theta})^2\right\} d\theta &= 0, \\ \int \tau(\theta)\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta - \hat{\theta})^2\right\} d\theta &= 0, \text{ y} \\ \int n\tau(\theta)(\theta - \hat{\theta})\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta - \hat{\theta})^2\right\} d\theta &= \mathcal{O}(n^{-1}). \end{aligned}$$

Al integrar ambos lados de (2.4),

$$\begin{aligned} I &= q(\hat{\theta})(2\pi/n)^{1/2}\Sigma(\hat{\theta})^{1/2}\exp\{-nh(\hat{\theta})\}\{1+\mathcal{O}(n^{-1})\} \\ &= \hat{I}\{1+\mathcal{O}(n^{-1})\}. \end{aligned}$$

■

2.4. Selección de modelos Bayesianos

En esta sección introduciremos algunos de los criterios propuestos por varios autores, algunos muy usados en la práctica, empleados para resolver el problema de pruebas de hipótesis y el problema de selección de modelos.

2.4.1. Verosimilitud marginal y factor de Bayes

La verosimilitud marginal de un modelo es simplemente la función de densidad evaluada en los datos observados dado cierto modelo, es decir, $p(\boldsymbol{y}|\mathcal{M})$, que es marginal de los parámetros del modelo. Cuando consideramos un conjunto de M modelos $\{\mathcal{M}_m\}_{m=1}^{M}$, sus respectivas verosimilitudes marginales pueden representarse como $p(\boldsymbol{y}|\mathcal{M}_m)$ para indicar que son distintas para cada uno de los distintos modelos. Cuando el modelo en cuestión es único es usual omitir la notación condicional en $\mathcal{M}$, dejando sólo $p(\boldsymbol{y})$. En general es difícil calcular la verosimilitud marginal debido a que requiere integrar todas las cantidades desconocidad, sean parámetros o variables latentes. En el contexto donde sólo los parámetros son desconocidos

$$p(\boldsymbol{y}) = \mathrm{E}_{\boldsymbol{\theta}}(p(\boldsymbol{y}|\boldsymbol{\theta})) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Mientras que en el contexto de Campos Gaussianos Markovianos Ocultos (ver la definición 1 en la Sección 1.4.3) donde $\boldsymbol{y}$ son las variables observables y $\boldsymbol{x}$ es el Campo Gaussiano Markoviano Oculto

$$p(\boldsymbol{y}) = \int \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}d\boldsymbol{\theta}.$$

Como vemos, en el modelo usual de INLA es relativamente claro como calcular una aproximación a la verosimilitud marginal, también llamada evidencia o densidad predictiva previa evaluada en los datos. En modelos que salen de este esquema no siempre es claro ni sencillo como calcular o aproximar la verosimilitud marginal. En el caso de métodos variacionales Bayesianos aprovechan una desigualdad por abajo (cota inferior) para aproximar esta evidencia a favor del modelo. En INLA podemos aproximar $p(\boldsymbol{y})$ directamente al integrar la función $p(\boldsymbol{y}|\boldsymbol{\theta})$ usando los puntos de integración de $\boldsymbol{\theta}$. Veremos en la Sección 3.2.3 que $p(\boldsymbol{y}|\boldsymbol{\theta})$ se calcula con la aproximación de Laplace.

Más aún, la verosimilitud marginal puede usarse para calcular la probabilidad posterior de un modelo ajustado mediante

$$p(\mathcal{M}_m|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mathcal{M}_m)p(\mathcal{M}_m),$$

donde $p(\mathcal{M}_m)$ es la probabilidad previa de cada modelo.

Por último, la verosimilitud marginal puede usarse para calcular el factor de Bayes para compara dos modelos dados.

Factor de Bayes (BF)

El factor de Bayes corresponde a la razón de dos verosimilitudes marginales de dos hipótesis a contrastar. La probabilidad posterior $p(\mathcal{M}|\boldsymbol{y})$ del modelo $\mathcal{M}$ dados los datos $\boldsymbol{y}$ esta dada por el Teorema de Bayes

$$p(\mathcal{M}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{M})p(\mathcal{M})}{p(\boldsymbol{y})}.$$

Para un problema de selección de modelos en que debemos escoger entre dos modelos basado en los datos $\boldsymbol{y}$, la plausibilidad de ambos modelos $\mathcal{M}_1$ y $\mathcal{M}_2$, parametrizados por el vector de parámetros $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$, se contrasta por el factor de Bayes K

$$K = \frac{p(\boldsymbol{y}|\mathcal{M}_1)}{p(\boldsymbol{y}|\mathcal{M}_2)} = \frac{\int p(\boldsymbol{\theta}_1|\mathcal{M}_1)p(\boldsymbol{y}|\boldsymbol{\theta}_1,\mathcal{M}_1)\,d\theta_1}{\int p(\boldsymbol{\theta}_2|\mathcal{M}_2)p(\boldsymbol{y}|\boldsymbol{\theta}_2,\mathcal{M}_2)\,d\boldsymbol{\theta}_2} = \frac{p(\mathcal{M}_1|\boldsymbol{y})}{p(\mathcal{M}_2|\boldsymbol{y})}\frac{p(\mathcal{M}_2)}{p(\mathcal{M}_1)}.$$

Si los dos modelos son igualmente probables inicialmente, se tiene que $p(\mathcal{M}_1) = p(\mathcal{M}_2)$, y entonces el factor de Bayes es igual a la razón de las probabilidades posteriores de $\mathcal{M}_1$ y $\mathcal{M}_2$.

Para interpretar el factor de Bayes podemos emplear la muy conocida escala de Jeffreys

$\log_{10}$ K	K	Contundencia de la evidencia
Menor a 0	Menor a 1	Negativa (Apoya a $\mathcal{M}_2$)
De 0 a 1/2	De 1 a 3.2	Apenas merece mención
De 1/2 a 1	De 3.2 a 10	Positiva
De 1 a 3/2	De 10 a 31.6	Fuerte
De 3/2 a 2	De 31.6 a 100	Muy Fuerte
Mayor a 2	Mayor a 100	Decisiva

Tabla 2.2: Escala de Jeffreys

o bien la escala sugerida en Kass & Raftery (1995).

ln K	K	Contundencia de la evidencia
De 0 a 1	De 1 a 3	No merece más que una breve mención
De 1 a 2.5	De 3 a 12	Positiva
De 2.5 a 5	De 12 a 150	Fuerte
> 5	> 150	Decisiva

Tabla 2.3: Escala de Kass & Raftery

2.4.2. Criterio de Información Bayesiano (BIC)

El Criterio de Información Bayesiano (BIC) en Claeskens & Hjort (2008) formalmente se define como

$$\text{BIC} = \ln(n)k - 2\ln(\widehat{L})$$

donde $\hat{L}$ es el valor maximizado de la función de verosimilitud del modelo $\mathcal{M}$, es decir $\hat{L} = p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}, \mathcal{M})$, donde $\widehat{\boldsymbol{\theta}}$ son los valores de los parámetros que maximizan la función de verosimilitud; $\boldsymbol{y}$ son los datos observados; n es el número de observaciones o tamaño de muestra y k es el número de parámetros estimados por el modelo.

Por ejemplo, en regresión lineal múltiple, los parámetros estimados son: el intercepto, los q parámetros de pendiente y la varianza (constante) de los errores. Por lo que $k = q + 2$.

En Konishi & Kitagawa (2008) derivan el BIC al aproximar la distribución de los datos, integrando respecto a los parámetros empleando la aproximación de Laplace empezando de la siguiente forma

$$p(\boldsymbol{y} \mid \mathcal{M}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M})\, d\boldsymbol{\theta}$$

donde $p(\boldsymbol{\theta} \mid \mathcal{M})$ es la previa para $\boldsymbol{\theta}$ bajo el modelo $\mathcal{M}$.

Como vimos en la Sección 2.3, las integrales que podemos calcular con la aproximación de Laplace son de la forma

$$I = \int q(\boldsymbol{\theta}) \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}$$

donde $q : \mathbb{R}^k \to \mathbb{R}$ y $h : \mathbb{R}^k \to \mathbb{R}$ son funciones suaves de $\boldsymbol{\theta}$. Siendo en este caso $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathcal{M})$ y $h(\boldsymbol{\theta}) = -\log p(y \mid \boldsymbol{\theta}, \mathcal{M})$ en el caso de una muestra independiente e idénticamente distribuida. El método de Laplace aproxima I a través de

$$\hat{I} = q(\hat{\boldsymbol{\theta}})(2\pi/n)^{k/2} |\Sigma(\hat{\boldsymbol{\theta}})|^{1/2} \exp\{-nh(\hat{\boldsymbol{\theta}})\},$$

donde

$$\Sigma(\hat{\boldsymbol{\theta}}) = \left\{ \frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}} \right\}^{-1}.$$

Al tratarse de la log verosimilitud (por observación), $\log(p(y|\boldsymbol{\theta}, \mathcal{M}))$, resulta que el va-

lor que maximiza $h(\boldsymbol{\theta})$ es el Estimador Máximo Verosímil (EMV), $\widehat{\boldsymbol{\theta}}$ y $\Sigma(\hat{\boldsymbol{\theta}})$ es información de Fisher promedio o información observada de Fisher por observación $\mathcal{I}(\boldsymbol{\theta})$.

De este modo,

$$\hat{I} = p(\hat{\boldsymbol{\theta}})(2\pi/n)^{k/2}|\mathcal{I}(\hat{\boldsymbol{\theta}})|^{1/2}\exp\{-n\log(p(y|\hat{\boldsymbol{\theta}},\mathcal{M}))\}$$

Conforme n aumenta, podemos ignorar $|\mathcal{I}(\widehat{\boldsymbol{\theta}})|$ y $p(\widehat{\boldsymbol{\theta}})$ ya que no dependen de n y son por tanto O grande ($\mathcal{O}(1)$). Si denotamos a la función de verosimilitud valuada en su máximo $\widehat{L} = L(\widehat{\boldsymbol{\theta}}) = p(\boldsymbol{y}|\hat{\boldsymbol{\theta}},\mathcal{M})$, entonces

$$p(\boldsymbol{y} \mid M) = \exp\{\ln\widehat{L} - (k/2)\ln(n) + O(1)\} = \exp(-\mathrm{BIC}/2 + O(1)),$$

donde el BIC queda definido como arriba.

Si bien aqui hemos justificado el uso de $\widehat{L}$ como el valor máximo de la función de verosimilitud, también es común encontrar y justificar las siguiente dos posibles alternativas:

(a) $\widehat{L}$ es la moda posterior Bayesiana de $L(\boldsymbol{\theta})$.

(b) es el EMV $\widehat{\boldsymbol{\theta}}$ valuado en $L(\boldsymbol{\theta})$ (siempre que $\frac{d}{d\boldsymbol{\theta}}p(\boldsymbol{\theta} \mid \mathcal{M})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \neq 0$).

Por lo tanto la posterior es

$$p(\mathcal{M} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \mathcal{M})p(\mathcal{M}) \approx \exp(-\mathrm{BIC}/2)p(\mathcal{M}).$$

2.4.3. Densidad Predictiva

Para lo que sigue, llamaremos indistintamente log densidad predictiva o log verosimilitud a $\log p(\boldsymbol{y}|\boldsymbol{\theta})$. A continuación presentamos un resultado asintótico de la densidad predictiva para modelos lineales normales que también es útil como punto comparación e interpretación para casos no lineales o no normales.

Bajo condiciones de regularidad estándar, la distribución posterior $p(\boldsymbol{\theta}|y)$, en el límite se aproxima a una distribución normal al aumentar el tamaño de muestra DeGroot (2005). En este límite asintótico la posterior es dominada por la veromilitud, por lo que la previa sólo contribuye un factor, mientras que a verosimilitud contribuye n factores, uno por cada observación. Así que la verosimilitud se aproxima a la misma distribución normal.

Conforme el tamaño de muestra $n \to \infty$ la distribución posterior $\boldsymbol{\theta}|y \to \text{Normal}(\boldsymbol{\theta}_0, V_0/n)$. En este límite, la log densidad predictiva es

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = c(\boldsymbol{y}) - \frac{1}{2}\left(k\log(2\pi) + \log|V_0/n| + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T(V_0/n)^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right),$$

donde $c(\boldsymbol{y})$ es una constante que solo depende de los datos $\boldsymbol{y}$ y la familia del modelo pero no de los parámetros y V_0/n corresponde a la matriz de covarianzas posterior de $\boldsymbol{\theta}$.

La distribución normal multivariada límite de $\boldsymbol{\theta}$ induce una distribución posterior de la log densidad predictiva que resulta ser igual a una constante (igual a $c(y) - \frac{1}{2}(k\log(2\pi) +$

$\log|V_0/n|$)) menos $\frac{1}{2}$ veces una variable aleatoria con distribución χ^2_k, donde k es la dimensión de $\boldsymbol{\theta}$, es decir, el número de parámetros en el modelo.

Es fácil ver que el máximo de esta distribución de la log densidad predictiva se alcanza cuando $\boldsymbol{\theta}$ es igual al estimador máximo verosímil (EMV) y la media posterior es un valor $\frac{k}{2}$ menor.

Para modelos no lineales normales, este resultado asintótico es sólo una aproximación, pero sera útil como punto de referencia para interpretar la log densidad predictiva como medida de la bondad del ajuste.

Podemos resumir la exactitud (accuracy) del modelo ajustado a los datos mediante

$$\begin{aligned}\text{lppd} &= \text{log densidad predictiva en un punto (log pointwise predictive density)}\\ &= \log p(\tilde{y}_i|\boldsymbol{y})\\ &= \log \int p(\tilde{y}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}\end{aligned}$$

donde $\tilde{y}_i$ son datos futuros a observar. Podemos definir el valor esperado para las observaciones fuera de la muestra $\tilde{y}_i$

$$\begin{aligned}\text{elpd} &= \text{log densidad predictiva esperada para una nueva observación}\\ &= \mathrm{E}_{f(\tilde{y}_i)} \log p(\tilde{y}_i|\boldsymbol{y})\\ &= \int \log p(\tilde{y}_i|\boldsymbol{y}) f(\tilde{y}_i) d\tilde{y}.-\end{aligned}$$

donde f es la distribución de los datos.

También podemos definir una medida de precisión para un conjunto de n datos

$$\begin{aligned}\text{elppd} &= \text{log densidad predictiva esperada para un nuevo conjunto de datos}\\ &= \sum_{i=1}^{n} \mathrm{E}_f \log p(\tilde{y}_i|\boldsymbol{y}).\end{aligned}$$

Por razones históricas, las medidas de exactitud predictiva son llamadas criterios de información y típicamente se definen en términos de la devianza (la log densidad predictiva de los datos dado un estimador puntual calculado del modelo ajustado, multiplicado por -2, es decir, $-2\log p(y|\hat{\boldsymbol{\theta}})$).

El estimador puntual calculado $\hat{\boldsymbol{\theta}}$ y la distribución posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ se obtienen de los datos y, y las predicciones para valores fuera de la muestra tipicamente son menos precisos que los de los datos dentro de la muestra.

En mucha de la literatura sobre presición predictiva, la inferencia de $\boldsymbol{\theta}$ no se resume usando una distribución posterior sino un estimador puntual $\hat{\boldsymbol{\theta}}$, típicamente el estimador de máxima verosimilitud (MLE). Así, la precisión para valores fuera de la muestra se define

mediante $\text{elpd}_{\hat{\theta}} = \text{E}_f(\log p(\tilde{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{y})))$ donde $\boldsymbol{y}$ y $\tilde{y}$ son ambos aleatorios.

2.4.4. Criterio de Información de Akaike (AIC)

Sea k el número de parámetros estimados en el modelo. La corrección del sesgo más simple está basado en la distribución asintótica posterior. Se resta el número de parámetros de la log densidad predictiva (verosimilitud valuada en el EMV):

$$\widehat{\text{elpd}}_{\text{AIC}} = \log p(y|\hat{\boldsymbol{\theta}}_{EMV}) - k.$$

Tal como lo definió Akaike (1973), el AIC es lo anterior multiplicado por -2, así

$$\text{AIC} = -2 \log p(y|\hat{\boldsymbol{\theta}}_{EMV}) + 2k.$$

2.4.5. Criterio de Información de Devianza (DIC)

El criterio de información de devianza es, por así decirlo, una versión Bayesiana del AIC y fue definido por Spiegelhalter, Best, Carlin & Van Der Linde (2002). Tiene dos diferencias respecto al AIC; en primer lugar remplaza el EMV $\hat{\boldsymbol{\theta}}$ por la media posterior $\hat{\boldsymbol{\theta}}_{\text{Bayes}} = \text{E}(\boldsymbol{\theta}|y)$ y en segundo remplaza k por una corrección del sesgo basada en los datos. Esta nueva medida de la precisión de la predicción es

$$\widehat{\text{elpd}}_{\text{DIC}} = \log p(y|\hat{\boldsymbol{\theta}}_{Bayes}) - p_{\text{DIC}},$$

donde el p_{DIC} es el número efectivo de parámetros definido por

$$p_{\text{DIC}} = 2\left(\log p(y|\hat{\boldsymbol{\theta}}_{Bayes}) - \text{E}(\log p(y|\boldsymbol{\theta})|\boldsymbol{y})\right),$$

donde la esperanza del segundo término es la media de la verosimilitud respecto a la distribución posterior de $\boldsymbol{\theta}$.

Una versión alternativa del DIC usa una definición ligeramente diferente del número efectivo de parámetros

$$p_{\text{DIC alt}} = 2\text{Var}(\log p(y|\boldsymbol{\theta})|\boldsymbol{y}).$$

La justificación del uso de estas cantidades para estimar el número efectivo de parámetros puede encontrarse en Spiegelhalter et al. (2002). La cantidad a la que llamamos el DIC se define en términos de la devianza en lugar de la log densidad predictiva,

$$\text{DIC} = -2 \log p(y|\hat{\boldsymbol{\theta}}_{\text{Bayes}}) + 2p_{\text{DIC}}.$$

2.4.6. Criterio de Información de Watanabe-Akaike (WAIC)

Introducido en Watanabe (2010), que entonces lo llamó también criterio de información ampliamente aplicable (Widely Applicable Information Criterion) es una aproximación más propiamente Bayesiana para estimar un valor esperado fuera de la muestra.

$$-2\text{elppd} = -2\sum_{i=1}^{n} \log p(y_i|\boldsymbol{y})$$

y definimos el criterio de Información de Watanabe-Akaike (WAIC) como

$$\text{WAIC} = -2\widehat{\text{elppd}}_{\text{WAIC i}}$$

donde $\text{elppd}_{\text{WAIC i}}$ depende de cuál de las dos posibles estimaciones del número de parámetros efectivos se use,$p_{\text{WAIC 1}}$ o $p_{\text{WAIC 2}}$. La primera aproximación es una diferencia, similar a como se construyó el p_{DIC}:

$$p_{\text{WAIC 1}} = 2\sum_{i=1}^{n} \left(\log(\text{E}(p(y_i|\boldsymbol{\theta})|\boldsymbol{y})) - \text{E}(\log p(y_i|\boldsymbol{\theta})|\boldsymbol{y})\right),$$

La otra medida usa la varianza de los términos individuales en la log densidad predictiva sumando sobre los n datos:

$$p_{\text{WAIC 2}} = \sum_{i=1}^{n} \text{Var}(\log p(y_i|\boldsymbol{\theta})|\boldsymbol{y}).$$

Esta expresión se ve similar a la formula para $p_{\text{DIC alt}}$ (salvo por el factor 2), pero es más estable pues calcula la varianza por separado para cada dato y luego los suma, sumar da estabilidad.

Podemos usar tanto $p_{\text{WAIC 1}}$ como $p_{\text{WAIC 2}}$ como corrección del sesgo:

$$\widehat{\text{elppd}}_{\text{WAIC}} = \text{lppd} - p_{\text{WAIC}}$$

Como hicimos para el AIC y el DIC, definimos el WAIC tal que

$$\text{WAIC} = -2\widehat{\text{elppd}}_{\text{WAIC}} = -2\text{lppd} + 2p_{\text{WAIC}},$$

de tal modo que este en la escala de la varianza. En la definición original de Watanabe, el WAIC es el negativo del promedio de las log densidades predictivas puntual (suponiendo la predicción de una nueva observación) y dividido por n son el factor 2.

Para más detalles sobre la densidad predictiva, el AIC, DIC y WAIC ver Gelman, Hwang & Vehtari (2014).

2.4.7. Ordenadas Predictivas Condicionales y Transformación Integral Predictiva

Consideremos la densidad posterior marginal de un elemento del campo Gaussiano Markoviano $\boldsymbol{x}$

$$p(x_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(x_i|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

del i-ésimo componente x_i de $\boldsymbol{x}$. Podemos aproximarlo mediante

$$\tilde{p}(x_i|\boldsymbol{y}) = \sum_k \tilde{p}(x_i|\boldsymbol{\theta}, \boldsymbol{y})\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})\Delta_k$$

usando la aproximación Gaussiana o de Laplace $\tilde{p}(x_i|\boldsymbol{y})$ de la densidad marginal posterior de un elemento del Campo Gaussiano Markoviano $p(x_i|\boldsymbol{y})$ y una aproximación de Laplace de $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$ de la densidad marginal posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ de los hiperparámetros. Los pesos Δ_k se escogen de manera apropiada de acuerdo al tipo de aproximación numérica que se eligió utilizar. Para aproximar $p(\boldsymbol{\theta}|\boldsymbol{y})$ tenemos que

$$p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \times p(\boldsymbol{\theta}|\boldsymbol{y}) \times p(\boldsymbol{y}),$$

se sigue claramente que

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \text{ para todo } \boldsymbol{x}.$$

Mientras que la marginal de x_i es

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

la densidad predictiva de y_i es

$$p(y_i|\boldsymbol{y}) = \int p(y_i|x_i, \boldsymbol{y})p(x_i|\boldsymbol{y})dx_i.$$

2.4.8. Ordenadas Predictivas Condicionales

Las ordenadas predictivas condicionales (ver Pettit, 1990) son un criterio tipo validación cruzada para la evaluación de modelos que se calcula para cada observación.

Así

$$\text{CPO}_i = p(y_i^{\text{obs}}|\boldsymbol{y}_{-i}),$$

donde y_i^{obs} denota el valor que en efecto se observó, a diferencia de un valor arbitrario y_i en la densidad de Y_i. Por lo tanto, para cada observación su correspondiente CPO es la probabilidad posterior de obtener dicha observación cuando el modelo se ajusta usando todos los datos salvo y_i^{obs}. Valores grandes del CPO indican un mejor ajuste del modelo a los datos, mientras que valores chicos indican mal ajuste del modelo a los datos, o quizás, que la observación y_i^{obs} es un valor atípico (outlier).

Notemos que

$$\text{CPO}_i = \int p(y_i^{\text{obs}}|\boldsymbol{y}_{-i}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})d\boldsymbol{\theta}.$$

El primer término en la integral del CPO_i es igual a

$$p(y_i^{\text{obs}}|\boldsymbol{y}_{-i},\boldsymbol{\theta}) = 1\Big/\int \frac{p(x_i|\boldsymbol{y},\boldsymbol{\theta})}{p(y_i^{\text{obs}}|x_i,\boldsymbol{\theta})}dx_i. \tag{2.5}$$

Para ver esto notemos que

$$p(x_i|\boldsymbol{y}_{-i},\boldsymbol{\theta}) = \frac{p(x_i|\boldsymbol{y},\boldsymbol{\theta})p(y_i^{\text{obs}}|\boldsymbol{y}_{-i},\boldsymbol{\theta})}{p(y_i^{\text{obs}}|x_i,\boldsymbol{\theta})}.$$

Integrando respecto a x_i obtenemos la igualdad (2.5). Esta expresión se aproxima mediante integración numérica.

Una medida que resume la información de los CPO_i es

$$-\sum_{i+1}^{n}\log(\text{CPO}_i) = -n\overline{\log \text{CPO}}$$

para la que valores más pequeños indica un mejor ajuste del modelo.

2.4.9. Transformación Integral Predictiva

La transformación integral predictiva mide, para cada observación, la probabilidad de que una nueva observación sea menor que la que en efecto se observó:

$$\text{PIT}_i = p(\tilde{Y}_i \leq y_i^{\text{obs}}|\boldsymbol{y}_{-i})$$

Para observaciones discretas, el PIT ajustado se calcula

$$\text{PIT}_i^{\text{ajustado}} = \text{PIT}_i - \frac{\text{CPO}_i}{2}$$

En este caso, el caso $\tilde{y}_i = y_i$ sólo se cuenta como una mitad. Si el modelo describe adecuadamente la observación, entonces la distribución de diferentes valores debería ser cercana a una distribución uniforme entre 0 y 1.

Notemos que

$$\text{PIT}_i = \int \text{Prob}(Y_i \leq y_i^{\text{obs}}|\boldsymbol{y}_{-i},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})d\boldsymbol{\theta}.$$

El primer término en la expresión del PIT puede escribirse

$$\text{Prob}(Y_i \leq y_i^{\text{obs}}|\boldsymbol{y}_{-i},\boldsymbol{\theta}) = \int \text{Prob}(Y_i \leq y_i^{\text{obs}}|\boldsymbol{x}_i,\boldsymbol{\theta})p(x_i|y_i,\boldsymbol{\theta})dx_i.$$

El primer término de esta integral puede calcularse fácilmente de la verosimilitud (es la función de distribución del modelo de salida). El segundo término se puede calcular de $p(x_i|\boldsymbol{y}_{-i},\boldsymbol{\theta})$ usando $p(y_i^{\text{obs}}|\boldsymbol{y}_{-i},\boldsymbol{\theta})$ como se calcula en el CPO.

Por último, necesitamos calcular

$$p(\boldsymbol{\theta}|\boldsymbol{y}_i) = \frac{p(\boldsymbol{\theta}|\boldsymbol{y})p(y_i^{\text{obs}}|\boldsymbol{y}_{-i})}{p(y_i^{\text{obs}}|\boldsymbol{y}_{-i},\boldsymbol{\theta})}$$

Por lo tanto, la constante de normalización es

$$p(y_i^{\text{obs}}|\boldsymbol{y}_{-i}) = 1 \Big/ \int \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{p(y_i^{\text{obs}}|\boldsymbol{y}_{-i},\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

Para más detalles sobre el CPO y el PIT ver Held, Schrödle & Rue (2010).

2.4.10. Pruebas de bondad de ajuste para los PIT's

Cuando los PIT's tienen distribución uniforme tenemos un indicador del buen ajuste del modelo. En esta sección justificaremos la distribución de uniforme de los PIT's bajo el modelo correcto y revisamos algunas de las pruebas de bondad de ajuste que pueden utilizarse para probar su cumplimiento en el análisis de datos.

Para justificar la distribución uniforme teórica de los PIT's mencionamos que algunas de pruebas de bondad de ajuste más conocidas se basan en el resultado conocido como Teorema de la Transformada Inversa.

Teorema 2.4.1. *Teorema de la Transformada Inversa*

Sea $Y_1, ..., Y_n$ una muestra aleatoria de variables continuas condicionalmente independiente dados los parámetros $\boldsymbol{\theta}$ e idénticamente distribuidas. Sea $F(y;\boldsymbol{\theta})$ su correspondiente función de distribución acumulada. Entonces $F(Y_1;\theta), ..., F(Y_n;\boldsymbol{\theta})$ se distribuye como una muestra aleatoria independiente Uniforme$(0,1)$.

Este resultado se usa de forma frecuente para simular de cualquier variable aleatoria continua. Además, no es difícil extender este método de simulación a variables discretas. Al realizar un ajuste de un modelo apropiado podemos transformar los datos observados y usar un criterio de bondad de ajuste donde la hipótesis nula es que los datos transformados tienen distribución uniforme.

Notemos que aplicar el Teorema de la Transformada Inversa a los casos en INLA tiene algunas complicaciones. Para empezar, la muestra aleatoria en el Teorema 2.4.2 es una muestra intercambiable, supuesto que generalmente no tenemos en los modelos con campos Markovianos Gaussianos latentes. Además, si nuestros modelos fueran directamente sobre los conteos de eventos violentos tendríamos el problema de que las variables aleatorias son discretas, por lo que las transformadas ya no tendrían una distribución uniforme continua. Afortunadamente, podemos resolver esta última dificultad trabajando no sobre los conteos, sino sobre los conteos ponderados por tamaño de la población, que si pueden modelarse con una variable aleatoria continua.

El problema que resta es el de la intercambiabilidad. Sin embargo, si suponemos que el campo aleatorio latente $\boldsymbol{x}$ es conocido podemos obtener un resultado similar.

Teorema 2.4.2. *Teorema de la Transformada Inversa para CMGL*

Sea $Y_1, ..., Y_n$ una muestra aleatoria de variables continuas donde la función de densidad de cada variable aleatoria es $p(y_i|x_i, \boldsymbol{\theta})$. Sea $F_{Y_i}(y_i|x_i, \theta)$ su correspondiente función de distribución acumulada. Entonces $F_{Y_1}(Y_1|x_1, \theta), ..., F_{Y_n}(Y_n|x_i, \theta)$ se distribuye como una muestra aleatoria independiente Uniforme$(0, 1)$.

Notemos que en general no podemos conocer precisamente el valor de $\boldsymbol{x}$, así como tampoco podemos conocer exactamente $\boldsymbol{\theta}$. Sin embargo, ahora sabemos que este es el comportamiento que esperaríamos si tuviéramos un conocimiento perfecto del fenómeno suponiendo que se comporta de acuerdo al modelo salvo por la incertidumbre intrínseca de $Y_i|x_i, \theta$.

Por otro lado, como se trata de modelos de inferencia en realidad no vamos a conocer exactamente $F_{Y_i}(Y_i|x_i, \theta)$, sino que podemos aproximarla. Notemos que los PIT's son básicamente una aproximación a estas cantidades, salvo que se usan las distribuciones posteriores dada una muestra $y_1, ..., y_n$ y para aproximar F_{Y_i} usan solo $\boldsymbol{y}_{-i}$. Esto debido a que si se usa la muestra completa estaríamos usando información que es en realidad incertidumbre intrínseca. Es decir, usamos información de la instancia especifica y no sólo del conocimiento del modelo. Por eso empleamos las predictivas acumuadas de tipo validación cruzada.

Por último, debemos evaluar que tan "distantes" están los PIT's observados de una muestra independiente Uniforme$(0, 1)$. Esto se puede hacer de forma inicial visualizando el histograma o los llamados gráficos P-P y los gráficos Q-Q. Otro recurso para evaluar (que además permite compara de forma sencilla varios modelos) es mediante pruebas de bondad de ajuste. Para cada una de ellas debe elegirse un estadístico de prueba y examinar los p-valores. En esta tesis utilizamos dos, la prueba χ^2 de Pearson y al prueba de Kolmogorov-Smirnov.

Notemos que asumimos que la variable observada era continua. Faltan examinar los casos discretos finitos y discretos infinito numerables. El caso discreto finito es sencillo, pues la única diferencia es que en lugar de distribuirse como uniformes continuas en $[0, 1]$ ahora son uniformes discretas con dominio en $\{\frac{1}{n}, \frac{2}{n}, ..., 1\}$. Es un poco más complicado el caso discreto infinito numerable, que es el caso del modelo Poisson. Afortunadamente, con el modelo Poisson tenemos algunos resultados asintóticos bien conocidos que nos pueden servir de aproximación y llevaran el caso valores esperados grandes al caso continuo.

Recordemos que es posible aproximar la distribución Poisson mediante una distribución normal cuando el parámetro λ es grande. Esta aproximación se justifica debido a que cuando λ tiende a infinito la distribución Poisson converge en distribución a una distribución Normal. Se suele denotar

$$\text{Poisson}(\lambda) \xrightarrow[\lambda\to\infty]{\text{F}} \text{Normal}\left(\lambda, \sqrt{\lambda}\right)$$

y queremos decir que si $X_\lambda \sim \text{Poisson}(\lambda)$ entonces

$$\lim_{\lambda\to\infty} \text{P}\left(\frac{X_\lambda - \lambda}{\sqrt{\lambda}} \leq z\right) = P\left(Z \leq z\right), \text{con } Z \sim \text{Normal}(0, 1).$$

Esto se puede demostrar de forma muy sencilla apelando a que la distribución Poisson es infinitamente divisible y empleando el Teorema del Límite Central.

Existe cierto consenso en que la aproximación es adecuada cuando $\lambda \geq 10$. Es por esto que si esperamos que la mayoría de las μ_i sean considerablemente mayores a 10 el comportamiento de los PIT sea parecido al caso continuo, es decir, que se distribuyan como una Uniforme continua en $[0, 1]$. La ventaja de esta aproximación es que no se requiere conocer de la distribución verdadera para conocer la distribución de los PIT bajo la hipótesis nula, solo se requiere para calcular los PIT's.

Sin embargo, en ejemplos numéricos se puede ver que incluso para valores no tan grandes como por ejemplo $\lambda = 20$, la distribución de los $F_Y(Y)$ dista mucho de ser Uniforme$(0, 1)$.

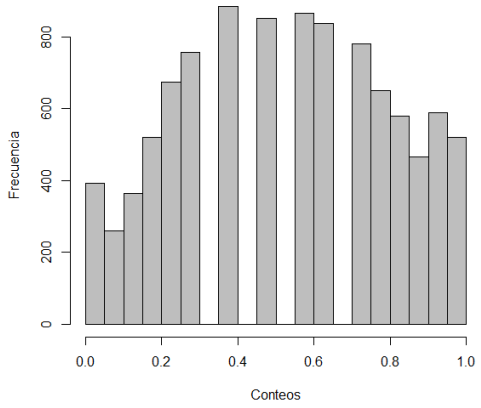


Figura 2.2: Histograma de $F_Y(Y)$ con $\lambda = 20$

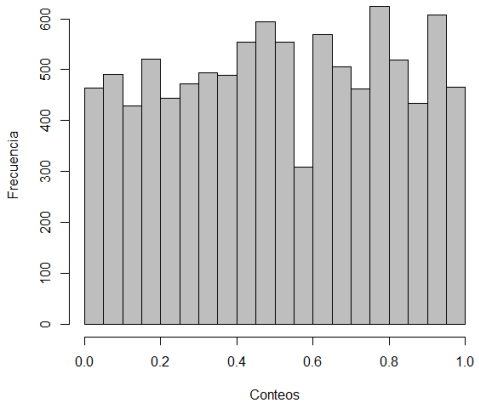


Figura 2.3: Histograma de $F_Y(Y)$ con $\lambda = 200$

En la Figura 2.2 observamos no sólo que hay intervalos sin observaciones, que es de esperarse pues el dominio de los $F_Y(Y)$ es discreto, si no que además se concentran más en el centro (cerca de 0.5) que en los extremos (o colas). Al aumentar el tamaño de λ (por ejemplo a 200) ambos efectos se corrigen, como podemos ver en la Figura 2.3. Sin embargo, ya no resulta muy adecuada justificación para valores no muy grandes de λ. Vale la pena notar que usualmente cuando la distribución estimada no es la distribución de la que vienen los datos los $F_Y(Y)$ tienden a concentrarse en alguno de los extremos o en ambos. Por ejemplo, una concentración alta en el centro indicaría falta de eventuales valores raros o en las colas de la distribución.

Es posible emplear una modificación de los $F_Y(Y)$ o PIT's que sea continua y se pueda comparar su distribución con la de una variable Uniforme$(0, 1)$. Esta se inspira en el método para simular distribuciones discretas partiendo de uniformes en $[0, 1]$.

Sea X una variable aleatoria discreta con probabilidad de masa

$$\mathrm{P}(X = x_j) = p_j, j = 0, 1, ...$$

Sea $U \sim$ Uniforme$(0, 1)$ una simulación de una variable aleatoria con distribución

uniforme. El método de la transformada inversa es

$$X = \begin{cases} x_0 & \text{si } U < p_0 \\ x_1 & \text{si } p_0 \leq U < p_0 + p_1 \\ \vdots & \\ x_i & \text{si } p_0 + \ldots + p_{j-1} \leq U < p_0 + \ldots + p_{j-1} + p_j \\ \vdots & \end{cases}$$

Así, $\text{P}(X = x_j) = \text{P}(\sum_{i=0}^{j-1} p_i \leq U < \sum_{i=0}^{j} p_i) = p_j$.

Sea $F(x) = \text{P}(X \leq x)$ la función de distribución acumulada. Sabemos que F es una función creciente, escalonada y que toma valores entre 0 y 1. Si se ordenan los valores de la variable en forma creciente.

$$x_0 < x_1 < \ldots < x_n < \ldots,$$

entonces

$$F(x_j) = \sum_{k=0}^{j} p_k.$$

Evaluar una muestra $X_1, .., X_n \sim F$ en F no nos dará una muestra que se distribuya Uniforme$(0, 1)$, pues es claro que de entrada el dominio no es $[0, 1]$. Sin embargo, esto puede corregirse simulando una uniforme en $(F(x_{j-1}), F(x_j)]$ o bien simulando una W v.a. continua con distribución uniforme y definir $U' = F^-(X) + W(F(X) - F^-(X))$ donde $F^-(z) = \lim_{x \to z^-} F(x)$. Es claro que en este caso discreto $F(X) - F^-(X) = p(X)$, es decir, la función de probabilidad. De modo que

$$U' = F^-(X) + W(p(X)) \tag{2.6}$$

En la Figura 2.4 podemos ver que el histograma de los U' sugiere de forma más contundente que los U' se distribuyen Uniforme$(0, 1)$. Esto es muy útil pues hecha la corrección no tenemos que tratar de forma especial cada uno de los casos de variables aleatorias (finita, discreta infinita, continua o mezcla de las anteriores). Basta encontrar una prueba de bondad de ajuste para el caso Uniforme$(0, 1)$.

La modificación es

$$Y = F_x(X) - W * \sum_{i=1}^{\infty} P_i \mathbb{1}(X = T_i)$$

La expresión en la Ecuación (2.6) es particularmente útil pues podemos expresar la corrección de los PIT en termino de los PIT y los CPO. Así, si bien en el caso discreto no

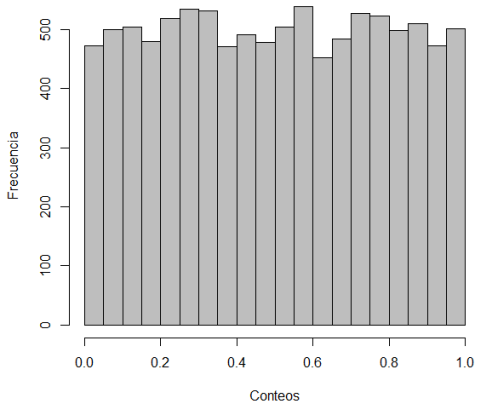


Figura 2.4: Histograma de $F_Y(Y)$ corregido con $\lambda = 200$

finito los PIT no se distribuyen Uniforme$(0, 1)$, se cumple que los PIT', definidos como

$$\text{PIT}' = (\text{PIT} - \text{CPO}) + \text{CPO} \times W = \text{PIT} - \text{CPO} \times (1 - W) \text{ con } W \sim \text{Uniforme}(0, 1)$$

se distribuyen Uniforme$(0, 1)$. Notemos que si $W \sim$ Uniforme$(0, 1)$ también $1 - W \sim$ Uniforme$(0, 1)$. Así, basta tomar

$$\text{PIT'=PIT - CPO} \times W \text{ con } W \sim \text{Uniforme}(0, 1).$$

Adicionalmente, si tomamos el valor esperado respecto a W tenemos que

$$\text{E(PIT')} = \text{PIT} - \text{CPO} \times (0.5)$$

que coincide con el PIT ajustado para observaciones discretas.

Por último, podemos proponer un par de pruebas muy usadas para pruebas de bondad de ajuste. Aún si debido al tamaño de la muestra los p-valores siempre resultan muy chicos, podemos emplear los estadísticos como un índice informal de la "cercanía" a la hipótesis nula (que los PIT's se distribuyan como una muestra independiente Uniforme$(0, 1)$).

Para justificar esta modificación consideremos el siguiente ejemplo: sea X una variable aleatoria con distribución multinomial $(1/4, 1/2, 1/8, 1/8)$. Con el fin de que esta variable tenga valores en los reales los valores de salida asociados serán $(1, 2, 3, 4)$. De este modo la función de distribución acumulada es

$$F_X(x) = \begin{cases} x < 1 & 0 \\ 1 \leq x < 2 & 1/4 \\ 2 \leq x < 3 & 3/4 \\ 3 \leq x < 4 & 7/8 \\ 4 \leq x & 1, \end{cases}$$

y podemos ver su gráfica en la Figura 2.5.

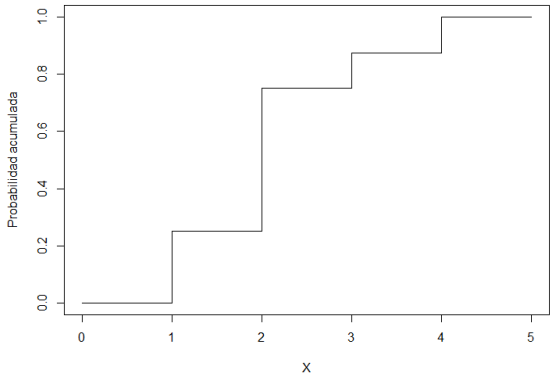


Figura 2.5: Función de densidad de la multinomial

Al transformar la variable aleatoria X aplicando su función de distribución acumulada, es decir, $Y = F_X(X)$, tenemos que la distribución acumuldada de Y esta dada por la gráfica en la Figura 2.6.

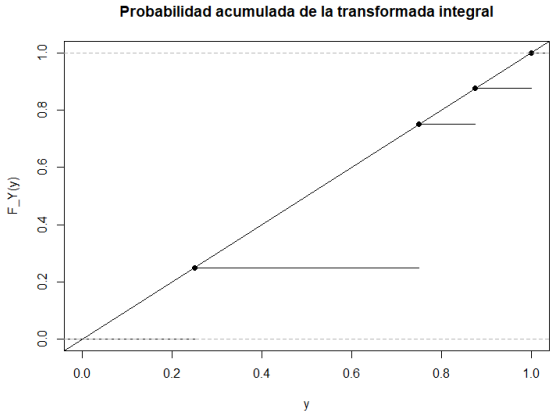


Figura 2.6: Función de densidad de la multinomial

Notemos que la función de dstribución al dar un salto discreto siempre vuelve a coincidir con la recta identidad. Al agregar estar la variable aleatoria $W * P(X = x)$ hacemos que ahora la función de distribución acumulada coincida perfectamente con la función identidad. Simulamos una muestra de tamaño 1000 y aplicamos la transformación a las mismas. Vemos en la Figura 2.7 que esta muestra se encuentra muy cerca de la recta identidad, por lo que la distribución de esta variable es muy cercana a la Uniforme $(0, 1)$.

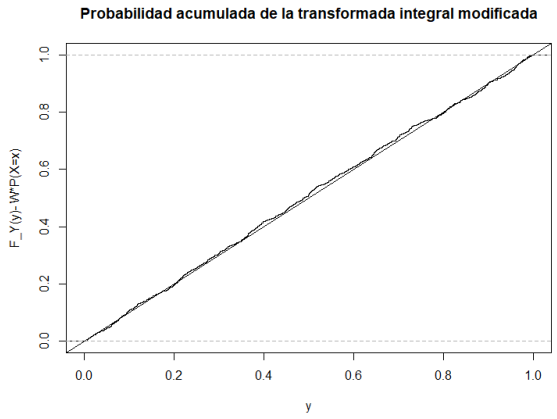


Figura 2.7: Función de densidad de la multinomial

2.4.10.1. Prueba χ^2 de Pearson

Supongamos un modelo multinomial, es decir, los resultados se clasifican en alguna de $k+1$ categorías mutuamente exclusivas $A_1, A_2, ..., A_{k+1}$. Definamos $p_j = P(A_j), j = 1, ..., k+1$. Se realizan n repeticiones independientes del experimento. Sea N_j el número de veces que resulta la categoría A_j, $j = 1, ..., k+1$ (Notemos que $\sum_{i=1}^{k+1} N_j = n$). Entonces

$$T_k = \sum_{j=1}^{k+1} \frac{(N_j - np_j)^2}{np_j}$$

tiene asintóticamente una distribución χ^2 con k grados de libertad.

Sea X variable aleatoria con función de distribución $F(x)$. Supongamos que el soporte de X es $s_1 \cup ... \cup s_k \cup s_{k+1}$ con $s_i \cap s_j = \emptyset$ y definamos $p_i = \int_{s_i} dF(z), i = 1, ..., k, k+1$. Si contamos con n observaciones $X_1, ..., X_n$ i.i.d. $\sim F(x)$ definamos

$$N_i = \sum_{m=1}^{n} \mathbb{1}_{s_i}(X_m), i = 1, ..., k, k+1.$$

Entonces $N_i \sim$ Binomial(n, p_i), y el vector $(N_1, ..., N_k, N_{k+1})$ es multinomial con parámetros $(n, p_1, ..., p_k, p_{k+1})$.

Probar la hipótesis $H_0 : F(x) = F_0(x)$ es equivalente a probar

$$H_0 : p_1 = p_{01}, ..., p_k = p_{0k}, p_{k+1} = p_{0(k+1)}$$

Bajo H_0, $(N_1, ..., N_k, N_{k+1}) \sim$ Multinomial$(n, p_{01}, ..., p_{0k}, p_{0(k+1)})$, por lo que

$$T_k \equiv \sum_{j=1}^{k+1} \frac{(N_j - np_{0j})^2}{np_j} \xrightarrow{\text{d}} \chi^2_k$$

cuando $n \to \infty$. Por lo tanto, si H_0 es cierta y se tiene n suficientemente grande, al calcular

el valor t_k con la muestra $x_1, ..., x_k$ se tiene el p-valor $P(\chi^2(k) > t_k)$. Si este valor es pequeño existe evidencia en contra de H_0, esto es, evidencia de que $F(x) \neq F_0(x)$.

En este contexto nos interesa probar $H_0 : F(x) = \text{Uniforme}(0, 1)$. Por simplicidad, podemos tomar s_i de longitud igual en el intervalo $[0, 1]$. Por ejemplo $s_i = \left(\frac{i-1}{k+1}, \frac{i}{k+1}\right)$ para $i = 1, ..., k + 1$. Así, bajo H_0, $p_i = \frac{1}{k+1}$.

2.4.10.2. Prueba de Kolmogorov-Smirnov

La función de distribución empírica F_n para n v.a.i.i.d. X_i se define

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$$

donde $I_{[-\infty,x]}(X_i)$ es la función indicadora, igual a 1 si $X_i \leq x$ e igual a 0 en otro caso.

El estadístico de Kolmogorov-Smirnov para una función de distribución acumulada $F(x)$ dada es

$$D_n = \sup_x |F_n(x) - F(x)|,$$

donde $\sup_x$ es el supremo del conjunto de distancias.

Por el Teorema de Glivenko-Cantelli, si la muestra viene de la distribución $F(x)$, entonces D_n converge a 0 casi seguramente en el límite cuando $n \to \infty$.

Definición 2.4.3. La distribución de Kolmogorov

La distribución de Kolmogorov es la distribución de la variable aleatoria

$$K = \sup_{t \in [0,1]} |B(t)|$$

donde $B(t)$ es un puente Browniano. La función de distribución acumulada de K está dada por

$$\Pr(K \leq x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/(8x^2)}.$$

Para más detalles sobre su derivación ver (Massey Jr, 1951)

Bajo la hipótesis nula (la muestra viene de la distribución $F(x)$)

$$\sqrt{n} D_n \xrightarrow{n\to\infty} \sup_t |B(F(t))|$$

en distribución.

La prueba de bondad de ajuste o prueba de Kolmogorov-Smirnov test puede construirse usando los valores críticos de la distribución de Kolmogorov. Esta prueba es asintóticamente válida cuando $n \to \infty$. Se rechaza la hipótesis nula con un nivel α si

$$\sqrt{n}\, D_n > K_\alpha$$

donde K_α es el siguiente cuantil tal que

$$\Pr(K \leq K_\alpha) = 1 - \alpha.$$

La potencia asintótica de la prueba es 1.

Podemos interpretar el modelo con covariables (fijas) y efectos espacio-temporales como un modelo de efectos mixtos donde los efectos asociados a las covariables son los efectos fijos y los efectos espacio-temporales son los aleatorios. Respecto a los fijos, es usual analizar la distribución posterior marginal de los coeficientes de las covariables. En el contexto de INLA, estas distribuciones posteriores se aproximan con una distribución Normal. Sin embargo, buscando la interpretabilidad de la inferencia es conveniente analizar una transformación de los coeficientes. Dado que se tratan de efectos lineales sobre la log-media, al exponenciarlos obtenemos los efectos multiplicativos sobre la media. Más aun, siendo muchos de estos porcentajes, esta nueva variable aleatoria se interpreta como el efecto multiplicativo sobre la media de aumentar en un punto porcentual el efecto en cuestión. Para conocer la densidad de cualquier transformación de una variable aleatoria continua, como es el caso normal, podemos usar el Teorema 2.4.4.

Teorema 2.4.4. *Transformadas de variables aleatorias Sea X una variable aleatoria continua y g una función medible uno a uno. La densidad de la variable aleatoria $Y = g(X)$ es*

$$p_Y(y) = p_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

Nos interesa el caso particular donde $g(\cdot) = \exp(\cdot)$ y $X \sim \text{Normal}(\mu, \sigma)$. Convenientemente, este caso particular ha sido muy estudiado y se dice que $Y \sim g(X) \sim$ log-Normal(μ, σ).

A diferencia del caso Normal, esta distribución no es simétrica respecto a la media y no es tan sencillo encontrar los mejores intervalos o regiones de estimación. Una posibilidad es emplear las llamadas regiones de máxima densidad/ *High Density Region*(HDR). Se puede mostrar que las HDR son los conjuntos de menor longitud que contienen al parámetro θ con probabilidad $1 - \alpha$, pues son la solución Bayesiana óptima al problema de decisión con espacio de decisión $\mathcal{A} = \{B; B \subseteq \Theta\}$ y función de perdida $L(B, \theta) = \alpha\lambda(B) + (1 - \alpha)\mathbb{1}_{B^c}(\theta)$ donde $\lambda(B)$ es la medida de Lebesgue o longitud de la región.

La definición de estas regiones pueden verse en la Definición 2.4.5, la cual se toma de Hyndman, 1996.

Definición 2.4.5. Región de máxima densidad/*High Density Region*(HDR)

Sea $p(\theta)$ la función de densidad de una variable aleatoria Θ. Entonces la región de máxima densidad de probabilidad $1 - \alpha$ (HDR) es el subconjunto $R(p_\alpha)$ del espacio muestral de X tal que

$$R(p_\alpha) = \{\theta : p(\theta) \geq p_\alpha\},$$

donde p_α es la mayor constante tal que

$$P\left(\Theta \in R(p_\alpha)\right) \geq 1 - \alpha.$$

En el paquete de Snow, 2016 existen dos funciones para encontrar estos intervalos o regiones. Si contamos con forma analítica de la marginal y además existe en R su función de cuantiles (como es el caso de la distribución log-Normal) podemos usar la función `hpd(posterior.icdf, conf, tol,...)`. Mientras que si contamos con una muestra de la posterior usamos la función `emp.hpd(x, conf)`. Por ejemplo, podemos calcular los intervalos HDR de probabilidad posterior $1 - \alpha = 0.95$.

Otra alternativa es utilizar una aproximación Normal a una transformada de una variable aleatoria con distribución Normal. Cuando se tiene que una sucesión de variables aleatorias se distribuye asintóticamente Normal, se tiene que una transformación, que cumple ciertas propiedades, también se distribuye asintóticamente Normal. Este resultado, conocido como método Delta, aplica en nuestro caso pues una variable que se distribuye Normal es también asintóticamente Normal.

Definición 2.4.6. Método Delta

Sea $X_1, X_2, ...,$ una sucesión de variables aleatorias tales que

$$\sqrt{n}[X_n - \theta] \xrightarrow{\text{d}} \text{Normal}(0, \sigma^2),$$

donde θ y σ^2 son constantes finitas y $\xrightarrow{D}$ denota convergencia en distribución, entonces

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{\text{d}} \text{Normal}(0, \sigma^2 \cdot [g'(\theta)]^2)$$

para toda función g que satisface la propiedad que $g'(\theta)$ existe y es distinta de cero.

Por lo tanto, si $g(\cdot) = \exp(\cdot)$ y $X \sim \text{Normal}(\mu, \sigma)$, la variable aleatoria $Y = g(X)$ se distribuye asintóticamente $\text{Normal}(\exp(\mu), (\sigma \exp(\mu))^2)$. La ventaja de este enfoque es que es muy sencillo encontrar los HDR, pues se trata de una distribución simétrica respecto a la media donde además la media y la moda coinciden.

CAPÍTULO 3

Aproximación de Laplace anidada integrada

En estadística Bayesiana muchos de los problemas de computo o cálculo se traducen finalmente a problemas de integración. La estrategia que ha ganado más popularidad es aproximar estas integrales mediante aproximaciones de tipo Monte Carlo (simulando mediante Cadenas de Markov). Más aún, teniendo una muestra simulada de los parámetros, el problema de las distribuciones marginales se simplifica al sólo tener que considerar la muestra del parámetro de interés. Sin embargo, obtener esta muestra de distribución posterior de los parámetros puede ser computacionalmente costoso, en especial si el número de parámetros desconocidos es muy grande. Por esto, se vuelve atractivo probar otros métodos de aproximación de integrales, como lo son las aproximaciones analíticas.

Estas aproximaciones analíticas, tales como la aproximación Gaussiana o la aproximación de Laplace, requieren resolver un problema de optimización. Si la función de la densidad distribución posterior es complicada es posible emplear métodos de optimización conocidos. La aproximación de Laplace es equivalente a representar la información del campo latente posterior (condicionado a los datos) mediante un vector de medias y una matriz de precisión. Esta matriz de precisión es originalmente rala por tratarse de un modelo Markoviano y la aproximación Gaussiana se mantiene con muchos ceros. Mediante algoritmos especiales para matrices ralas los cálculos son eficientes y precisos. Posteriormente, mediante herramientas de optimización y selección de puntos para hacer integración numérica, podemos integrar los hiperparámetros. Como siguiente paso, podemos obtener integrales de las componentes del campo aleatorio latente. Finalmente, mediante una pequeña modificación para retirar la información dada por la observación en un estado espacial y temporal específico, podemos aproximar la distribución predictiva de tipo validación cruzada o de dejar uno fuera (en inglés Leave One Out o por sus siglas LOO). Esto permite calcular los criterios de selección de modelos LOO, CPO y PIT. Adicionalmente, basados en el resultado de transformación integral, un criterio de bondad de ajuste es que los PIT se distribuyan como una muestra uniforme estándar y podemos aplicar pruebas de bondad

de ajuste, como Kolmogorov-Smirnov o Chi cuadrada de Pearson, para obtener resúmenes numéricos de ajuste, basados en los estadísticos o en los p-valores de las respectivas pruebas.

3.1. Introducción

Los modelos dentro del marco de INLA se definen como en la Sección 2.1.1, donde $\boldsymbol{y} = (y_1, ..., y_n)$ es el vector de variables de respuesta observadas y la media μ_i está ligada a un predictor lineal η_i usando una función liga apropiada. Este predictor lineal puede incluir coeficientes en covariables (efectos fijos) y distintos tipos de efectos aleatorios. Retomemos la Ecuacion 2.1 que es

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \varepsilon_i, \tag{3.1}$$

donde α es el intercepto, $\beta_j, j = 1, ..., n_\beta$ los coeficientes para algunas covariables $\boldsymbol{z}_i = z_{1,i}, \ldots, z_{n_\beta i}$, las funciones $f^{(k)}$ definen n_f efectos aleatorios sobre covariables $\boldsymbol{u}_i$. Por último ε es un término de error.

El vector de todos los efectos latentes $\boldsymbol{x}$ incluye el predictor lineal, coeficientes asociados a las covariables y los efectos aleatorios, de modo que

$$\boldsymbol{x}_i = (\boldsymbol{\mu}_i, \alpha, \boldsymbol{\beta}_i, \boldsymbol{u}_i).$$

Además de depender de $\boldsymbol{x}_i$, la distribución de y_i dependerá del vector de hiperparámetros $\boldsymbol{\theta}_1$.

Como en la Sección 2.1.1, suponemos que, los efectos latentes $\boldsymbol{x}$ son un Campo Aleatorio Gaussiano Markoviano (GMRF) con media cero y matriz de precisión $\boldsymbol{Q}(\boldsymbol{\theta}_2)$, donde $\boldsymbol{\theta}_2$ es un vector de hiperparámetros. Denotamos el vector de todos los hiperparámetros del modelo como $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Ahora, suponemos que dado el vector de los efectos latentes y el de los hiperparámetros, las observaciones son independientes entre sí. Es decir,

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} p(y_i|\eta_i, \boldsymbol{\theta}),$$

donde η_i es el predictor lineal latente (3.1) y el conjunto $\mathcal{I}$ contiene los índices de todos los valores observados de $\boldsymbol{y}$ (aunque es posible que algunos de los valores no hallan sido observados).

El objetivo de la metodología de INLA es aproximar las posteriores marginales de los efectos e hiperparámetros explotando las propiedades de los GMRF y de la aproximación de Laplace para integración multidimensional.

La distribución posterior conjunta de los efectos y los hiperparámetros puede expresarse como

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} p(y_i|x_i, \boldsymbol{\theta})$$
$$\propto p(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} \exp\left\{ -\frac{1}{2} x^T \boldsymbol{Q}(\boldsymbol{\theta}) x + \sum_{i \in \mathcal{I}} \log(p(y_i|x_i, \boldsymbol{\theta})) \right\}.$$

La notación se simplifica al representar la matriz de precisión de los efectos latentes por $\boldsymbol{Q}(\boldsymbol{\theta})$, mientras que $|\boldsymbol{Q}(\boldsymbol{\theta})|$ denota el determinante de la matriz de precisión.

El cálculo de las distribuciones posteriores marginales de los efectos latentes y los hiperparámetros puede hacerse considerando

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta},$$

y

$$p(\boldsymbol{\theta}_j|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j},$$

donde $\boldsymbol{\theta}_{-j}$ denota el vector $\boldsymbol{\theta}$ excluyendo su entrada j-ésima. Notemos que en ambas expresiones la integración se hace sobre el espacio de los hiperparámetros y que se requiere una buena aproximación de la distribución posterior conjunta. En Rue, Martino & Chopin, 2009 los autores aproximan $p(\boldsymbol{\theta}|\boldsymbol{y})$ en una rejilla, denotando esta aproximación por $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$, la que usan para aproximar la marginal posterior del parámetro latente x_i con

$$\tilde{p}(x_i|\boldsymbol{y}) = \sum_k \tilde{p}(x_i|\boldsymbol{\theta}_k, \boldsymbol{y}) \tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{y}) \Delta_k,$$

donde las Δ_k son los pesos asociados a un vector de valores de $\boldsymbol{\theta}_k$ de los hiperparámetros en la rejilla.

La aproximación $p(\boldsymbol{\theta}_j|\boldsymbol{y})$ puede tomar distintas formas y ser calculada de distintos modos. En Rue et al., 2009 también discuten como esta aproximación debe realizarse para reducir el error numérico.

La importancia de definir la forma en que se hará la inferencia de las variables, o en el caso de INLA, campos latentes, viene de que esta inferencia es un paso intermedio antes de obtener las distribuciones posteriores de los hiperparámetros. Las marginales de las componentes del campo latente:

$$p(x_i|\boldsymbol{y}) = \int \int p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{x}_{-i} d\boldsymbol{\theta}$$
$$= \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, i = 1, \ldots, n$$

Las marginales de los hiperparámetros:

$$\begin{aligned} p(\theta_j|\boldsymbol{y}) &= \int\int p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{\theta}_{-j} \\ &= \int p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, j = 1,\ldots,m \end{aligned}$$

Por otro lado, en algunos contextos es relevante hacer inferencia sobre las variables latentes, como en los modelos de estudios longitudinales (efectos aleatorios) y nos interesa la inferencia de uno o varios individuos particulares.

A continuación describiremos el procedimiento para integrar los hiperparámetros que se usa en INLA. Primero se explora $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$, se localiza la moda, se usa la matriz Hessiana para construir nuevas variables y se hace una búsqueda por rejilla.

Para cada θ_j se evalúa la aproximación de Laplace para valores elegidos de x_i. Se construye una aproximación Skew-Normal o log-spline Normal corregida para representar la densidad marginal condicional. Integramos θ_j Para cada i se suma (integra) $\boldsymbol{\theta}$

$$\tilde{\pi}(x_i|\boldsymbol{y}) \propto \sum_i \tilde{\pi}(x_i|\boldsymbol{y},\theta_j)\tilde{\pi}(\boldsymbol{\theta}_j|\boldsymbol{y})$$

Se construye una distribución Gaussiana corregida por log-spline

$$\text{Normal}(x_i;\mu_i,\sigma_i^2) \times \exp(\text{spline})$$

para representar $\tilde{\pi}(x_i|\boldsymbol{y})$.

Estos casos son particularmente interesantes. Por un lado, según el modelo, las variables observadas dependen de valores fijos que han tomado las variables latentes. Por otro lado, estas son realizaciones aleatorias de la distribución en nivel superior y ya que no se observarán eventualmente deberían tratarse como observaciones faltantes.

En la inferencia Bayesiana esto no supone mucha novedad ya que en general los parámetros son tratados matemáticamente como variables aleatorias, si bien su distribución inicial es la expresión de las creencias subjetivas del tomador de decisiones. Para llevar a cabo la actualización de las creencias mediante los datos, llevada a cabo usualmente a través del Teorema de Bayes, requiere postular una distribución previa de todos los parámetros.

En estadística Bayesiana ésta es la forma de cuantificar la incertidumbre a través de distribuciones condicionales. Sin embargo, en el contexto de modelos jerárquicos es posible cuantificar parte la incertidumbre sin necesidad de postular una previa a los parámetros desconocidos, y sólo aplicar el Teorema de Bayes usando las distribuciones especificadas por el modelo. En general los cálculos pueden ser complicados, pero pueden simplificarse muchos si se adoptan modelos jerárquicos análogos a los modelos Bayesianos con previas conjugadas.

3.2. Implementación INLA

En esta sección, definiremos una clase de modelos ligeramente más especifica/simple que la definida en la Ecuación (3.1). Mencionaremos por qué los campos Gaussianos dan lugar a matrices de precisión ralas y como podemos permutar los vectores aleatorios $\boldsymbol{x}$ para que las componentes distintas de cero estén cerca de la diagonal. Usaremos la aproximación de Laplace para obtener la densidad marginal posterior de los hiperparámetros que después sera explorada de forma numérica y encontrar su moda y algunos puntos de integración adecuados. Además recordaremos la descomposición de Cholesky, introduciremos su versión para matrices ralas y veremos como puede usarse para aproximar la matriz de varianza. Esta a su vez puede usarse para aproximar las distribuciones marginales del Campo Gaussiano Markoviano posterior, ya sea como aproximación Gaussiana o su aproximación de Laplace simplificada. Finalmente usaremos todas las distribuciones marginales posteriores encontradas para calcular distribuciones predictivas de las variables de salida. En especial aquellas de tipo Leave One Out que serviran como criterios de selección de modelos.

3.2.1. Campos de Markov Gaussianos en modelos de efectos mixtos

Si al modelo lineal generalizado con predictor lineal, tal y como esta definido en la Ecuación (3.1), lo restringimos a que las funciones $f^{(k)}$ sean lineales, entonces el predictor lineal tiene la forma

$$\boldsymbol{\eta} = \alpha\mathbf{1} + \boldsymbol{A}\boldsymbol{\beta} + \sum_{j=1}^{m} \boldsymbol{B}_j \boldsymbol{v}_j + \boldsymbol{\varepsilon}$$

donde $\boldsymbol{A}$ es la matriz de covariables de los efectos fijos β, $\boldsymbol{B}_j$ es la matriz de pesos asociada a los efectos aleatorios $\boldsymbol{v}_j$ con $\boldsymbol{v}_j \sim \text{Normal}(\mathbf{0}, \sigma^2\boldsymbol{I})$, de modo que $B_j\boldsymbol{v}_j \sim \text{Normal}(0, \sigma^2 B_j^T B_j)$, y $\boldsymbol{\varepsilon}$ es un posible ruido Gaussiano. A este tipo de modelo donde el predictor lineal depende tanto de efectos fijos como de efectos aleatorios se le conoce como modelo de efectos mixtos. Dado el predictor lineal $\boldsymbol{\eta}$ (contenido en el vector $\boldsymbol{x}$) las observaciones se distribuyen i.i.d.

$$\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{\eta}) = \prod_i p(y_i|\eta_i).$$

La intercambiabilidad pues será útil para entender los modelos que queremos utilizar.

Definición 4. *Intercambiabilidad finita. Las variables aleatorias $X_1, ..., X_n$ son (finitamente) intercambiables bajo una medida de probabilidad P si la distribución inducida por P satisface*

$$p(x_1, ..., x_n) = p(x_{\pi(1)}, ..., x_{\pi(n)})$$

para toda permutación π definida en el conjunto $\{1, 2, ..., n\}$.

Dicho en otras palabras, las "etiquetas" que identifican a cada una de las variables no proporcionan información alguna. Es claro que si las variables aleatorias $X_1, ..., X_n$ son independientes e idénticamente distribuidas entonces son intercambiables.

Definición 5. *Intercambiabilidad infinita. La sucesión infinita de variables aleatorias $X_1, X_2, ...$ es (infinitamente) intercambiable si toda subsucesión finita es intercambiable en el sentido finito.*

En el caso de modelos donde la muestra no es intercambiable el orden de los datos es importante. Ya sea que estén ordenados en un vector o bien en una matriz. El caso de los modelos espacio-temporales pueden presentarse en una estructura matricial, por lo que vale la pena continuar con este caso específico.

Podemos decir que el predictor lineal se compone de

$$\boldsymbol{\eta} = \text{covariables} + \text{efecto fila} + \text{efecto columna} + \text{ruido},$$

es decir,

$$\eta_{ij} = \alpha + \sum_{k=1}^{K} c_{ij}^{(k)} \beta_k + u_i + v_j + w_{ij}$$

con K covariables $\boldsymbol{c}^{(k)}$ (también matrices) y efectos aleatorios $\boldsymbol{u}$, $\boldsymbol{v}$ y $\boldsymbol{w}$. En los modelos espacio-temporales, usualmente $\boldsymbol{u}$ será el efecto espacial Gaussiano, $\boldsymbol{v}$ el efecto temporal Gaussiano y $\boldsymbol{w}$ un ruido Gaussiano. Hasta aquí tenemos la especificación del modelo, al que podemos hacer inferencia mediante un enfoque frecuentista o Bayesiano.

Sea $\mathbb{1}$ un vector con todas sus componentes iguales a 1, $J = \mathbb{1}\mathbb{1}^T$ una matriz con todas sus componentes iguales a 1, $\mathbb{I}$ la matriz identidad y sea vect$(\cdot)$ la operación que transforma una matriz $n \times m$ en un vector nm concatenando sus columnas, entonces el modelo se puede presentar como

I Observables

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} \sim \prod_i p(y_i|\eta_i, \boldsymbol{\theta}_1)$$

II Variables latentes

$$\boldsymbol{\eta}|\alpha, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{u} \sim \text{Normal}\left(\text{vect}\left(\alpha J + \sum_{k=1}^{K} \boldsymbol{C}^{(k)}\boldsymbol{\beta} + \boldsymbol{u}\mathbb{1}^T + \mathbb{1}\boldsymbol{v}^T\right), \tau_w \mathbb{I}\right)$$

$$\boldsymbol{v} \sim \text{Normal}\left(\boldsymbol{0}, \tau_v \boldsymbol{Q}_v(\boldsymbol{\theta}_v)\right)$$

$$\boldsymbol{u} \sim \text{Normal}\left(\boldsymbol{0}, \tau_u \boldsymbol{Q}_u(\boldsymbol{\theta}_u)\right)$$

III Hiperparámetros

$$\boldsymbol{\theta}_1, \alpha, \boldsymbol{\beta}, \tau_v, \tau_u, \tau_w, \boldsymbol{\theta}_v, \boldsymbol{\theta}_u$$

En esta sección empleamos la parametrización de la densidad normal multivariada con matriz de precisión $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. La estructura de las matrices $Q_v(\boldsymbol{\theta}_v)$ y $Q_u(\boldsymbol{\theta}_u)$ es conocida en el sentido de que sus componentes son funciones conocidas de los parámetros $\boldsymbol{\theta}_v$ y $\boldsymbol{\theta}_u$ respectivamente. Esta estructura esta dada por el modelo de efectos aleatorios que elijamos.

Si además asumimos un enfoque Bayesiano podemos asignar previas sobre μ y β. Para poder incorporarlas al campo Gaussiano Markoviano (y no tener que tratarlas como hiperparámetros) conviene asumir que tienen distribución previa normal con media 0. Así $(\alpha, \boldsymbol{\beta}) \sim \text{Normal}(0, \boldsymbol{Q}_\beta)$. Entonces el vector concatenado

$$\boldsymbol{x} = (\alpha, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\eta})$$

es conjuntamente Gaussiano.

Podemos interpretar este modelo Bayesiano como uno jerárquico donde

I Observaciones

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} \sim \prod_i p(y_i|\eta_i, \boldsymbol{\theta}_1)$$

II Campo Latente

$$\boldsymbol{x}|\boldsymbol{\theta} \sim p(\boldsymbol{x}|\boldsymbol{\theta}) = \text{Normal}(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\theta}_2))$$

III Hiperparámetros

$$(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \sim p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \text{ con } \boldsymbol{\theta}_2 = (\tau_v, \tau_u, \tau_w, \boldsymbol{\theta}_v, \boldsymbol{\theta}_u)$$

Debido las estrategias algorítmicas empleadas en INLA para obtener las varianzas marginales del campo Gaussiano Markoviano la dimensión de $\boldsymbol{x}$ puede ser grande, del orden de 10^2 a 10^5, mientras que la dimensión de $\boldsymbol{\theta}$ debe ser pequeña, de 1 a 5, pues se usan estrategias de integración estándar (regla de Simpson o del trapezoide) para obtener estas distribuciones marginales.

3.2.2. Obtener matriz de precisión previa dado el modelo condicional

La matriz de precisión del campo latente

$$\boldsymbol{Q}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$$

se define como la inversa de la matriz de varianzas y covarianzas ($\Sigma(\boldsymbol{\theta})$). Podríamos decir que en términos prácticos que los Campos Aleatorios de Markov Gaussianos (GMRFs) son medidas Gaussianas con matriz de precisión rala. Modelos que involucran estos campos aleatorios tiene buenas propiedades computacionales mediante algoritmos para matrices ralas. Resultan muy útiles también para hacer inferencia basada en algoritmos MCMC. Un GMRF tiene una construcción simple

- Un vector con distribución normal

$$\boldsymbol{x} = (x_1, ..., x_n)^T$$

- Propiedad de Markov adicional

$$x_i \perp\!\!\!\perp x_j | \boldsymbol{x}_{-ij}$$

  es decir, x_i y x_j son condicionalmente independientes.

Si $x_i \perp\!\!\!\perp x_j | \boldsymbol{x}_{-ij}$ para un conjunto de $\{i, j\}$, entonces necesitamos restringir la parametrización del GMRF. Para esto, trabajar con la matriz de precisión es mucho más fácil y eficiente que hacerlo con la de varianza, que no necesariamente es rala.

La función de densidad de una distribución Gaussiana con media cero y matriz de precisión $\boldsymbol{Q}$ corresponde a

$$p(\boldsymbol{x}) \propto |\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^t \boldsymbol{Q} \boldsymbol{x}\right)$$

Como nos muestra el siguiente teorema, efectivamente el espacio parametral que cumple con la independencia condicional considera matrices de precisión ralas.

Teorema 3.2.1. *Independencia condicional implica ceros en matriz de precisión (y viceversa)*

Sea $\boldsymbol{x}$ un Campo Aleatorio Gaussiano Markoviano,

$$x_i \perp\!\!\!\perp x_j | \boldsymbol{x}_{-ij} \iff Q_{ij} = 0.$$

Podemos mencionar cierto paralelismo interesante entre las matrices $\boldsymbol{Q}$ y $\boldsymbol{\Sigma}$. Mientras que ceros en la matriz implica independencia condicional (dado el resto de las variables), ceros en la matriz $\boldsymbol{Q}$ implica independencia en su sentido usual, que podemos llamar independencia marginal. Además, notemos que conocer la matriz $\boldsymbol{\Sigma}$ implica conocer las distribuciones marginales de X_i respecto a $\boldsymbol{X}_{-i}$, es decir, respecto a resto del campo Gaussiano Markoviano. De este modo, sólo resta integrar los hiperparámtros para obtener las distribuciones marginales respecto a todas las cantidades desconocidas.

Por el Teorema 3.2.1, la matriz de precisión de un Campo Gaussiano Markoviano es rala. Esto es importante pues trae consigo dos beneficios. Por un lado es más fácil construir modelos condicionando (modelos jerárquicos) y por otro tiene beneficios computacionales. Sin embargo, en general no es sencillo obtener para un modelo jerárquico cualquiera la matriz de precisión Q de forma explícita.

Algunos componentes como el predictor lineal $\boldsymbol{\eta}$ pueden agregarse fácilmente si ya se conoce la matriz de precisión del resto de las variables. Podemos hacer uso del siguiente resultado:

Teorema 3.2.2. *Construir matriz de precisión en efectos aditivos Gaussianos*

Si

$$\boldsymbol{X} \sim \textit{Normal}(\boldsymbol{0}, \boldsymbol{Q}_{\boldsymbol{X}}^{-1})$$

$$y\ \boldsymbol{Y}|\boldsymbol{X} \sim \textit{Normal}(\boldsymbol{X}, \boldsymbol{Q}_{\boldsymbol{Y}}^{-1})$$

entonces $(\boldsymbol{X}, \boldsymbol{Y}) \sim \textit{Normal}(\mathbf{0}, \boldsymbol{Q}^{-1}_{\boldsymbol{X},\boldsymbol{Y}}))$ *con*

$$\boldsymbol{Q}_{\boldsymbol{X},\boldsymbol{Y}} = \begin{bmatrix} \boldsymbol{Q}_{\boldsymbol{X}} + \boldsymbol{Q}_{\boldsymbol{Y}} & -\boldsymbol{Q}_{\boldsymbol{Y}} \\ -\boldsymbol{Q}_{\boldsymbol{Y}} & \boldsymbol{Q}_{\boldsymbol{Y}} \end{bmatrix}$$

Demostración. La densidad conjunta es

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \mathbf{0})^T Q_{\boldsymbol{X}} (\boldsymbol{x} - \mathbf{0}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T Q_{\boldsymbol{Y}} (\boldsymbol{y} - \boldsymbol{x}) \right\} \\ &\propto \exp\left\{ -\frac{1}{2}\Big(\boldsymbol{x}^T Q_{\boldsymbol{X}} \boldsymbol{x} + \boldsymbol{y}^T Q_{\boldsymbol{Y}} \boldsymbol{y} - \boldsymbol{x}^T Q_{\boldsymbol{Y}} \boldsymbol{y} - \boldsymbol{y}^T Q_{\boldsymbol{Y}} \boldsymbol{x} + \boldsymbol{x}^T Q_{\boldsymbol{Y}} \boldsymbol{x} \Big) \right\} \\ &\propto \exp\left\{ -\frac{1}{2}\left(\begin{bmatrix} \boldsymbol{x}^T & \boldsymbol{y}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{\boldsymbol{X}} + \boldsymbol{Q}_{\boldsymbol{Y}} & -\boldsymbol{Q}_{\boldsymbol{Y}} \\ -\boldsymbol{Q}_{\boldsymbol{Y}} & \boldsymbol{Q}_{\boldsymbol{Y}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \right) \right\} \end{aligned}$$

■

Queda, por último, explicar como se obtienen los bloques de la matriz Q correspondientes a un segmento de $\boldsymbol{x}$ asociado a un efecto aleatorio (por ejemplo v). Los efectos aleatorios usualmente se definen a través de la especificación de las condicionales completas. Por la propiedad de Markov, estas sólo dependen de los vecinos.

En general, esta distribución puede calcularse a través de las distribuciones condicionales completas mediante el Teorema 3.2.3 ver Robert & Casella, 2013.

Teorema 3.2.3. *Hammersley-Clifford (Como en Robert & Casella)*

La distribución conjunta asociada a las densidades condicionales $p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})$ *y* $p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ *tiene densidad conjunta*

$$p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})}{\int [p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) / p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})] d\boldsymbol{y}}$$

Demostración. Sabemos que

$$p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) = p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) p_{\boldsymbol{Y}}(\boldsymbol{y}) = p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}),$$

por lo que

$$\frac{p_{\boldsymbol{Y}}(\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})}{p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})}.$$

Integrando respecto a $\boldsymbol{y}$

$$\int \frac{p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})}{p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})} d\boldsymbol{y} = \int \frac{p_{\boldsymbol{Y}}(\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x})} d\boldsymbol{y} = \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x})}.$$

Podemos obtener de forma explícita la densidad conjunta a partir de las condicionales com-

pletas mediante

$$p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\boldsymbol{X}}(\boldsymbol{x})p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})}{\int[p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})/p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})]d\boldsymbol{y}}$$

■

El problema es que este cálculo formal supone que todos los objetos involucrados existen, hecho que no siempre es cierto. Además, debido a que trabajamos con el caso Gaussiano, nos interesara tener una expresión más sencilla de calcular estas distribuciones conjuntas que aproveche que estamos en este caso. Podríamos decir que nos interesa tener el inverso al famoso resultado de las distribución condicional para una distribución normal multivariada.

Teorema 3.2.4. *Distribuciones condicionales de una normal multivariada.*
Sea $\boldsymbol{X}$ un vector aleatorio Gaussiano de dimensión N

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_p \\ \boldsymbol{X}_q\end{bmatrix} \text{ de dimensión } \begin{bmatrix} r \times 1 \\ (N-r)\times 1\end{bmatrix}$$

con $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ particionados de la siguiente forma

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q\end{bmatrix} \text{ de dimensión } \begin{bmatrix} r \times 1 \\ (N-r)\times 1\end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{pp} & \boldsymbol{\Sigma}_{pq} \\ \boldsymbol{\Sigma}_{qp} & \boldsymbol{\Sigma}_{qq}\end{bmatrix} \text{ de dimensión } \begin{bmatrix} r \times r & r \times (N-r) \\ (N-r)\times r & (N-r)\times(N-r)\end{bmatrix}$$

entonces la distribución de $\boldsymbol{X}_p$ condicionado en $\boldsymbol{X}_q = \boldsymbol{a}$ es una Normal ($\bar{\boldsymbol{\mu}}_p, \overline{\boldsymbol{\Sigma}}_p$) donde

$$\bar{\boldsymbol{\mu}}_p = \boldsymbol{\mu}_p + \boldsymbol{\Sigma}_{pq}\boldsymbol{\Sigma}_{qq}^{-1}\left(\mathbf{a} - \boldsymbol{\mu}_q\right)$$

y matriz de varianzas y covarianzas

$$\overline{\boldsymbol{\Sigma}}_p = \boldsymbol{\Sigma}_{pp} - \boldsymbol{\Sigma}_{pq}\boldsymbol{\Sigma}_{qq}^{-1}\boldsymbol{\Sigma}_{qp}.$$

Es decir, queremos obtener $\mu_p, \mu_q, \boldsymbol{\Sigma}_{pp}, \boldsymbol{\Sigma}_{pq}, \boldsymbol{\Sigma}_{qp}$ y $\boldsymbol{\Sigma}_{qq}$ a partir de $\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\mu}}_q, \overline{\boldsymbol{\Sigma}}_p$ y $\overline{\boldsymbol{\Sigma}}_q$. No hay un resultado general que sirva para todos los casos, por lo que para cada efecto aleatorio debe examinarse por separado para obtener la forma de la matriz Q.

Si la relación de dependencia de las variables aleatorias puede representarse con un grafo dirigido acíclico o *directed acyclic graph* (DAG), la construcción de la función de densidad conjunta puede hacerse mediante el producto de las densidades condicionales, pero aún faltaría obtener las distribuciones marginales.

Consideremos un ejemplo donde es fácil obtener la matriz de varianzas y covarianzas, para luego invertirla y obtener la matriz de precisión.

Ejemplo 3.2.5. *Autorregresivo de orden* 1

Sea

$$X_t|X_{t-1},...,X_1 \sim \textit{Normal}(\phi X_{t-1}, \tau^{-1}), \textit{ para } t = 2,...,n$$

y $X_1 \sim \textit{Normal}(0, \tau^{-1}(1-\phi^2)^{-1})$.

Mostraremos que el vector $\boldsymbol{X} = X_1, \ldots, X_t \sim \textit{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$, *con*

$$\boldsymbol{\Sigma} = \frac{\tau^{-1}}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \ldots & \phi^{t-1} \\ \phi & 1 & \phi & \ldots & \phi^{t-2} \\ \phi^2 & \phi & 1 & \ldots & \phi^{t-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{t-1} & \phi^{t-2} & \phi^{t-3} & \ldots & 1 \end{pmatrix}$$

Notemos que $X_2 = \phi X_1 + \varepsilon$, *con* $\varepsilon \sim \textit{Normal}(0, \tau^{-1})$. *Pero* $\phi X_1 \sim \textit{Normal}\left(0, \phi^2\left(\frac{\tau^{-1}}{1-\phi^2}\right)\right)$. *Como* $\phi X_1 \perp\!\!\!\perp \varepsilon$, *entonces* $X_2 \sim \textit{Normal}\left(0, \tau^{-1}\left(\frac{\phi^2}{1-\phi^2}+1\right)\right)$. *Pero* $\frac{\phi^2}{1-\phi^2}+1 = \frac{\phi^2+1-\phi^2}{1-\phi^2} = \frac{1}{1-\phi^2}$. *Así,* $\text{Var}(X_2) = \frac{\tau^{-1}}{1-\phi^2}$. *Por inducción,* $\text{Var}(X_i) = \frac{\tau^{-1}}{1-\phi^2}$ *para* $i = 1,...,t$.

Por otro lado,

$$\begin{aligned} \text{Cov}(X_1, X_2) &= \text{Cov}(X_1, \phi X_1 + \varepsilon) \\ &= \text{Cov}(X_1, \phi X_1) + \text{Cov}(X_1, \varepsilon) \\ &= \phi \text{Cov}(X_1, X_1) + 0 \\ &= \phi \text{Var}(X_1) \end{aligned}$$

Por inducción se puede mostrar que $\text{Cov}(X_1, X_j) = \phi^{j-1}\text{Var}(X_1)$. *Pero por simetría de la matriz de varianzas y covarianzas* $\text{Cov}(X_j, X_1) = \phi^{j-1}\text{Var}(X_1)$. *De nuevo por inducción se puede mostrar que* $\text{Cov}(X_i, X_j) = \phi^{j-i}\text{Var}(X_i)$ *si* $i < j$ *y* $\text{Cov}(X_i, X_j) = \phi^{j-i}\text{Var}(X_i)$ *si* $i > j$.

Como $\text{Var}(X_i) = \frac{\tau^{-1}}{1-\phi^2}$ *para* $i = 1,...,t$, *entonces*

$$\Sigma_{ij} = \frac{\tau^{-1}}{1-\phi^2}\phi^{|j-i|}$$

Para obtener $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ *debemos invertir esta matriz. En general, hacer esto es complicado, pero se puede aprovechar que se trata de una matriz simétrica de Toeplitz. En Dow (2002) se les conoce como matriz de Kac-Murdock-Szego a la matriz simétrica* A *como aquella que es de Toeplitz y cuyas componentes son de la forma*

$$A_{ij} = \rho^{|i-j|} \textit{ con } i, j = 1,...,n.$$

Su inversa es una matriz tridiagonal de la forma

$$A^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \dots & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots & 0 & -\rho & 1 \end{bmatrix}$$

Como

$$\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1} = \left(\frac{\tau^{-1}}{1-\phi^2}\right)^{-1}\left[\phi^{|j-i|}\right]^{-1}$$

concluimos que la matriz de precisión es

$$\boldsymbol{Q} = \tau\begin{pmatrix} 1 & -\phi & 0 & \dots & 0 \\ -\phi & 1+\phi^2 & -\phi & \dots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \dots & -\phi & 1+\phi^2 & -\phi \\ 0 & \dots & 0 & -\phi & 1 \end{pmatrix}$$

Destaquemos que por el Teorema 3.2.1, sabíamos que las componentes fuera de la tridiagonal son 0*, pues componentes no consecutivas son condicionalmente independientes.*

Un caso más complicado es donde las relaciones de dependencia condicional no definen un grafo acíclico. Un ejemplo de esto es el modelo de Besag. Las distribuciones condicionales (completas) del modelo Besag (cuyo caso particular es la caminata aleatoria de orden 1) para v_i es

$$v_i|\boldsymbol{v}_{-i}, \tau_v \sim \text{Normal}\left(\frac{1}{n_{\mathcal{N}(i)}}\sum_{j\in\mathcal{N}(i)} v_i, \frac{1}{n_{\mathcal{N}(i)}\tau_v}\right),$$

donde τ_v es el parámetro de precisión, $\boldsymbol{v}_{-i} = (v_1, \dots, v_{i-1}, v_{i+1}, v_n)$, $\mathcal{N}(i)$ son los vecinos de v_i y $n_{\mathcal{N}(i)}$ es el número de elementos en $\mathcal{N}(i)$, es decir, la cardinalidad del conjunto $\mathcal{N}(i)$. En general, tener distribuciones condicionales normales no implica tener una distribución conjunta normal multivariada. Veremos que las distribuciones condicionales del modelo Besag en particular sí lo cumple y la estructura de la matriz de precisión es conocida.

La densidad conjunta para $\boldsymbol{v}$ (ver Riebler, Sørbye, Simpson & Rue, 2016) es

$$p(\boldsymbol{v}|\tau_v) \propto \exp\left(-\frac{\tau_v}{2}\sum_{i\sim j}(v_i - v_j)^2\right) \propto \exp\left(-\frac{\tau_v}{2}\boldsymbol{v}^T\boldsymbol{G}\boldsymbol{v}\right),$$

donde la matriz $\boldsymbol{G}$ tiene componentes

$$G_{ij} = \begin{cases} n_{\mathcal{N}(i)} & i = j \\ -1 & i \sim j \\ 0 & \text{en otro caso.} \end{cases}$$

Definimos la relación de equivalencia $\sim$ tal que $i \sim j$ si y sólo sí $j \in \mathcal{N}(i)$. En otras palabras, estamos diciendo que $\boldsymbol{v}$ se distribuye Normal$(\mathbf{0}, (\tau_v \boldsymbol{G})^{-1})$. Podemos aplicar estos resultados para conocer la forma de la matriz de precisión $\boldsymbol{Q}$ de un caso particular útil para series de tiempo.

Ejemplo 3.2.6. *Caminata aleatoria de orden* 1

Sea

$$v_t|v_{-t} \sim \textit{Normal}(v_{t+1}, \tau_v) \textit{ para } t = 1$$
$$v_t|v_{-t} \sim \textit{Normal}\left(\frac{v_{t+1} + v_{t-1}}{2}, \frac{\tau_v}{2}\right) \textit{ para } t = 2, ..., T-1$$
$$v_t|v_{-t} \sim \textit{Normal}(v_{t-1}, \tau_v) \textit{ para } t = T.$$

Entonces, $\boldsymbol{v} \sim \textit{Normal}(\mathbf{0}, \boldsymbol{Q})$ *con*

$$\boldsymbol{Q} = \tau_v \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \dots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}.$$

3.2.3. Forma explícita de la aproximación de Laplace en INLA

Notemos que en realidad interesa obtener una aproximación de la función de verosimilitud que es complicada pues involucra integrar el campo Gaussiano. Nos limitaremos a demostrar que emplear la aproximación de Laplace clásica es equivalente a la aproximación empleada en INLA donde se emplea una aproximación Gaussiana del denominador.

Nos interesa evaluar

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Necesitamos una aproximación de $p(\boldsymbol{y}|\boldsymbol{\theta})$ que es una integral pues

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\boldsymbol{x}} p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} = \int_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}.$$

Para emplear la aproximación de Laplace notemos que esta integral es de la forma

$$I \equiv p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\boldsymbol{x}} \exp\left\{ n \left(\frac{1}{n} \log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) + \frac{1}{n} \log p(\boldsymbol{x}|\boldsymbol{\theta}) \right) \right\} d\boldsymbol{x}.$$

Recordemos que la forma de las integrales de la aproximación de Laplace clásica es

$$I = \int q(\boldsymbol{\theta}) \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}$$

donde $q : \mathbb{R}^d \to \mathbb{R}$ y $h : \mathbb{R}^d \to \mathbb{R}$ son funciones suaves de $\boldsymbol{\theta}$. Supongamos que $h(\cdot)$ tiene un mínimo en $\hat{\boldsymbol{\theta}}$.

Como vimos en la Sección 2.2, el método de Laplace aproxima I a través de

$$\hat{I} = q(\hat{\boldsymbol{\theta}})(2\pi/n)^{d/2}|\Sigma(\hat{\boldsymbol{\theta}})|^{1/2}\exp\{-nh(\hat{\boldsymbol{\theta}})\},$$

donde

$$\Sigma(\hat{\boldsymbol{\theta}}) = \left\{\frac{\partial^2 h(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T\partial\boldsymbol{\theta}}\right\}^{-1}.$$

En este caso $q(\boldsymbol{\theta}) \equiv 1$ y $h(\boldsymbol{\theta}) = \frac{1}{n}\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{x}|\boldsymbol{\theta})$.

Para igualar la notación denotaremos $\boldsymbol{x}_0 \equiv \hat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ y también denotaremos

$$f(\boldsymbol{x}) \equiv \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}).$$

Observemos que

$$\log p(\boldsymbol{x}|\boldsymbol{\theta}) \propto -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu}) \propto -\frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{\mu}^T\boldsymbol{Q}\boldsymbol{x} - x^T\boldsymbol{Q}\boldsymbol{\mu}\right) \tag{3.2}$$

Conviene recordar algunas reglas de derivación vectorial tales como

$$\begin{aligned}\frac{d}{d\boldsymbol{x}}b^T\boldsymbol{x} &= b\\ \frac{d}{d\boldsymbol{x}}\boldsymbol{x}^Tb &= b^T\\ \frac{d}{d\boldsymbol{x}}\boldsymbol{x}^TA\boldsymbol{x} &= (A+A^T)\boldsymbol{x}\end{aligned}$$

Entonces

$$\begin{aligned}\frac{d}{d\boldsymbol{x}}h(\boldsymbol{x}) &= \frac{1}{n}f'(\boldsymbol{x}) - \frac{1}{n}\frac{1}{2}\left((\boldsymbol{Q}+\boldsymbol{Q}^T)\boldsymbol{x} - \boldsymbol{Q}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{Q}^T\right)\right)\\ &= \frac{1}{n}f'(\boldsymbol{x}) - \frac{1}{n}\left(\boldsymbol{Q}\boldsymbol{x} - \frac{1}{2}\left(\boldsymbol{Q}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{Q}^T\right)\right)\\ &= \frac{1}{n}f'(\boldsymbol{x}) - \frac{1}{n}\left(\boldsymbol{Q}\boldsymbol{x} - \frac{1}{2}\left(\boldsymbol{Q}\boldsymbol{\mu} + (\boldsymbol{Q}\boldsymbol{\mu})^T\right)\right)\\ &= \frac{1}{n}f'(\boldsymbol{x}) - \frac{1}{n}\left(\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right).\end{aligned}$$

La condición necesaria de primer orden de los máximos es $h'(\boldsymbol{x}) = 0$ y esto se cumple si y sólo si

$$f'(\boldsymbol{x}) = \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})$$

Calculando la matriz de segundas derivadas

$$\begin{aligned}\frac{d^2}{(d\boldsymbol{x})^2}h(\boldsymbol{x}) &= \frac{d}{d\boldsymbol{x}}\left[\frac{1}{n}f'(\boldsymbol{x}) - \frac{1}{n}\left(\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right] \\ &= \frac{1}{n}\text{diag}(f''(\boldsymbol{x})) - \frac{1}{n}\boldsymbol{Q}^T \\ &= \frac{1}{n}\text{diag}(f''(\boldsymbol{x})) - \frac{1}{n}\boldsymbol{Q}^T\end{aligned}$$

Entonces, por la aproximación clásica de Laplace

$$\begin{aligned}\hat{I} &= q(\hat{\boldsymbol{\theta}})(2p/n)^{d/2}|\Sigma(\hat{\boldsymbol{\theta}})|^{1/2}\exp\{-nh(\hat{\boldsymbol{\theta}})\} \\ &= (2p/n)^{d/2}\left|\left(\frac{1}{n}(\text{diag}(f''(\boldsymbol{x}_0)) - \boldsymbol{Q}^T)\right)^{-1}\right|^{1/2}\exp\left\{\left(f(\boldsymbol{x}_0) + -\frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}_0-\boldsymbol{\mu})\right)\right\} \\ &= (2p/n)^{d/2}\left|\left(\frac{1}{n}(\text{diag}(f''(\boldsymbol{x}_0)) - \boldsymbol{Q}^T)\right)^{-1}\right|^{1/2}\exp\left\{-nh(\boldsymbol{x}_0)\right\} \\ &= (2p)^{d/2}\left|\left((\text{diag}(f''(\boldsymbol{x}_0)) - \boldsymbol{Q}^T)\right)^{-1}\right|^{1/2}\exp\left\{-nh(\boldsymbol{x}_0)\right\}.\end{aligned}$$

Se puede demostrar que la aproximación clásica de Laplace da la misma ecuación que la aproximación

$$p(\boldsymbol{y}|\boldsymbol{\theta}) \approx \left.\frac{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})}{p_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}\right|_{x=\text{xmoda}(\boldsymbol{\theta})}$$

donde $p_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ es una aproximación Gaussiana.

Proposición 2. *El término $p_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ es una aproximación Gaussiana con media y matriz de covarianzas:*

$$\text{E}_{\boldsymbol{x}_0}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \approx \left(\boldsymbol{Q} - \textit{diag}(f''(\boldsymbol{x}_0))\right)^{-1}\left(\boldsymbol{Q}\boldsymbol{\mu} + f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0\right)$$

$$\text{Var}_{\boldsymbol{x}_0}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \approx \left(\boldsymbol{Q} - \textit{diag}(f''(\boldsymbol{x}_0))\right)^{-1},$$

donde $\boldsymbol{x}_0 = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$.

Demostración. Observemos que para observaciones no Gaussianas tenemos que

$$\log p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{\theta}) + c.$$

Usando una aproximación de Taylor de $f(\boldsymbol{x}) = \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$ alrededor de x_0 y (3.2) obtenemos la aproximación Gaussiana $p_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$.

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^T f'(\boldsymbol{x}_0) + \frac{(\boldsymbol{x}-\boldsymbol{x}_0)^T f''(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)}{2!}.$$

Luego

$$
\begin{aligned}
&\log p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \\
&= \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{\theta}) + c \\
&\propto \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) - (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&\propto f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^T f'(\boldsymbol{x}_0) + \frac{(\boldsymbol{x}-\boldsymbol{x}_0)^T f''(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)}{2!} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&\propto -\frac{1}{2}\left\{-2f(\boldsymbol{x}_0) - 2(\boldsymbol{x}-\boldsymbol{x}_0)^T f'(\boldsymbol{x}_0) - (\boldsymbol{x}-\boldsymbol{x}_0)^T f''(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
&= -\frac{1}{2}\left\{-2f(\boldsymbol{x}_0) - 2\boldsymbol{x}^T f'(\boldsymbol{x}_0) + \boldsymbol{x}_0^T f'(\boldsymbol{x}_0) - \boldsymbol{x}^T f''(\boldsymbol{x}_0)\boldsymbol{x} + 2\boldsymbol{x}^T f''(\boldsymbol{x}_0)\boldsymbol{x}_0 - \boldsymbol{x}_0^T f''(\boldsymbol{x}_0)\boldsymbol{x}_0 + \right. \\
&\quad \left. \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{Q}\boldsymbol{\mu}\right\} \\
&\propto -\frac{1}{2}\left\{\boldsymbol{x}^T(\boldsymbol{Q} - f''(\boldsymbol{x}_0))\boldsymbol{x} - 2\boldsymbol{x}^T(f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{\mu})\right\} \\
&\propto -\frac{1}{2}\Big\{ \left[\boldsymbol{x}^T - (\boldsymbol{Q} - f''(\boldsymbol{x}_0))^{-1}(f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{\mu})\right]^T \Big(\boldsymbol{Q} - f''(\boldsymbol{x}_0)\Big) \\
&\quad \left[\boldsymbol{x}^T - (\boldsymbol{Q} - f''(\boldsymbol{x}_0))^{-1}(f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{\mu})\right] \Big\}.
\end{aligned}
$$

Así

$$\boldsymbol{X}_G|\boldsymbol{y},\boldsymbol{\theta} \sim \text{Normal}((\boldsymbol{Q} - f''(\boldsymbol{x}_0))^{-1}(f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{\mu}), (\boldsymbol{Q} - f''(\boldsymbol{x}_0))^{-1}).$$

■

Por otro lado, la relación

$$p(\boldsymbol{y}|\boldsymbol{\theta}) \approx \frac{p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})}{p_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})},$$

equivale

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) \approx \log(p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})) + \log(p_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})).$$

En particular, si $x = x_0$,

$$
\begin{aligned}
&\log p(\boldsymbol{y}|\theta) = \\
&\quad = \log(p(\boldsymbol{x}_0, \boldsymbol{y}|\boldsymbol{\theta})) + \log(p_G(\boldsymbol{x}_0, \boldsymbol{y}|\boldsymbol{\theta})) \\
&\quad = \log(p(\boldsymbol{y}|\boldsymbol{x}_0, \boldsymbol{\theta})) + \log(p(\boldsymbol{x}_0|\boldsymbol{\theta})) - \log(p_G(\boldsymbol{x}_0, \boldsymbol{y}|\boldsymbol{\theta})) \\
&\quad = f(\boldsymbol{x}_0) + (d/2)\log(2p) + 1/2(\log(|Q|)) + (x_0 - \mu)^T Q(x_0 - \mu) - (d/2)\log(2p) \\
&\qquad -1/2(\log(|(Q^{-1} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}|)) + \\
&\qquad \Big(\boldsymbol{x}_0 - (\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}(\boldsymbol{Q}\mu + f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0)\Big)^T \Big(\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0))\Big) \\
&\qquad \Big(\boldsymbol{x}_0 - (\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}(\boldsymbol{Q}\mu + f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0)\Big) \\
&\quad = -nh(\boldsymbol{x}_0) - (d/2)\log(2p) - 1/2(\log(|(Q^{-1} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}|)) - \\
&\qquad ((\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))\, x_0 - \boldsymbol{Q}\mu - f'(\boldsymbol{x}_0) + f''(\boldsymbol{x}_0)\boldsymbol{x}_0)^T (\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1} \\
&\qquad ((\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))\, x_0 - \boldsymbol{Q}\mu - f'(\boldsymbol{x}_0) + f''(\boldsymbol{x}_0)\boldsymbol{x}_0)\,.
\end{aligned}
$$

Recordemos que por condición de primer orden de $h(\boldsymbol{x})$ todo máximo debe cumplir

$$f'(x_0) = \boldsymbol{Q}(x_0 - \mu).$$

Luego

$$
\begin{aligned}
&\log p(\boldsymbol{y}|\boldsymbol{\theta}) = \\
&\quad = -nh(\boldsymbol{x}_0) - (d/2)\log(2p) - 1/2(\log(|(\boldsymbol{Q}^{-1} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}|)) - \\
&\qquad \Big((\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))\, x_0 - \boldsymbol{Q}\mu - f'(\boldsymbol{x}_0) + f''(\boldsymbol{x}_0)\boldsymbol{x}_0\Big)^T (\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1} \\
&\qquad \Big((\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))\, x_0 - \boldsymbol{Q}\mu - f'(\boldsymbol{x}_0) + f''(\boldsymbol{x}_0)\boldsymbol{x}_0\Big) \\
&\quad = -nh(\boldsymbol{x}_0) - (d/2)\log(2p) - 1/2(\log(|(\boldsymbol{Q}^{-1} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}|)) - \\
&\qquad (\boldsymbol{Q}(x_0 - \mu) - f'(\boldsymbol{x}_0))^T (\boldsymbol{Q} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1} (\boldsymbol{Q}(x_0 - \mu) - f'(\boldsymbol{x}_0)) \\
&\quad = -nh(\boldsymbol{x}_0) + (d/2)\log(2p) + 1/2(\log(|(\boldsymbol{Q}^{-1} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}|)) - 0.
\end{aligned}
$$

En conclusión

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) \approx \exp\{-nh(\boldsymbol{x}_0)\}(2p)^{(d/2)}|(\boldsymbol{Q}^{-1} - \text{diag}(f''(\boldsymbol{x}_0)))^{-1}|^{1/2}.$$

Por lo tanto, hemos demostrado que la aproximación clásica de Laplace y la aproximación de marginales mediante la aproximación Gaussiana del denominador son equivalentes.

3.2.4. Optimizar la densidad posterior de los hiperparámetros

Recordemos que un campo Markoviano Gaussiano latente (Sección 1.4.3) es un modelo con la siguiente estructura

I Datos observados $\boldsymbol{y}$, $y_i|x_i \sim p(y_i|x_i, \boldsymbol{\theta})$

II Campo Gaussiano $\boldsymbol{x} \sim \text{Normal}(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\theta})^{-1})$

III Hiperparámetros $\boldsymbol{\theta}$

Los hiperparámetros $\boldsymbol{\theta}$ pueden incluir los parámetros en la verosimilitud $p(\boldsymbol{y}|\boldsymbol{x})$. Además, si aparecen en la matriz Q pueden afectar la fuerza o extensión de la dependencia, así como la variabilidad del modelo. La densidad conjunta es de la forma

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} p(y_i|x_i, \boldsymbol{\theta})$$

y las primeras marginales que nos interesan son

$$p(\theta_i|\boldsymbol{y}), \text{ para todas o algunas } i$$

Se requiere poder integrar estas marginales para obtener las marginales de la forma

$$p(x_i|\boldsymbol{y}), \text{ para todas o algunas } i.$$

A su vez, estas se usan para calcular distribuciones predictivas, que de nuevo, son distribuciones marginales.

Podemos calcular (aproximadamente) las marginales directamente sin tener que usar MCMC. Mediante aproximaciones analíticas ganamos velocidad y precisión pues no son métodos computacionalmente intensivos. Aprovechamos completamente la estructura de los modelos Gaussianos latentes.

En el caso de tener observaciones Gaussianas no hace falta recurrir a la aproximación de Laplace, pero siguen siendo útiles todos los algoritmos que hemos introducido para integrar de forma eficiente el campo Markoviano Gaussiano latente.

Como

$$\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta} \sim \text{Normal}(\cdot, \cdot)$$

podemos calcular (numéricamente) todas las marginales, usando que

$$p(\theta|\boldsymbol{y}) \propto \frac{\overbrace{p(\boldsymbol{x}, \boldsymbol{y}|\theta)}^{\text{Gaussiana}}\, p(\boldsymbol{\theta})}{\underbrace{p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})}_{\text{Gaussiana}}}.$$

Como

$$\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta} \sim \text{Normal}(\cdot, \cdot)$$

entonces

$$p(x_i|\boldsymbol{y}) = \int \underbrace{p(x_i|\boldsymbol{\theta}, \boldsymbol{y})}_{\text{Gaussiana}} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}.$$

En el caso de modelo con observaciones no Gaussianas, el procedimiento no es muy distinto del caso de observaciones Gaussianas. Lo que cambia es que la distribución condicional posterior del campo latente $p(\boldsymbol{x}|\boldsymbol{y}, \theta)$ debe aproximarse mediante la llamada aproximación Gaussiana. Afortunadamente, esta aproximación solo requiere modificar en la diagonal la matriz de precisión $\boldsymbol{Q}$ del campo latente. La modificación en la matriz de covarianzas $\boldsymbol{\Sigma}$ no es tan sencilla, y por ello trabajamos con la de precisión $\boldsymbol{Q}$. Sin embargo, nos interesa obtener las marginales del campo latente, que es equivalente a obtener la matriz de covarianza $\boldsymbol{\Sigma}$ a partir de la matriz de precisión modificada.

En este caso

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\overbrace{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}^{\text{No Gaussiana conocida}} \quad p(\boldsymbol{\theta})}{\underbrace{p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})}_{\text{No Gaussiana desconocida}}}.$$

Por lo que empleamos una aproximación Gaussiana para la condicional

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \approx c \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p_G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})} \Bigg|_{x=\text{xmoda}(\boldsymbol{\theta})}$$

donde c es una constante de normalización.

Para las marginales de $x_i|\boldsymbol{y}, \boldsymbol{\theta}$, podemos aproximar de manera similar

$$p(x_i|\boldsymbol{y}, \boldsymbol{\theta}) \approx c \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{p_G(\boldsymbol{x}_{-i}|x_i, \boldsymbol{y}, \boldsymbol{\theta})} \Bigg|_{\boldsymbol{x}_{-i}=\text{xmoda}(\boldsymbol{\theta}, x_i)}$$

donde, de nuevo, c es una constante de normalización.

Esta es la parte más difícil pues es potencialmente muy lenta. Las ideas principales son simples y se basan en la identidad

$$p(\boldsymbol{z}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{x}|\boldsymbol{z})} \quad \text{que conduce a} \quad \tilde{p}(\boldsymbol{z}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{\tilde{p}(\boldsymbol{x}|\boldsymbol{z})}.$$

Cuando $\tilde{p}(\boldsymbol{x}|\boldsymbol{z})$ es la aproximación Gaussiana, $\tilde{p}(\boldsymbol{z})$ es la aproximación de Laplace.

La idea es construir las aproximaciones Gausianas de

- $p(\boldsymbol{\theta}|\boldsymbol{y})$
- $p(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

para posteriormente integrar

$$p(x_i|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})p(x_i|\boldsymbol{\theta},\boldsymbol{y})d\boldsymbol{\theta}$$

y

$$p(\theta_j|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}.$$

Teorema 3.2.7. *Con los mismos supuestos de la aproximación de Laplace Clásica, podemos aproximar la distribución condicional mediante una aproximación Gaussiana, es decir,*

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_i \log p(y_i|x_i)\right) \\ &\approx \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{Q} + \mathit{diag}(c_i))(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= p_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}). \end{aligned}$$

El desarrollo puede verse en 3.2.3.

Sabemos que

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}).$$

Por lo tanto,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \underbrace{\frac{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}}_{p(\boldsymbol{y}|\boldsymbol{\theta})} p(\boldsymbol{\theta}) \quad \text{para cualquiera} \quad \boldsymbol{x}.$$

Necesitamos una buena aproximación de $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ para posteriormente integrar numéricamente $\boldsymbol{\theta}$. Para observaciones no Gaussianas tenemos que

$$\log p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{x},\theta) + \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \text{contante}.$$

Usando una aproximación de Taylor de segundo orden

$$f(\boldsymbol{x}) = \log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) \text{ alrededor de } x_0$$

obtenemos su aproximación Gaussiana $p_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$, con

$$\mathrm{E}_{\boldsymbol{x}_0}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \approx \left(\boldsymbol{Q} - \mathrm{diag}(f''(\boldsymbol{x}_0))\right)^{-1}\left(\boldsymbol{Q}\boldsymbol{\mu} + f'(\boldsymbol{x}_0) - f''(\boldsymbol{x}_0)\boldsymbol{x}_0\right)$$

$$\mathrm{Var}_{\boldsymbol{x}_0}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \approx \left(\boldsymbol{Q} - \mathrm{diag}(f''(\boldsymbol{x}_0))\right)^{-1}$$

donde $\boldsymbol{x}_0 = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$.

La demostración de este resultado puede encontrarse en la Sección 3.2.3. Afortunadamente, las propiedades de Markov se conservan pues sólo se modifica la diagonal, y los ceros de la precisión siguen siendo ceros.

En resumen, para evaluar $p(\boldsymbol{\theta}|\boldsymbol{y})$ usando la aproximación de Taylor/ Laplace:

1. Para un valor de $\boldsymbol{\theta}$ encontrar la moda

$$\boldsymbol{x}_0 = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}).$$

2. Calcular la expansión de Taylor de $f(\boldsymbol{x})$ alrededor de $\boldsymbol{x}_0$.

3. La aproximación de $p(\boldsymbol{\theta}|\boldsymbol{y})$ es

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{y}|\boldsymbol{x}_0, \boldsymbol{\theta})p(\boldsymbol{x}_0|\boldsymbol{\theta})p(\theta)}{p_G(\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{\theta})}.$$

   Así, el estimador máximo verosímil (aproximado) es

$$\theta_{ML} \approx \arg\max_{\theta} \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}).$$

   Podemos usar integración numérica sobre θ para obtener las posteriores de $[x|y]$

$$p(x_i|\boldsymbol{y}) = \int p(x_i|\boldsymbol{\theta}, \boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \approx \sum_k p_G(x_i|\boldsymbol{\theta}_k, \boldsymbol{y})\tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{y}).$$

En general, tenemos que lidiar con más de un solo hiperparámetro y que las observaciones no sean Gaussianas. Las complicaciones son usualmente sólo prácticas, no cambia mucho el esquema. Notemos que la complicación principal es que el denominador ya no es Gaussiano y no podemos emplear las herramientas que tenemos para calcular condicionales o marginales del campo Markoviano Gaussiano. Sin embargo, podemos emplear una aproximación Gaussiana. Para ello, debemos explorar $p(\boldsymbol{\theta}|\boldsymbol{y})$ y

1. *Localizar la moda de* $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$ *al optimizar* $\log \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$ *con respecto a* $\boldsymbol{\theta}$. Esto puede hacerse mediante un método cuasi-Newton que construye una aproximación de la segunda derivada de $\log \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$ usando las diferencias entre vectores gradiente sucesivos. El gradiente se aproxima usando diferencias finitas. Sea $\boldsymbol{\theta}^*$ la moda aproximada (hasta la iteración elegida).

2. *Emplear el Hessiano para construir nuevas variables.* En la moda aproximada $\boldsymbol{\theta}^*$ calculamos el negativo de la matriz Hessiana $\boldsymbol{H} > 0$, usando diferencias finitas. Sea $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1}$. que corresponderia a la matriz de covarianza de $\boldsymbol{\theta}$ si su densidad fuera Gaussiana. Para facilitar la exploración se usan variables estandarizadas $\boldsymbol{z}$ en lugar de $\boldsymbol{\theta}$: sea $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ la descomposición en valores propios de $\boldsymbol{\Sigma}$, y definimos $\boldsymbol{\theta}$ a través de $\boldsymbol{z}$, de modo que

$$\boldsymbol{\theta}(\boldsymbol{z}) = \boldsymbol{\theta}^* + \boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}.$$

Si $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$ es una densidad Gaussiana, entonces $\boldsymbol{z}$ tiene distribución Normal$(\mathbf{0}, \boldsymbol{I})$. Esta reparametrización corrige la escala y la rotación, y así simplifica la integración numérica. En la Figura 3.1 éstas nuevas coordenadas estan representadas por los ejes z_1 y z_2.

3. *Buscar en una celosía de puntos de integración para poder aproximar las marginales para cada θ_i.* Primero, buscamos una aproximación de $p(\theta_j|\boldsymbol{y})$. Las densidades marginales posteriores de θ_j pueden obtenerse directamente de $\tilde{p}(\boldsymbol{\theta}_j|\boldsymbol{y})$ mediante integración numérica. Sin embargo, esto es computacionalmente costoso, ya que tenemos que evaluar $\tilde{p}(\boldsymbol{\theta}_j|\boldsymbol{y})$ para un gran número de combinaciones. Un enfoque más factible es utilizar los puntos ya calculados para la integración numérica para construir un interpolador a $\log \tilde{p}(\boldsymbol{\theta}_j|\boldsymbol{y})$, y calcular los marginales utilizando la integración numérica de este interpolador. Si se requiere alta precisión, necesitamos en la práctica un conjunto de puntos más denso (por ejemplo $\delta_z = 1/2$ o $1/4$) que la requerida para el campo latente $\boldsymbol{x}$. En la Figura 3.2 puede verse esta selección de puntos tanto sobre los ejes (negros) como combinación de ellos (grises).

   Siguiendo el desarrollo de la selección de la celosía desarrollada en Rue, Riebler, Sørbye, Illian, Simpson & Lindgren, 2017.

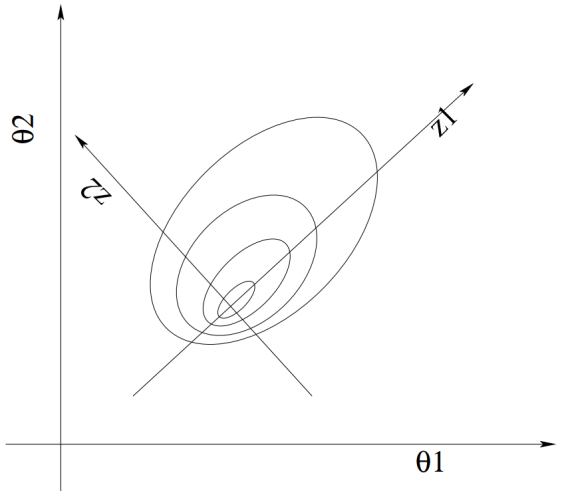


Figura 3.1: Se localiza la moda, se calcula el Hessiano y el sistema de coordenadas de z.

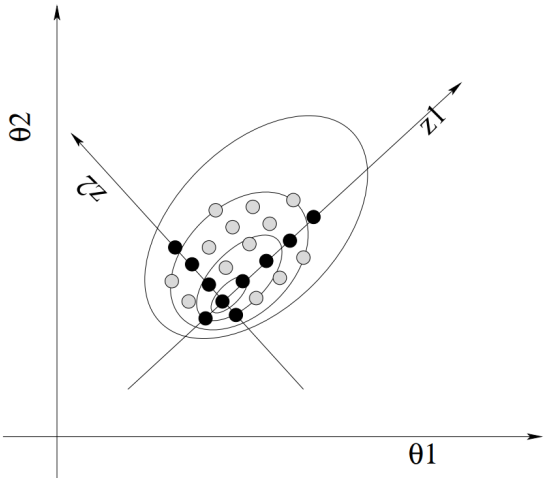


Figura 3.2: Se explora cada dirección de coordenadas (puntos negros) hasta que la densidad logarítmica cae por debajo de un cierto límite. Finalmente se exploran los puntos grises.

Tanto las ideas como los detalles computacionales de INLA pueden encontrarse repartidos en Lindgren, Rue & others, 2015; Rue & Martino, 2007; Rue et al., 2009.

3.2.5. La factorización de Cholesky

Para obtener las densidades marginales del campo latente conviene aprovechar qye tenemos una matriz de precisión rala. La estrategia en general consiste en primero hacer una

permutación sobre el vector $\boldsymbol{x}$ para que la matriz $\boldsymbol{Q}$, además de rala, esté concentrada cerca de la diagonal. Mediante la aproximación Gaussiana, veremos que el efecto de la verosimilitud en la densidad del campo Markoviano Gaussiano modifica de forma sencilla la matriz de precisión. Una vez que tenemos esta nueva $\boldsymbol{Q}'$ podemos obtener las marginales de forma inmediata si obtenemos la matriz $\boldsymbol{\Sigma}' = (\boldsymbol{Q}')^{-1}$. Una forma computacionalmente eficiente de hacerlo es mediante la descomposición de Cholesky.

En su forma usual, esta descomposición es la usualmente usada en estadística, ya que se obtiene para matrices simétricas definidas positivas, como es el caso tanto de la matriz de precisión y la matriz de covarianzas. Además, esta misma factorización permite simular de nuestro campo Markoviano Gaussiano partiendo de normales estándar independientes.

Usualmente deseamos factorizar $\boldsymbol{Q}$ en $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$ (Cholesky) para resolver $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{L}\boldsymbol{x} = \boldsymbol{b}$ o $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{b}$ y calcular diag$(\boldsymbol{Q}^{-1})$, es decir, las varianzas marginales.

Además, existen modificaciones en este algoritmo que aprovechan la raleza de $\boldsymbol{Q}$ consiguiendo importantes mejoras en los tiempos de ejecución. La factorización de Cholesky para una matriz simétrica definida positiva (SPD) rala

- Modelo temporal: $\mathcal{O}(n)$
- Modelo espacial: $\mathcal{O}(n^{3/2})$
- Modelo espacio-temporal: $\mathcal{O}(n^2)$

Comparemos estas complejidades con el orden $\mathcal{O}(n^3)$ del algoritmo de factorización de Cholesky usual para una matriz SPD densa. A continuación veremos como la factorización de Cholesky en su versión general para matrices SDP aparece en los métodos de solución de ecuaciones lineales para después mostrar la modificación del algoritmo para matrices ralas. Los detalles sobre la complejidad algorítmica de estos resultados puede consultarse en Rue et al., 2009.

Si $\boldsymbol{A}$ es una matriz $n \times n$ definida positiva, entonces existe una única matriz triangular inferior (de Cholesky) que cumple

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T.$$

Notemos que como $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$, se cumple que

$$A_{ij} = \sum_{k=1}^{j} L_{ik}L_{jk}, \ i \geq j.$$

y definamos

$$u_i \equiv A_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk}, \ i \geq j.$$

Entonces puede mostrarse que

- $L_{jj}^2 = u_j$, y
- $L_{ij}L_{jj} = u_j$ para $i > j$.

Si conocemos $\{u_i\}$ para j fijo, entonces

$$L_{jj} = \sqrt{u_j} \text{ y } L_{ij} = u_i/\sqrt{u_j}, \text{ para } i = j+1, ..., n.$$

Esto nos da la j-ésima columna de $\boldsymbol{L}$. Ver Algoritmo 1.

También, la factorización puede ser calculada directamente a través de las siguientes fórmulas

$$L_{ii}^2 = A_{ii} - \sum_{k=1}^{i-1} L_{ik}^2 \text{ para los elementos de la diagonal principal}$$

$$L_{ij} = \frac{A_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk}}{L_{jj}} \text{ para el resto de los elementos.}$$

Algoritmo 1 Factorización de Cholesky $\boldsymbol{L}$ de $\boldsymbol{A}$

For $j = 1$ **to** n

$u_{j:n} = A_{j:n,j}$

For $k = 1$ **to** $j - 1$

$u_{j:n} = u_{j:n} - L_{j:n,k}L_{jk}$

$L_{j:n,j} = u_{j:n}/\sqrt{u_j}$

Return L

Calcular $\boldsymbol{L}$ cuesta $n^3/3$ flops (*floating point operations per second*). Esta factorización es la base para resolver sistemas de la forma

$$\boldsymbol{Ax} = \boldsymbol{b} \text{ o } \boldsymbol{AX} = \boldsymbol{B}$$

o equivalentemente para resolver

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \text{ o } \boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{B}.$$

Al paso dos del algóritmo para resolver sistemas de ecuaciones usando la descomposición de Cholesky se le llama sustitución por delante y cuesta $\mathcal{O}(n^2)$ flops. La solución de $\boldsymbol{v}$ se calcula iterando hacia adelante

$$v_i = \frac{1}{L_{ii}}(b_i - \sum_{j=1}^{i-1} L_{ij}v_j),\ i = 1, ..., n$$

Al paso tres se le llama sustitución hacia atrás y cuesta $\mathcal{O}(n^2)$ flops. La solución de $\boldsymbol{x}$

se calcula iterando hacia atrás

$$x_i = \frac{1}{L_{ii}}(v_i - \sum_{j=i+1}^{n} L_{ji}x_j), i = n, ..., 1$$

Para calcular $\boldsymbol{A}^{-1}\boldsymbol{B}$ donde $\boldsymbol{B}$ es una matriz $n \times k$, podemos calcular la solución $\boldsymbol{X}$ de

$$\boldsymbol{A}\boldsymbol{X}_j = \boldsymbol{B}_j$$

para cada una de las k columnas de $\boldsymbol{X}$. Notemos que si podemos resolver este tipo de problemas, eligiendo $\boldsymbol{B} = \mathbb{I}$ podemos recuperar directamente $\boldsymbol{A}^{-1}$, o cualquier elemento de la diagonal de $\boldsymbol{A}^{-1}$, según las cantidades que necesitemos.

Para algunos problemas es útil poder simular de nuestro Campo Markoviano Gaussiano. Por ejemplo para emplear aproximaciones Monte Carlo de transformaciones no lineales de x. Por esto requerimos saber simular de $\boldsymbol{x} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$.

Si $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$ y $\boldsymbol{z} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I})$, entonces $\boldsymbol{x}$ definido por

$$\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$$

tiene covarianza

$$\text{Cov}(\boldsymbol{x}) = \text{Cov}(\boldsymbol{L}^{-T}z) = (\boldsymbol{L}\boldsymbol{L}^T)^{-1} = \boldsymbol{Q}^{-1}.$$

A continuación ahondamos sobre los métodos numéricos que se utilizan.

3.2.6. Métodos numéricos para matrices ralas

Los cálculos en Campos Markovianos Gaussianos pueden expresarse de tal modo que las tareas a realizar son

1. Calcular la factorización de Cholesky de $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$, y

2. Resolver $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{b}$ y $\boldsymbol{L}^T\boldsymbol{x} = z$.

La segunda tarea es mucho más rápida que la primera, pero la raleza nos será de ayuda en la primera. Podemos dar una interpretación útil de $\boldsymbol{L}$. Recordemos la solución de

$$\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z} \text{ donde } \boldsymbol{z} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I})$$

se distribuye Normal$(\mathbf{0}, \boldsymbol{Q}^{-1})$. Como $\boldsymbol{L}$ es triangular inferior, entonces

$$\begin{aligned} x_n &= \frac{1}{L_{nn}} z_n \\ x_{n-1} &= \frac{1}{L_{n-1,n-1}} (z_{n-1} - L_{n,n-1} X_n) \\ x_{n-2} &= \frac{1}{L_{n-2,n-2}} (z_{n-2} - L_{n,n-2} X_n - L_{n-1,n-2} X_{n-1}) \\ &\vdots \end{aligned}$$

Básicamente, da una forma iterativa de construir las variables x_i dadas las $x_{i+1}, ..., x_n$

Debido a que son combinaciones lineales de cantidades conocidas es fácil encontrar la esperanza condicional y la varianza (y correspondientemente la precisión) de las x_m.

Teorema 3.2.8. *Sea $\boldsymbol{x}$ un GMRF con respecto al grafo etiquetado $G = (\mathcal{V}, E)$ y con media $\mathbf{0}$ y matriz de precisión $Q > 0$. Sea $\boldsymbol{L}$ el triángulo de Cholesky de $\boldsymbol{Q}$, entonces para $i \in \mathcal{V}$,*

$$\mathrm{E}(x_i | \boldsymbol{x}_{(i+1):n}) = -\frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji} x_j \text{ y}$$

$$\textit{Prec}(x_i | \boldsymbol{x}_{(i+1):n}) = L_{ii}^2,$$

donde $\boldsymbol{x}_{(i+1):n} = (x_{i+1}, x_{i+2}, ..., x_n)$.

Demostración. Como

$$x_i = \frac{1}{L_{ii}} \left(z_i - \sum_{j=i+1}^{n} L_{ji} x_j \right)$$

y $\mathrm{E}(z_i) = 0$, entonces

$$\mathrm{E}(x_i | \boldsymbol{x}_{(i+1):n}) = 0 - \frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji} x_j.$$

También, como $\mathrm{Var}(z_i) = 1$

$$\begin{aligned} \mathrm{Var}(x_i | \boldsymbol{x}_{(i+1):n}) &= \mathrm{Var}\left(\frac{z_i}{L_{ii}} \right) - \mathrm{Var}\left(\frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji} x_j \right) \\ &= \frac{1}{L_{ii}^2} - 0. \end{aligned}$$

Por lo que Prec $(x_i | \boldsymbol{x}_{(i+1):n}) = L_{ii}^2$. ■

Adicionalmente, al ser $X_i | \boldsymbol{x}_{(i+1):n}$ combinaciones lineales de z_i Normales estándar, se cumple que $X_i | \boldsymbol{x}_{(i+1):n}$ también tienen distribución Normal con la media y precisión ya mencionados.

Teorema 3.2.9. *Sea $\boldsymbol{x}$ un GMRF con respecto a G, con media $\mathbf{0}$ y matriz de precisión $Q > 0$. Sea $\boldsymbol{L}$ el triangulo de Cholesky de $\boldsymbol{Q}$, y definamos para $1 \leq i \leq j \leq n$ el conjunto*

$$F(i,j) = \{i+1, ..., j-1, j+1, ..., n\},$$

que es el futuro de i excepto j. Entonces

$$x_i \perp\!\!\!\perp x_j | \boldsymbol{x}_{F(i,j)} \iff L_{ji} = 0.$$

Notemos que si podemos verificar que L_{ji} es cero, no tenemos que calcularlo cuando factorizamos $\boldsymbol{Q}$. Como es usual en los campos GMRF en INLA, supongamos que $\boldsymbol{\mu} = \mathbf{0}$ y fijemos $1 \leq i < j \leq n$. El Teorema 3.2.8 nos dice que

$$\begin{aligned}\pi(\boldsymbol{X}_{i:j}) &\propto \exp\left(-\frac{1}{2}\sum_{k=i}^{n} L_{kk}^2 \left(x_k + \frac{1}{L_{kk}}\sum_{k=i}^{n} L_{jk}x_j\right)^2\right) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{x}_{i:n}^T \boldsymbol{Q}^{(i:n)} \boldsymbol{x}_{i:n}\right),\end{aligned}$$

donde la matriz $Q^{(i:n)}$ (la expresión en el exponente es un superíndice) de dimensión $(n - i + 1) \times (n - i + 1)$ está definida componente a componente por

$$Q_{ij}^{(i:n)} = Q_{ji}^{(i:n)} = \Big\{ L_{ii}L_{ji}.$$

Entonces

$$x_i \perp\!\!\!\perp x_j | \boldsymbol{x}_{F(i,j)} \iff L_{ii}L_{ji} = 0,$$

que es equivalente a que $L_{ji} = 0$ ya que $L_{ii} > 0$ así como $\boldsymbol{Q}^{(i:n)} > 0$.

El objetivo de identificar estas propiedades es saber donde se localizan de los ceros en $\boldsymbol{L}$ y así no tener que calcularlos. La propiedad global de Markov da un criterio simple y suficiente para comprobar que $L_{ji} = 0$.

A continuación definiremos un concepto útil para caracterizar donde están los ceros en la matriz triangular inferior L. Este viene de la teoría de grafos. Sea $S \subset G$ un subconjunto de vértices, S es un separador de vértices (o corte de vértices, conjunto de separación) para vértices no adyacentes i y j si la eliminación de S del grafo separa i y j en componentes conectados distintos. En este contexto, existe un vértice entre i y j si y sólo si $i \sim j$, es decir, i es vecino de j y viceversa.

Corolario 1. *Si $F(i,j)$ separa $i < j$ en G, entonces $L_{ji} = 0$.*

Corolario 2. *Si $i \sim j$ entonces $F(i,j)$ no separa $i < j$.*

El método para ahorrar cálculos es

- Usar la propiedad global de Markov para revisar si $L_{ji} = 0$.
- Calcular sólo los términos no ceros en $\boldsymbol{L}$, tal que $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$.

Notemos que así el ancho de banda se preserva. Si $\boldsymbol{Q}$ tiene ancho de banda p, entonces $\boldsymbol{L}$ tiene ancho de banda inferior p.

Teorema 1. *Sea $\boldsymbol{Q} > 0$ una matriz banda (diagonal) con ancho de banda p y dimensión n, entonces el triangulo de Cholesky de $\boldsymbol{Q}$ tiene un ancho de banda (inferior) p.*

Es fácil modificar el código original de factorización de Cholesky para sólo calcular las entradas donde $|i-j| \leq p$. De esta forma evitamos calcular L_{ij} y ver Q_{ij} para $|i-j| > p$. Ver Algoritmo 2. El costo computacional ahora es del orden $\mathcal{O}(n(p^2+3p))$ flops suponiendo que $n \gg p$ (ver Rue et al., 2009).

Algoritmo 2 Factorización de Cholesky para $\boldsymbol{Q}$ matriz banda con ancho de banda p

For $j = 1$ **to** n

$\lambda = \text{mín}\{j+p, n\}$

$u_{j:\lambda} = Q_{j:\lambda,j}$

For $k = \text{máx}\{1, j-p\}$ **to** $j-1$

$i = \text{mín}\{k+p, n\}$

$u_{j:i} = u_{j:i} - L_{j:i,k} L j k$

$L_{j:\lambda,j} = u_{j:\lambda}/\sqrt{u_j}$

Return L

Además, es posible permutar los vértices para hacer más eficiente el algoritmo. Sabemos que $\boldsymbol{Q}$ es rala, pero los valores distintos de 0 pueden estar dispersos por toda la matriz y por tanto la banda puede ser grande. En general, se puede elegir una de las $n!$ posibles permutaciones, definir la correspondiente matriz de permutación $\boldsymbol{P}$, tal que $\boldsymbol{i}^P = \boldsymbol{P}\boldsymbol{i}$, donde $\boldsymbol{i} = (1, ..., n)^T$, es el nuevo ordenamiento (orden) de los vertices.

Se busca escoger $\boldsymbol{P}$ tal que

$$\boldsymbol{Q}^P = \boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}^T$$

es una matriz banda con un ancho de banda pequeño.

En general es imposible obtener la permutación óptima pues se trata de un problema de optimización combinatoria y $n!$ es demasiado grande. Sin embargo, un ordenamiento sub-optimal puede ser suficiente. Recordemos que se busca que tenga ancho menor ancho de banda para hacer menos operaciones, pero las soluciones serán las mismas salvo permutación.

Una forma de resolver el problema de encontrar una permutación útil es mediante el llamado reordenado por disección anidada. La idea de este esquema de reordenado es

1. Elegir un (pequeño) número de vertices cuya remoción dividiría el grafo en dos subgrafos disconexos de tamaño parecido casi igual.

2. Ordenar los vértices elegidos **después** de ordenar el resto de los vértices en ambos subgrafos.

3. Aplicar este procedimiento recursivamente a todos los nodos en cada subgrafo.

Notemos que no debemos tratar a las matrices ralas como matrices ordinarias en la computadora. Tanto la forma de almacenarlas en memoria como los algoritmos para manipularlas deben ser especiales para matrices ralas. Es recomendable usar software existente. Si estamos trabajando en R, se recomienda usar el paquete *Matrix*, donde se pueden crear matrices ralas con *sparseMatrix()*.

3.2.7. Calcular las covarianzas marginales de un GMRF

Sea

$$\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T \text{ y } \boldsymbol{\Sigma} = \boldsymbol{Q}^{-1} = \boldsymbol{L}^{-T}\boldsymbol{L}^{-1},$$

donde $\boldsymbol{L}$ es una matriz triangular inferior de la descomposición de Cholesky. La igualdad

$$\boldsymbol{\Sigma} = \boldsymbol{L}^{-1} + (\boldsymbol{I} - \boldsymbol{L}^T)\boldsymbol{\Sigma} \tag{3.3}$$

define una recursión que puede usarse para calcular $\mathrm{Var}(x_i)$ y $\mathrm{Cov}(x_i, x_j)$ para $i \sim j$ esencialmente sin costo cuando se conoce el triangulo de Cholesky $\boldsymbol{L}$. La prueba es sencilla. Dado que $Q\Sigma = I$, entonces $LL^T\Sigma = I$. Multiplicando por la izquierda por L^{-1} y restando Σ en ambos lados da $(L^T - I)\Sigma = L^{-1} - \Sigma$. Tras ordenar los términos obtenemos (3.3). Este es un resultado de 1973, dentro del artículo llamado, *Formation of a sparse bus impedance matriz and its aplication to short circuit study* por Takahashi, Fagan y Mo-Shing Chen.

Recordemos que, como se demuestra en el Teorema 3.2.8, para un GMRF con media cero que

$$x_i | x_{i+1}, ..., x_n \sim \text{Normal}\left(-\frac{1}{L_{ii}}\sum_{k=i+1}^{n} L_{ki}x_k, 1/L_{ii}^2\right), i = n, ..., 1.$$

provee una representación secuencial del GMRF de reversa en el "tiempo" i.

Multiplicando por x_j, $j \geq i$, y tomando esperanzas tenemos que

$$\Sigma_{ij} = \delta_{ij}/L_{ii}^2 - \frac{1}{L_{ii}}\sum_{k\in\mathcal{I}(i)}^{n} L_{ki}\Sigma_{kj}, j \geq i, i = n, ..., 1, \tag{3.4}$$

donde $\mathcal{I}(i)$ son aquellos $k > i$ donde L_{ki} es distinto de cero,

$$\mathcal{I}(i) = \{k > i : L_{ki} \neq 0\}$$

y δ_{ij} es uno si $i = j$ y cero en otro caso.

Podemos usar (3.4) para calcular Σ_{ij} para cada ij. Supongamos que nos interesa calcular todas las varianzas marginales. Para hacer esto, necesitamos calcular Σ_{ij} (o Σ_{ji}, para

todas las ij en un conjunto arbitrario de aristas $S \subset \mathcal{V} \times \mathcal{V}$). Si las recursiones pueden resolverse solamente calculando Σ_{ij} para todo $ij \in S$ entonces decimos que las recursiones tienen solución usando S.

De (3.4) es evidente que para que S tenga solución S debe satisfacer

$$ij \in S \text{ y } k \in \mathcal{I}(i) \Rightarrow kj \in S.$$

También necesitamos que $ii \in S$ para todo $i = 1, ..., n$.

Notemos que $S = \mathcal{V} \times \mathcal{V}$ es un conjunto válido, pero nos interesa que $|S|$ sea pequeño para evitar cálculos innecesarios. Sin embargo, un conjunto mínimo depende de los valores numéricos en $\boldsymbol{L}$, o $\boldsymbol{Q}$ implícitamente. Afortunadamente, un conjunto ligeramente más grande, que contiene al mínimo, resulta ser aquel que ya utilizamos para calcular $\boldsymbol{L}$.

Teorema 2. *El conjunto S, definido abajo, permite calcular las recursiones (ver Takahashi, 1973)*

Se puede demostrar que

$$S = \{ij \in \mathcal{V} \times \mathcal{V} : j \geq i, i \text{ y } j \textit{ no son separados por } F(i,j)\}$$

es un conjunto con el que podemos calcular todas las $\Sigma_{ij}, ij \in \mathcal{V}$.

La prueba de este resultado puede consultarse en Rue & Martino, 2007. Su importancia está en que nos permite identificar de forma más rápida todos los posibles elementos no nulos de L. Algunos de ellos podrían resultar cero dependiendo de las propiedades de independencia condicional de la densidad marginal. Por tanto, una posible interpretación de S es que es el conjunto de L_{ji}'s que se requieren calcular cuando obtenemos $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$. Ya que $L_{ji} \neq 0$ en general cuando $i \sim j$, entonces calculamos también $\text{Cov}(x_i, x_j)$ para $i \sim j$. Algunos de los L_{ij}'s pueden resultar ser cero dependiendo de las propiedades de independencia condicional de la densidad marginal de $\boldsymbol{x}_{i:n}$ para $i = n, ..., 1$ aunque $ij \in S$.

Adicionalmente, es posible calcular varianza marginal bajo restricciones duras como las que se acostumbran contrastar en pruebas de hipótesis. Estas marginales pueden ser de utilidad para calcular factores de Bayes. Sea $\tilde{\boldsymbol{\Sigma}}$ la covarianza con restricciones y $\boldsymbol{\Sigma}$ sin restricciones. Entonces $\tilde{\boldsymbol{\Sigma}}$ se relaciona con $\boldsymbol{\Sigma}$ mediante

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{Q}^{-1},$$

por tanto

$$\tilde{\Sigma}_{ii} = \Sigma_{ii} - \left(\boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{Q}^{-1}\right)_{ii}, i = 1, ..., n.$$

para las restricciones duras $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

3.2.8. Cálculo de densidades marginales del campo latente

Muchos de los problemas en la inferencia Bayesiana son en realidad cálculo de marginales, es decir, básicamente son problemas de integración. En general, puede ser complicado integrar numéricamente en especial si la dimensión de los parámetros es muy grande.

Afortunadamente podemos emplear la aproximación de Laplace para aproximar buena parte de los parámetros, aquellos que aparecen en el campo Markoviano Gaussiano latente.

Consideremos el problema general

- $\boldsymbol{\theta}$ es los hiperparámetros con previa $p(\boldsymbol{\theta})$
- $\boldsymbol{x}$ son las variables latentes con densidad $p(\boldsymbol{x}|\boldsymbol{\theta})$
- $\boldsymbol{y}$ es la variable observada cuya verosimilitud es $p(\boldsymbol{y}|\boldsymbol{x})$

Entonces, notemos que

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}$$

para cualquier $\boldsymbol{x}$.

Ahora,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\ &\propto \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\ &\approx \left.\frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x})}{p_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \end{aligned}$$

donde $p_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ es la aproximación Gaussiana de $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ y $\boldsymbol{x}^*(\boldsymbol{\theta})$ es su moda.

Con n medidas repetidas de $\boldsymbol{y}$ del mismo $\boldsymbol{x}$, entonces

$$\frac{\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_n)}{p(\boldsymbol{\theta}|\boldsymbol{y}_n)} = 1 + \mathcal{O}(n^{-3/2})$$

tras normalización. Este error relativo es muy impresionante y útil. Sin embargo, los supuestos usualmente no se cumplen.

Siguiendo el desarrollo de Gómez-Rubio, 2020, la aproximación de las distribuciones marginales de los hiperparámetros pueden calcularse al marginalizar sobre $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ para obtener $\hat{\pi}(\theta_i|\boldsymbol{y})$. La aproximación de las marginales de los efectos latentes requiere integrar respecto a los hiperparámetros y marginalizar sobre los efectos latentes. INLA usa la siguiente aproximación

$$p(x_i|\boldsymbol{y}) \simeq \sum_{k=1}^{K} \tilde{p}(x_i|\boldsymbol{\theta}^{(k)}, \boldsymbol{y})\tilde{p}(\theta^{(k)})\Delta_k$$

Aquí, $\{\theta^{(k)}\}_{k=1}^{K}$ representa el conjunto de valores de $\boldsymbol{\theta}$ que son usados en la integración numérica y cada uno de ellos tiene asociado una ponderación de integración Δ_k. INLA

obtiene estos puntos de integración al colocar una rejilla regular alrededor de la moda posterior de $\boldsymbol{\theta}$ o usando un diseño central compuesto centrado (ver Box & Draper, 1987) en la media posterior.

En Rue et al., 2009 se describen tres diferentes aproximaciones para $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$. La primera consiste en una aproximación Gaussiana, la cual estima la media $\mu_i(\boldsymbol{\theta})$ y la varianza $\sigma_i^2(\boldsymbol{\theta})$. Ésto es computacionalmente rápido ya que en la exploración de $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ se calcula la distribución $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$, por lo que $\tilde{\pi}_G(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ puede calcularse fácilmente marginalizando esta normal multivariada. Esto es trivial si se conoce la matriz de varianza, ya que $\sigma_i^2(\boldsymbol{\theta}) = \Sigma_{ii}(\boldsymbol{\theta})$. En caso de tenerse la matriz de precisión primero se debe invertir esta. La segunda aproximación consiste en la aproximación de Laplace, de tal modo que la aproximación de $p(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ es

$$p_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{p}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-1}=\boldsymbol{x}^*_{-1}(x_i,\boldsymbol{\theta})}$$

La distribución $\tilde{p}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ representa la aproximación Gaussiana de $\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y}$ y $\boldsymbol{x}_{-i} = \boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})$ es su moda. Esta aproximación es mucho más costosa en términos computacionales que la aproximación Gaussiana ya que debe calcularse para cada valor de x_i. Por esta razón, Rue et al., 2009 proponen una aproximación modificada que depende de

$$\pi_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \text{Normal}(x_i|\mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})) \exp(\text{spline}(x_i)).$$

Por lo que la aproximación de Laplace depende del producto de la aproximación Gaussiana y un spline cúbico spline(x_i) en x_i. El spline se calcula en los valores elegidos de x_i y su objetivo es corregir la aproximación Gaussiana. La tercer aproximación $\tilde{\pi}_{SLA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ es llamada la aproximación simplificada de Laplace y depende de una expansión en series de $\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ alrededor de $x_i = \mu_i(\boldsymbol{\theta})$. Con esto, a la aproximación Gaussiana $\tilde{\pi}_G(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ puede corregírsele su localización y asimetría (skewdness) y es muy rápida computacionalmente.

3.2.9. Selección de modelos y detección de datos atípicos

Elegir y comparar varios modelos es una tarea importante pero usualmente difícil. Los criterios más naturales no siempre son accesibles de calcular.

3.2.9.1. Verosimilitud Marginal, Factor de Bayes y probabilidad posterior

Entre los posibles criterios de selección está la verosimilitud marginal que es la constante de normalización de $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\tilde{p}(\boldsymbol{y}) = \int \left. \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

En muchos modelos jerárquicos la previa $p(\boldsymbol{\theta})$ es impropia, lo que dificulta usar este criterio. La aproximación dada por INLA es

$$\tilde{p}(\boldsymbol{y}) = \int \left. \frac{p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})}{\tilde{p}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

En general, esta aproximación es bastante precisa. Es interesante que en INLA sea relativamente fácil de calcular, con adecuada precisión, una cantidad importante que con otros métodos (MCMC) sería bastante difícil de aproximar.

Recordemos que la verosimilitud marginal puede usarse para calcular la probabilidad posterior de un modelo ajustado mediante

$$p(\mathcal{M}_m|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mathcal{M}_m)p(\mathcal{M}_m),$$

donde $p(\mathcal{M}_m)$ es la probabilidad previa de cada modelo.

3.2.9.2. Criterio de Información de Devianza

Recordemos que Criterio de Información de Devianza (DIC) está basado en la devianza

$$D(\boldsymbol{y}; \boldsymbol{\theta}) = -2 \sum_i \log(y_i|x_i, \boldsymbol{\theta})$$

y se define

$$\text{DIC} = 2\text{E}[D(\boldsymbol{y}; \boldsymbol{\theta})] - D(\text{E}[\boldsymbol{y}]; \boldsymbol{\theta}^*),$$

donde $\boldsymbol{\theta}^*$ es la media posterior o bien la moda posterior de $\boldsymbol{\theta}$. Ha sido muy usado pues en general es fácil de calcular si se cuenta con una muestra MCMC.

3.2.9.3. Criterios tipo Leave One Out (LOO)

3.2.9.4. Ordenadas Predictivas Condicionales (CPO)

Para identificar observaciones atípicas y en general evaluar el poder predictivo del modelo podemos emplear validación cruzada Bayesiana. En particular, mediante el enfoque de INLA es relativamente fácil calcular

$$p(y_i|\boldsymbol{y}_{-i}) = \int_{\boldsymbol{\theta}} \left\{ \int_{x_i} p(y_i|x_i, \boldsymbol{\theta}) p(x_i|\boldsymbol{y}_{-i}, \boldsymbol{\theta}) dx_i \right\} p(\boldsymbol{\theta}|\boldsymbol{y}_{-i}) d\boldsymbol{\theta}$$

donde

$$p(x_i|\boldsymbol{y}_{-i}, \boldsymbol{\theta}) \propto \frac{p(x_i|\boldsymbol{y}, \boldsymbol{\theta})}{p(y_i|x_i, \boldsymbol{\theta})}$$

Notemos que sólo requiere una integral unidimensional para cada i y cada componente de $\boldsymbol{\theta}$.

Recordemos que

$$\text{CPO}_i = \int p(y_i^{\text{obs}}|\boldsymbol{y}_{-i}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}_{-i}) d\boldsymbol{\theta}$$

donde y_i^{obs} se refiere al valor en efecto observado, el cual se retira para construir la densidad predictiva de tipo *Leave One Out*. El primer término en la integral del CPO_i es igual a

$$p(y_i^{\text{obs}}|\boldsymbol{y}_{-i}, \boldsymbol{\theta}) = 1 \Big/ \int \frac{p(x_i|\boldsymbol{y}, \boldsymbol{\theta})}{p(y_i^{\text{obs}}|x_i, \boldsymbol{\theta})} dx_i.$$

Esta expresión se aproxima mediante integración numérica pues ya hemos visto como podemos obtener $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\theta})$ mediante INLA.

3.2.9.5. Transformación Integral Predictiva (PIT)

Podemos también hacer detección automatica de observaciones "sorprendentes" al calcular

$$P(y_i^{\text{nueva}} \leq y_i|\boldsymbol{y}_{-i})$$

e identificar valores inusualmente altos o bajos. Notemos que aqui y_i^{nueva} denota una variable aleatoria que queremos predecir, a pesar de conocer el valor y_i^{obs}.

Recordemos que ya hemos visto como podemos calcular de forma aproximada $p(x_i|\boldsymbol{y}, \theta)$ mediante INLA. El denominador $p(y_i^{\text{obs}}|\boldsymbol{y}_{-i}, \boldsymbol{\theta})$ ya lo hemos calculado. Tenemos también ya la aproximación de $p(\boldsymbol{\theta}|\boldsymbol{y})$.

El término $p(y_i^{\text{obs}}|\boldsymbol{y}_{-i})$ puede calcularse, aproximadamente, con

$$\tilde{p}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}) = \left(\sum_k \frac{\tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{y})}{\tilde{p}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}, \boldsymbol{\theta}_k)} \Delta_k \right)^{-1}.$$

Aquí los $\boldsymbol{\theta}_k$'s son puntos de soporte de la densidad marginal posterior aproximada $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$, la cuál ya se obtuvo en el primer paso del procedimiento de ajuste de INLA. Los pesos Δ_k se escogen de manera apropiada de acuerdo al tipo de aproximación numérica que se eligió utilizar. Por lo que el estimado $\tilde{p}(y_i^{\text{obs}}|\boldsymbol{y}_{-i})$ es una media armónica ponderada de los $\tilde{p}(y_i^{\text{obs}}|\boldsymbol{y}_{-i}, \boldsymbol{\theta}_k)$'s, $k = 1, ..., K$ con pesos $w_k = \tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k$. Ahora ya hemos calculado todos los términos en las expresiones de los CPO y los PIT. La aproximación final de cada PIT_i mediante la expresión primera se basa en los puntos de soporte θ_k como en la última expresión al remplazar la integral con una suma finita. Respecto a los CPO_i, notemos que $p(y_i^{\text{obs}}|\boldsymbol{y}_{-i})$ ya es calculada en $\tilde{p}(y_i^{\text{obs}}|\boldsymbol{y}_{-i})$, por lo que la integración adicional no es necesario.

3.3. R-INLA

Podemos ejemplificar el uso de R-INLA mediante una aplicación sencilla de modelos con efectos aleatorios. En este trabajo exploraremos especialmente la componente espacial

del fenómeno de los eventos violentos, pero también podemos plantear un análisis solamente temporal.

Podemos tomar el total de eventos ocurridos en cierto lapso de tiempo (mes o día) en todo el país. Así tendremos una serie de tiempo que podemos analizar mediante un efecto temporal autorregresivo o bien, mediante una caminata aleatoria. Adicionalmente, podemos ajustar un par de modelos más sencillos, donde el efecto temporal se incorpora como una covariable o efecto fijo, ya sea en escala lineal o logarítmica respecto al predictor lineal.

El modelo autorregresivo de orden 1 se define tal que

$$X_i|\phi, \tau, X_1, ..., X_{i-1} \sim \text{Normal}(\phi X_{i-1}, \tau), i = 1, ..., n$$

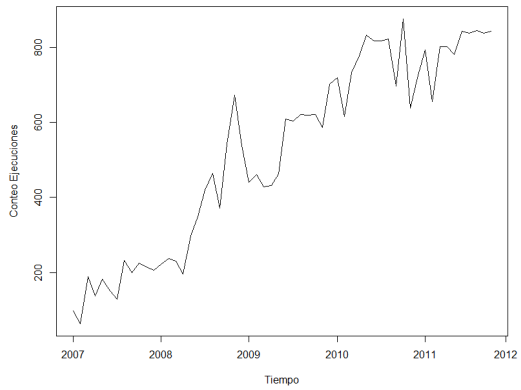
o bien

$$p(\boldsymbol{x}|\tau) \propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \phi x_{i-1})^2\right).$$

El modelo caminata aleatoria de orden 1 se define tal que

$$\begin{aligned} v_t|v_{-t} &\sim \text{Normal}(v_{t+1}, \tau_v) \text{ para } t = 1 \\ v_t|v_{-t} &\sim \text{Normal}\left(\frac{v_{t+1} + v_{t-1}}{2}, \frac{\tau_v}{2}\right) \text{ para } t = 2, ..., T-1 \\ v_t|v_{-t} &\sim \text{Normal}(v_{t-1}, \tau_v) \text{ para } t = T. \end{aligned}$$

Por ahora, nos concentraremos en el caso de las ejecuciones. Podemos separar por meses o bien por días. Notemos que, en particular, estamos tratando series de tiempo con conteos.



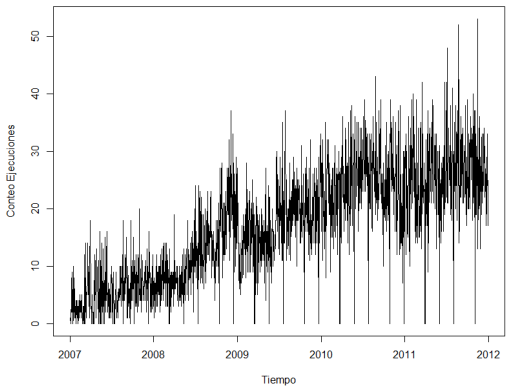


Figura 3.3: Ejecuciones a nivel nacional por mes **Figura 3.4:** Ejecuciones a nivel nacional por día.

Primero debemos definir la formula (forma del regresor lineal). En este caso nos interesan dos modelo con efectos fijos, un modelo de efectos aleatorios autorregresivo o un modelo de efectos aleatorios caminata aleatoria. Además, a los cuatro modelos incluimos efecto

fijo α que es el intercepto. Esta regresión es sobre la log media del modelo Poisson. Los efectos deben exponenciarse para tener la tasa marginal del efecto. Sea $y_i \sim$ Poisson(μ_i) con respectivamente para cada modelo

1. $\log(\mu_i) = \alpha + \beta t_i$ (Efecto exponencial del tiempo sobre la media)

2. $\log(\mu_i) = \alpha + \beta \log(t_i)$ (Efecto lineal del tiempo sobre la media)

3. $\log(\mu_i) = \alpha + X_i$ (Efecto autoregresivo)

4. $\log(\mu_i) = \alpha + v_i$ (Efecto caminata aleatoria)

Recordemos que las salidas son conteos, por lo que las distribuciones predictivas son discretas. Esto puede suponer dificultades respecto al análisis de los PIT, cuestión a la que daremos respuesta en el Capítulo 2.

La función *summary* nos regresa algunos resúmenes de la distribución posterior de los hiperparámetros. También podemos obtener algunos de los criterios para comparar modelos Bayesianos mencionados en el Capítulo 2. Debemos especificar al llamar a R-INLA que criterios queremos que calcule.

```
data<- data.frame(y= STD[TimeD], ID.Time = TimeD)

formula.ST1<- y ~ 1 + TimeD
formula.ST2<- y ~ 1 + log(TimeD)
formula.ST3<- y ~ 1 + f(ID.Time,model="ar1")
formula.ST4<- y ~ 1 + f(ID.Time,model="rw1")

modelo.inla.ST <- inla(formula.ST,family="poisson",data=data,
control.predictor=list(compute=TRUE),
control.compute=list(dic=TRUE,waic=T,cpo=TRUE))

summary(modelo.inla.ST)
modelo.inla.ST$dic$dic
modelo.inla.ST$dic$p.eff
modelo.inla.ST$waic$waic
modelo.inla.ST$mlik
sum(model.inla.ST$cpo$cpo)/length(STD[TimeD])
sum(model.inla.ST$cpo$pit)/length(STD[TimeD])
```

En la Tabla 3.1 podemos ver que todos los criterios indican que el modelo menos adecuado para los datos es el que llamamos Exponencial. Mientras que los indicadores de los modelos AR1 y RW1 son muy parecidos. Vale la pena mencionar que ninguna de las pruebas de bondad de ajuste para estos modelos rechazaron la hipótesis de que los PIT sean Uniformes. Esto se debe en parte a que al ser conteos por mes son números grandes y la aproximación Normal a la distribución Poisson es adecuada.

Modelo	DIC	Eff(P_D)	WAIC	logMLIK	$\overline{CPO}$	$\overline{PIT}$	D_{KS}	p_v KM	$T_{\chi^2_{29}}$	$p_v\chi^2$
Exponencial	1957.1	2.17	2012.2	-994.2	0.0026	0.4463	0.426	1.5e-09	205.6	0^*
Lineal	1206.5	2.18	1234.2	-614.4	0.0038	0.4539	0.363	2.3e-07	168.4	0^*
AR1	579.6	50.78	579.2	-356.8	0.0049	0.5100	0.161	0.09	16.4	0.127
RW1	580.8	50.14	582.4	-352.5	0.0050	0.5103	0.159	0.096	16.4	0.127

Tabla 3.1: Criterios Bayesianos de los modelos temporales para Ejecuciones por mes

Mientras que en la Tabla 3.1 todos los criterios indican que el peor modelo es el Exponencial, luego el Lineal, y muy cercanos siguen los modelos AR1 y RW1. Todas las pruebas de bondad de ajuste tienen p-valores muy pequeños, pero podemos emplear los estadísticos como índices informales de cercanía a la hipótesis nula. Que los PIT no tengan distribución Uniforme continua en $[0, 1]$ no necesariamente debe preocuparnos, pues al ser conteos por día los números no son tan grandes la aproximación Normal a la distribución Poisson no siempre será adecuada.

Modelo	DIC	Eff(P_D)	WAIC	logMLIK	$\overline{\log \text{CPO}}$	$\overline{\text{PIT}}$	D_{KS}	p_v KM	$T_{\chi^2_{29}}$	$p_v\chi^2$
Exponencial	13471.9	2.19	13475.4	-6753.1	0.055	0.514	0.124	0^*	3178.5	0^*
Lineal	12718.1	2.19	12721.7	-6369.5	0.064	0.524	0.105	0^*	1858.8	0^*
AR1	11505.0	443.3	11734.2	-5983.1	0.064	0.533	0.066	2.1e-07	717.6	0^*
RW1	11553.4	386.5	11774.1	-5983.4	0.065	0.533	0.068	1.1e-07	730.4	0^*

Tabla 3.2: Criterios Bayesianos de los modelos temporales para Ejecuciones por día

Comparación de resultados

A fin de examinar más a detalle los resultados que puede darnos INLA compararemos dos modelos. Comparamos el modelo que incorpora el factor tiempo como un efecto fijo con escala logarítmica del tiempo al modelo con efecto aleatorio de tipo autorregresivo.

Por día con modelo lineal respecto al valor esperado

Vemos en la Tabla 3.3 que el tiempo empleado es muy poco. Incluso el proceso de correr INLA, que es usualmente el más tardado, es muy rápido. Esto se debe a que en general los efectos fijos requieren menos poder computacional y la dimensión de este problema es muy pequeña.

Pre-procesamiento	Corriendo INLA	Post-procesamiento	Total
0.8019 seg	0.6702 seg	0.0788 seg	1.5509 seg

Tabla 3.3: Tiempos de ejecución en INLA para modelo lineal respecto al valor esperado

En la Tabla 3.4 podemos ver la inferencia hecha para los parámetros de los efectos fijos. Podemos decir que el efecto del tiempo es positivo, es decir creciente, y que tiene un efecto de $\exp(0.696) = 2.006$ respecto al tiempo. Es decir, al mes i el efecto multiplicativo del tiempo es aproximadamente $2 \times i$.

Parámetro	Media	Desv. Est.	Cuantil 0.025	Mediana	Cuantil 0.975	Moda
Intercepto	-1.8696	0.0670	-2.0016	-1.8694	-1.7385	-1.8691
log(Tiempo)	0.6962	0.0096	0.6773	0.6962	0.7152	0.6961

Tabla 3.4: Estimación de coeficientes para modelo lineal respecto al valor esperado

El modelo no tiene efectos aleatorios.
El modelo no tiene hiperparámetros.

Por día con modelo autoregresivo de orden 1

Vemos en la Tabla 3.5 que el tiempo empleado es también pequeño. Al tratarse de un modelo con efectos aleatorios tarda un poco más que su contraparte con efectos fijos, pero al ser la dimensión de este problema es pequeña el tiempo total es corto.

Pre-procesamiento	Corriendo INLA	Post-procesamiento	Total
1.0242 seg	9.1955 seg	0.1406 seg	10.3603 seg

Tabla 3.5: Tiempos de ejecución en INLA para el modelo AR(1)

En la Tabla 3.6 vemos la inferencia para el único efecto fijo de este modelo, el intercepto. Este puede interpretarse como la log media en el tiempo 0.

Parámetro	Media	Desv. Est.	Cuantil 0.025	Mediana	Cuantil 0.975	Moda
Intercepto	2.5984	0.2091	2.1644	2.603	3.0037	2.6092

Tabla 3.6: Estimación de coeficientes para el modelo AR(1)

Mientras que en la Tabla 3.7 podemos ver la inferencia asociada a los hiperparámetros de los efectos aleatorios. Valores pequeños de la precisión indican un efecto reducido de la componente estrictamente estocástica del modelo autorregresivo. Mientras que un valor menor a 1 del coeficiente del autorregresivo indica que no crecerá de forma explosiva.

Parámetro	Media	Desv. Est.	Cuantil 0.025	Mediana	Cuantil 0.975	Moda
Precisión del AR(1)	2.0419	0.4885	1.1842	2.0137	3.0869	1.9652
Coeficiente del AR(1)	0.9836	0.0050	0.9725	0.9839	0.9921	0.9848

Tabla 3.7: Estimación de coeficientes para modelo lineal respecto al valor esperado

CAPÍTULO 4

Análisis exploratorio de los datos del CIDE-PPD

La seguridad ciudadana y el desarrollo humano mantienen una relación de mutua retroalimentación. A través de su monopolio del ejercicio legítimo de la violencia, el Estado es el responsable de proveer la seguridad ciudadana, en tanto que es un bien público. En PNUD, 2013 se reportaron hallazgos y comentarios a un estudio comparativo (entre países de América Latina) de la violencia debido a distintos factores (políticos, económicos y sociales) y cómo esta afecta a la seguridad ciudadana. El estudio se hace a través de distintos periodos de tiempo (pues no se tiene información homogénea de todos los países) e incluye el periodo que es de nuestro interés. Sin embargo, no se ahonda en la información espacial y temporal con la que pretendemos trabajar pues entonces no se tenían los datos de lugar y tiempo. Además, no se pretendía incorporar este tipo de información ya que se requirío usar sólo los reportes oficiales que los países proveían, los cuales eran a nivel nacional y agregados en el tiempo.

En Atuesta, Siordia & Lajous, 2018 se expone la forma en que se integró la base de datos del CIDE-PPD, con la que trabajaremos. Tuvo un proceso de validación de los eventos y los investigadores involucrados consideraron adecuado la clasificación de los eventos en tres categorías:

Ejecuciones: Cualquier homicidio intencional de un miembro de una organización criminal. No incluye ninguno de los otros dos casos.

Confrontaciones: Actos violentos contra autoridades o victimas en los que la fuerza pública responde con armas de fuego o bien enfrentamientos entre grupos criminales.

Agresiones: Actos violentos contra autoridades o victimas que no obtuvieron respuesta armada por parte de la fuerza pública.

Esta base de datos intenta contabilizar todos los eventos violentos registrados durante el periodo de interés. Atuesta et al. realizan algunas visualizaciones del total de eventos en el

país para cada categoría, en forma de series de tiempo, y también algunas visualizaciones espaciales por estado, pero no han ajustado un modelo propiamente espacio temporal.

En este trabajo, primero se realizó un análisis exploratorio en la que se emplearon diversas herramientas del análisis multivariado y de visualización (aplicadas en la Sección 4.1). En seguida, en la Sección 4.2, se hace la estimación de la función de intensidad para un proceso puntual Poisson no homogéneo. En la Sección 4.3 se analizan los datos con una interpolación por Kriging ordinario. En esta interpolación para cada área se obtiene el centroide del polígono asociado al área y el valor observado es el total de eventos ocurridos en dicha área para cierto intervalo temporal. En la Sección 4.4 se proponen algunos modelos jerárquicos simples, para descomponer los datos en elementos como los temporales y espaciales.

Todo el código usado para generar los distintos análisis pueden encontrarse en el siguiente enlace al repositorio en Github https://github.com/mcarranzba/Tesis_INLA.

4.1. Análisis exploratorio

En la siguiente sección se mencionan algunas de las herramientas del análisis multivariado y la ciencia de datos que utilizamos para visualizar la base de datos de eventos violentos, así como su aplicación en nuestros datos de interés. Las descripciones de los métodos así como demostraciones de sus propiedades pueden encontrarse en Hastie, Tibshirani, Friedman & Franklin, 2005. Entre las herramientas usadas podemos mencionar:

1. Análisis de Componentes Principales

2. Gráficas de lineas

Para cada uno de los tipos de eventos contamos el total de eventos en cierto periodo de tiempo, en nuestro caso en un semestre. Tenemos una ventana de tiempo total de 5 años, por lo que tenemos 10 periodos temporales (semestrales) para cada área.

Podemos tomar el conteo de fallecidos o de eventos por semestre para todos los estados para cada uno de los tipos de eventos (ejecución, enfrentamiento o agresión). De esta forma, obtenemos sucesiones temporales para cada estado y podemos gratificarlas. En este caso, estandarizamos por el tamaño de población del estado estimado en el periodo en cuestión. Para manejar cifras más interpretables lo manejamos en fallecidos o eventos por millón de habitantes. Finalmente, conviene evaluar las ejecuciones en término del número de fallecidos ya que da una medida de la magnitud. En el caso de enfrentamientos y agresiones conviene considerar el número de eventos ya que muchos de ellos no tuvieron fallecidos. Si tomaramos el número de fallecidos no se contabilizarían muchos eventos. En la Figura 4.1 puede observarse que el estado de Chihuahua destaca por su alta tasa (por millón de habitantes) de ejecuciones. Altas también, aunque en menor medida y especialmente al final del sexenio, están Durango, Sinaloa, Guerrero y Nayarit.

Respecto a la Figura 4.2 observamos que Tamaulipas se destaca especialmente los últimos semestres. Constantes a través del sexenio son los estados de Durango y Chihuahua,

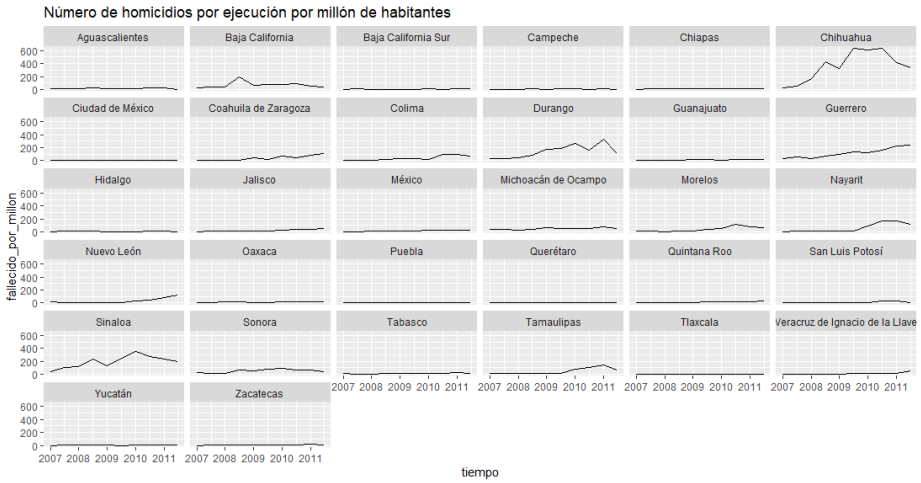


Figura 4.1: Series de tiempo (semestrales) estandarizado por millón de habitantes

mientras que Nuevo León, Nayarit y Guerrero tienen en el cierre del sexenio un número ajustado de eventos muy elevado.

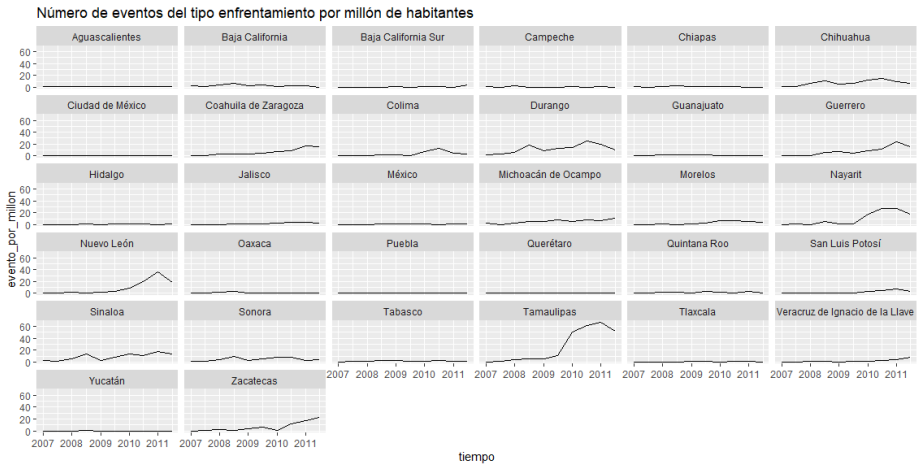


Figura 4.2: Series de tiempo (semestrales) estandarizado por millón de habitantes

Finalmente en la Figura 4.3 se destaca Chihuahua, Nuevo León, Tamaulipas, Guerrero, Durango y Coahuila. Es interesante notar cómo es prácticamente exclusiva la presencia de los estados del norte del pais y el caso especial de Guerrero.

Los conteos de eventos por millón por semestre y por estado pueden verse como un vector. Podemos aplicar las herramientas de análisis multivariado para visualizar el parecido de su comportamiento. Se aplicó PCA clásico. En la Figura 4.4 puede observarse el parecido del comportamiento de la series temporales con el de la proyección mediante PCA. Al igual que en las series temporales, observamos que el estado de Chihuahua se distingue mucho del resto respecto a ejecuciones y agresiones. Durango y Sinaloa son cercanos. Mientras que Nuevo León y Tamaulipas estan más alejados de la media (el centro) respecto a los enfrentamientos y las agresiones. El tamaño de las letras (que representa los semestres más recientes) van creciendo conforme nos alejamos del centro. La excepción es el último

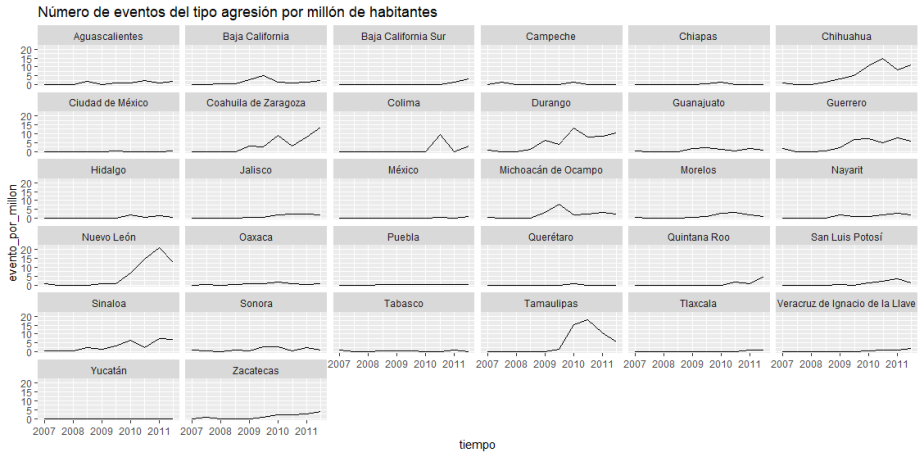


Figura 4.3: Series de tiempo (semestrales) estandarizado por millón de habitantes

semestre que no tiene el conteo completo de los eventos de ese periodo de tiempo.

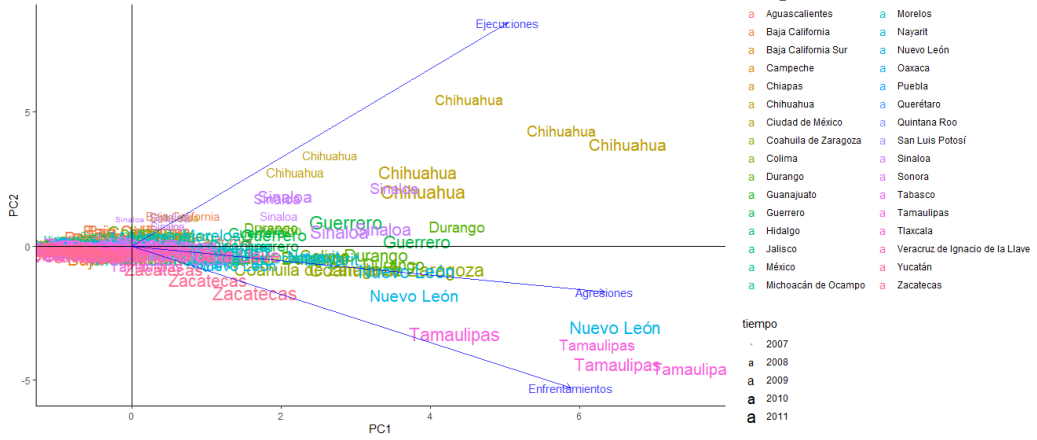
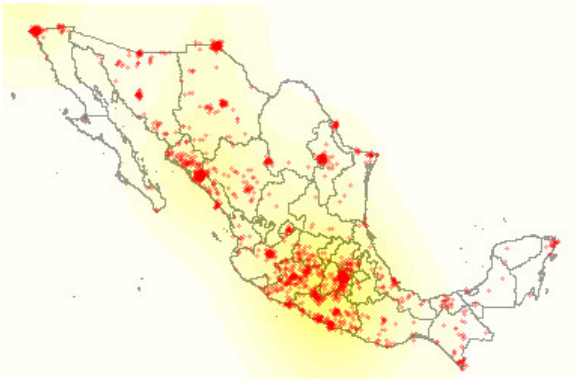


Figura 4.4: PCA de eventos por millón de habitantes por semestre-estado

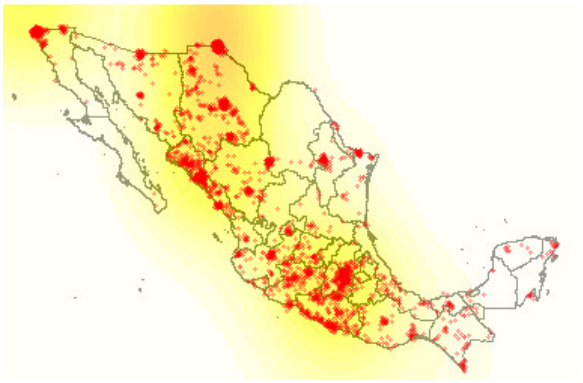
4.2. Resultados del análisis por procesos puntuales

Para cierto periodo, por ejemplo por semestre, podemos graficar los eventos observados y también la función de intensidad estimada mediante kernels. Esto sirve como una visualización similar al Kriging. Además, al hacerlo para varios periodos y observarlos en secuencia nos ayuda a darnos una idea de la dinámica de la aparición de eventos (para cualquiera de los tres tipos). Mostraremos los últimos dos años de cada uno de los tipos de evento violento.

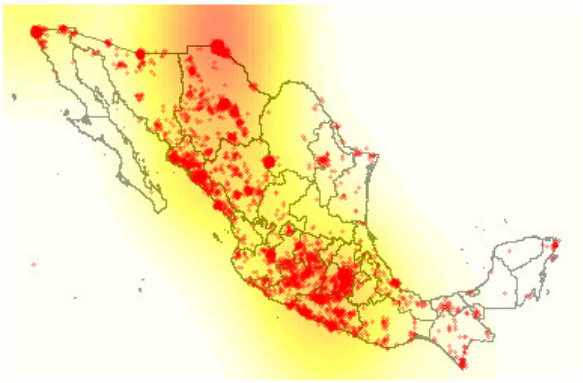
En la Figura 4.5d puede observarse que durante el 2010 la función de intensidad estimada toma valores más altos en los estados de Chihuahua, Sinaloa y Durango, seguido por Guerrero. Mientras que en 2011 como se ve en 4.5e los conteos se intensifican en Guerrero. Recordemos que estos datos solo incluyen el tiempo que duró el sexenio, por lo que los conteos reales del año sería aun mayores.



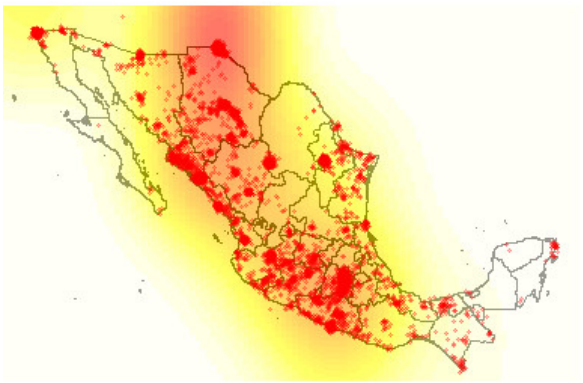
(a) Ejecuciones durante el 2007.



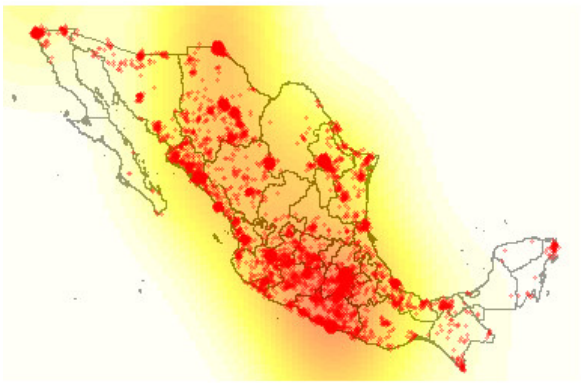
(b) Ejecuciones durante el 2008.



(c) Ejecuciones durante el 2009.



(d) Ejecuciones durante el 2010.

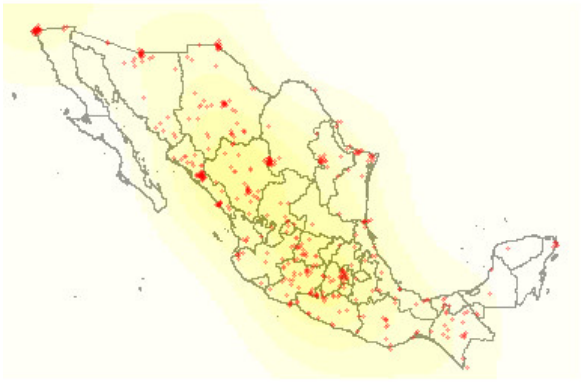


(e) Ejecuciones durante el 2011.

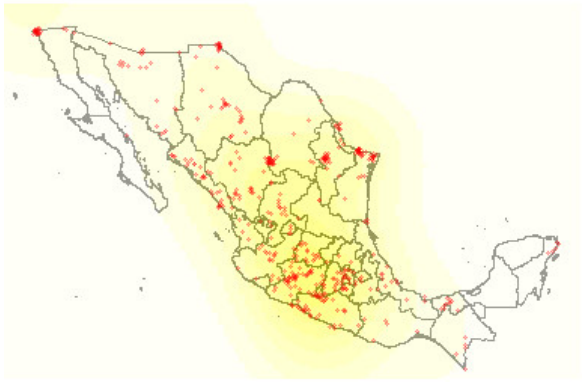
Figura 4.5: Función de intensidad de un proceso Poisson no homogéneo en Ejecuciones

Respecto a los últimos años de las Ejecuciones en la Figura 4.5d puede observarse en el 2010 que la función de intensidad toma valores más altos alrededor del estado de Monterrey y Tamaulipas. Más aún, continuando al 2011 en 4.5e vemos que las apariciones se intensifican en estos dos estados a pesar de que no se contabilizan los Enfrentamientos fuera del periodo de tiempo del sexenio.

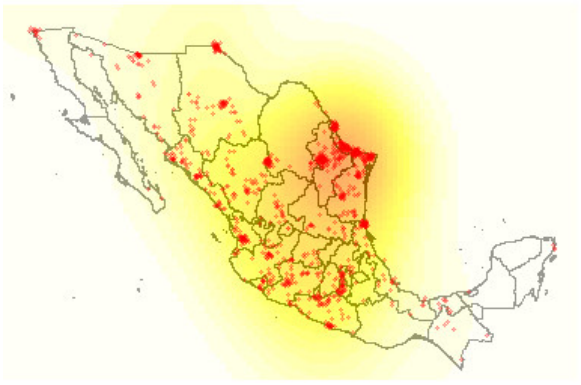
(a) Enfrentamientos durante el 2007.



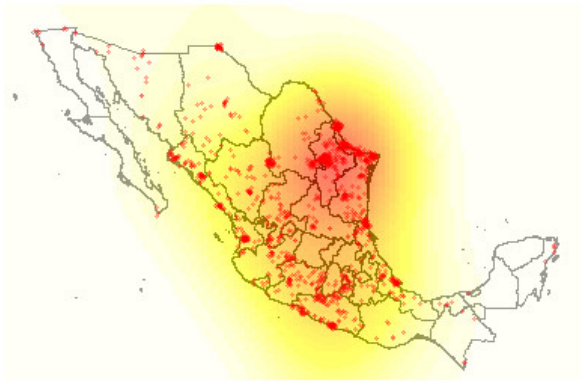
(b) Enfrentamientos durante el 2008.



(c) Enfrentamientos durante el 2009.



(d) Enfrentamientos durante el 2010.



(e) Enfrentamientos durante el 2011.

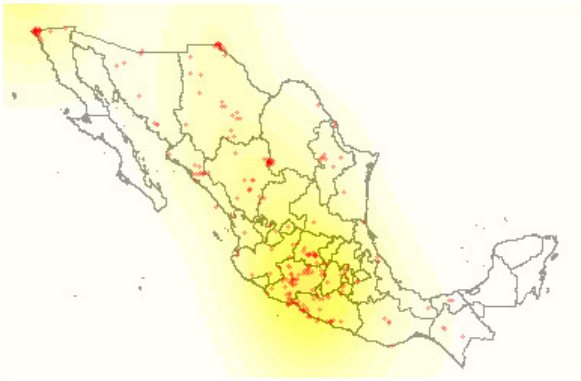
Figura 4.6: Función de intensidad de un proceso Poisson no homogéneo en Enfrentamientos

Finalmente, respecto a las agresiones vemos en la Figura 4.7d en el 2010 las observaciones se concentran en Monterrey y Tamaulipas, similar a lo observado en el caso de las Ejecuciones. También de forma similar a las Ejecuciones, vemos que las Agresiones en

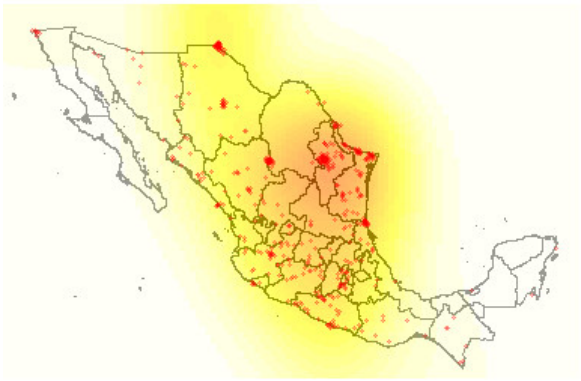
2011 en 4.5e las observaciones se intensifican.

(a) Agresiones durante el 2007.

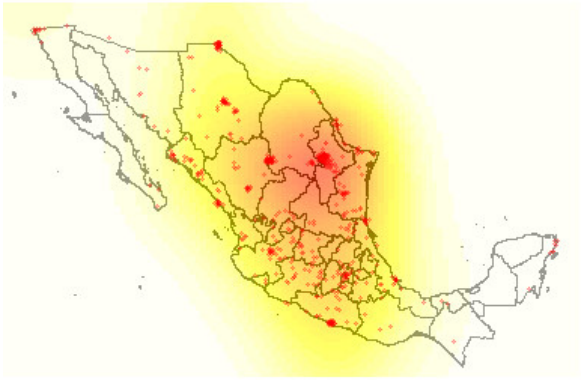
(b) Agresiones durante el 2008.



(c) Agresiones durante el 2009.



(d) Agresiones durante el 2010.



(e) Agresiones durante el 2011.

Figura 4.7: Función de intensidad de un proceso Poisson no homogéneo en Agresiones

Estas gráficas de la función de intensidad estimada representan la distribución de los eventos en el espacio. De este modo la intensidad esta ponderada por la extensión geográfica. En las siguientes gráficas se ponderará por el tamaño de la población. Tener en consideración ambos enfoques nos permitirá tener una imagen más completa de la distribución del fenómeno de la violencia.

4.3. Centroides para hacer Kriging

Explicamos en la Sección 1.5 cómo podemos obtener predictores para puntos en el espacio no observados.Para poder usar Kriging necesitamos una variable aleatoria X_{s_i} que esté instanciada en varios puntos del espacio s_i. Podemos construir dicha variable considerando el conteo de eventos en un polígono. Esta transformación tiene la ventaja de que podemos estandarizar los conteos con respecto a la población. Para cada área (estados o municipios) contamos los eventos que ocurrieron en cierto subperiodo de tiempo o en todo el periodo de interés. Si consideramos los conteos directamente el codominio de X_{s_i} es $\mathcal{N}$ (variables discretas). Si estandarizamos respecto al tamaño de la población entonces el codominio es $\mathcal{R}^+$ (variables continuas). Sin embargo, resta asignarle una localización espacial en coordenadas a cada una de las áreas. Por simplicidad, podemos optar por asignarles su centroide, el cual definiremos a continuación.

El centroide (también llamado centro de masa o centro de gravedad) de un polígono se calcula obteniendo la suma ponderada de los centriodes de una partición del polígono en polígonos más sencillos, usualmente triángulos, de modo que

$$(\bar{x}, \bar{y}) = \left(\frac{\sum_{i=1}^{n} A_i \bar{x}_i}{\sum_{i=1}^{n} A_i}, \frac{\sum_{i=1}^{n} A_i \bar{y}_i}{\sum_{i=1}^{n} A_i} \right),$$

donde A_i es el área del triángulo i. El centroide de cada triangulo es simplemente el promedio de sus tres vértices, es decir,

$$\bar{x}_{abc} = \frac{x_a + x_b + x_c}{3} \quad \text{y} \quad \bar{y}_{abc} = \frac{y_a + y_b + y_c}{3}.$$

Esto sugiere primero triangular el polígono, después obtener la suma de los centroides ponderada por el área de cada uno de los triángulos, luego normalizar por el área del polígono (es decir, la suma del área de los triángulos).

El área de un triángulo con vértices a, b, c es fácil de obtener. Notemos que podemos centrar las aristas AB y AC al restar (x_a, y_a) a (x_b, y_b) y (x_c, y_c). Hecho esto, el área del triangulo es la mitad del paralelogramo cuyos lados son AB y AC. Sabemos que el área del paralelogramo es igual al valor absoluto del determinante de $\begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix}$, por lo que

$$\begin{aligned} A_{abc} &= \frac{1}{2} \left| \det \left(\begin{bmatrix} x_b - x_a & x_c - x_a \\ y_b - y_a & y_c - y_a \end{bmatrix} \right) \right| \\ &= \frac{1}{2} \left[(x_b - x_a) \times (y_c - y_a) - (x_c - x_a) \times (y_b - y_a) \right] \\ &= \frac{1}{2} \left[x_a y_b + x_b y_c + x_c y_a - x_a y_c - x_b y_a - x_c y_b \right] \end{aligned}$$

Existe, además, un método más eficiente. La triangularización no debe ser necesariamente una partición, sino que podemos definir triángulos con orientación positiva y negativa (con áreas positivas y negativas), tal y como se hace al calcular el área de un polígono.

Esta estrategia nos lleva a un algoritmo muy simple para calcular el centroide, basado en la suma de los centroides de los triángulos ponderados por su área con signo. Los triángulos pueden tomarse de tal modo que sean los formados por cualquier punto fijo, por ejemplo el vértice v_0 del polígono, y los dos puntos vértices opuestos de dos aristas consecutivas del polígono: $(v_1, v_2), (v_2, v_3)$, etc. Hecho esto, sólo resta estandarizar respecto al área del polígono, que puede calcularse de forma sencilla empleando

$$A_{1:n} = \frac{1}{2}\sum_{i=1}^{n}(x_i y_{i+1} - x_{i+1}y_i) \text{ con } x_{n+1} = x_1 \text{ y } y_{n+1} = y_1.$$

También podemos generar una visualización de los datos por semestre a través de predicción por Kriging. Tomaremos los centroides de los municipios así como los conteos de eventos violentos observados en tal municipio en cierto semestre y dividiremos por el número de habitantes estimados en el municipio en dicho semestre.

En la Figura 4.8 podemos observar la ubicación de los centroides para cada uno de los municipios. Destacamos que en general todos parecen ser buenos representantes de la localización espacial de sus respectivos municipios, salvo un municipio en Yucatán que al tener una parte en la península y otra en una isla el centroide queda en el océano. Esto no es necesariamente incorrecto, pues puede seguirnos dando información razonable respecto a la localización espacial.

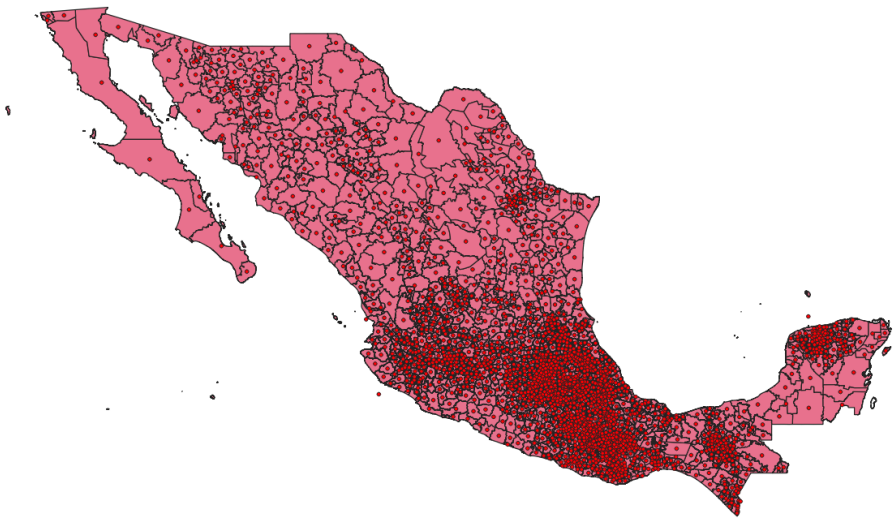


Figura 4.8: Centroide calculado para cada municipio

Como vimos en la Sección 1.5, antes de poder calcular los ponderadores de los predictores lineales de Kriging debemos elegir un modelo de semivariograma sensato para los datos y estimar sus parámetros. Al principio, se propuso ajustar los semivariogramas

esférico, Gaussiano y exponencial, pero al ajustar los semivariogramas esférico y exponencial aparecieron errores numéricos al estimar los parámetros. Por ello, nos limitaremos a analizar el ajuste del semivariograma esférico para los tres tipos de eventos violentos.

En la Figura 4.9 podemos ver que en distancias cortas el semivarigrama empírico esta por debajo del estimado y en general la forma es muy diferente. Esto nos hace sospechar que el ajuste no es muy bueno y no debemos tomarnos muy en serio las predicciones del mapa de calor para este modelo.

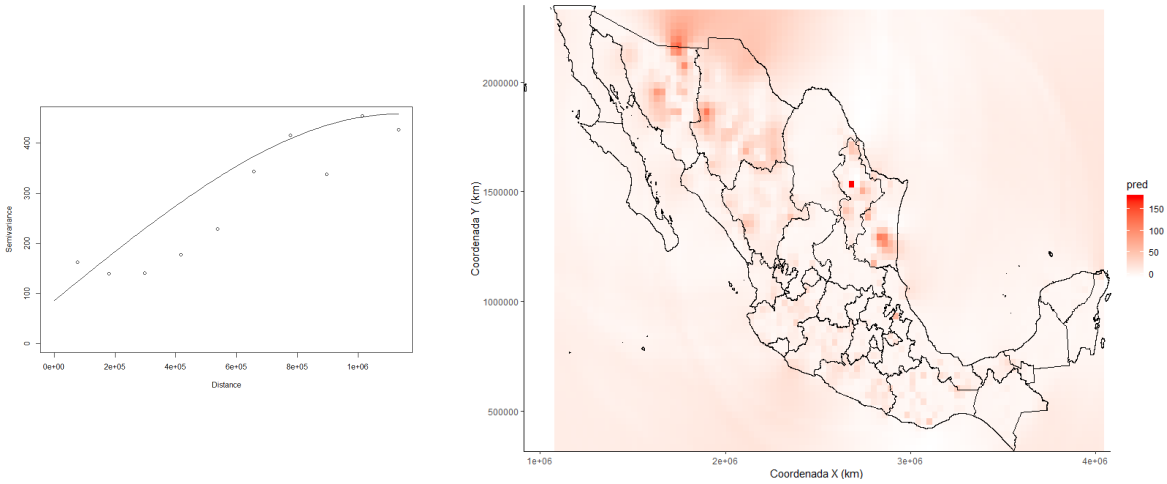


Figura 4.9: Semivariograma y predicción de Kriging Esférico para Ejecuciones noveno semestre

Mientras que en la Figura 4.10 podemos ver que la forma del semivariograma estimado es un poco más parecida a la del estimado, al menos respecto al caso de Ejecuciones. Las zonas de mayor riesgo de enfrentamientos parecen conglomerarse en la región del noroeste del país.

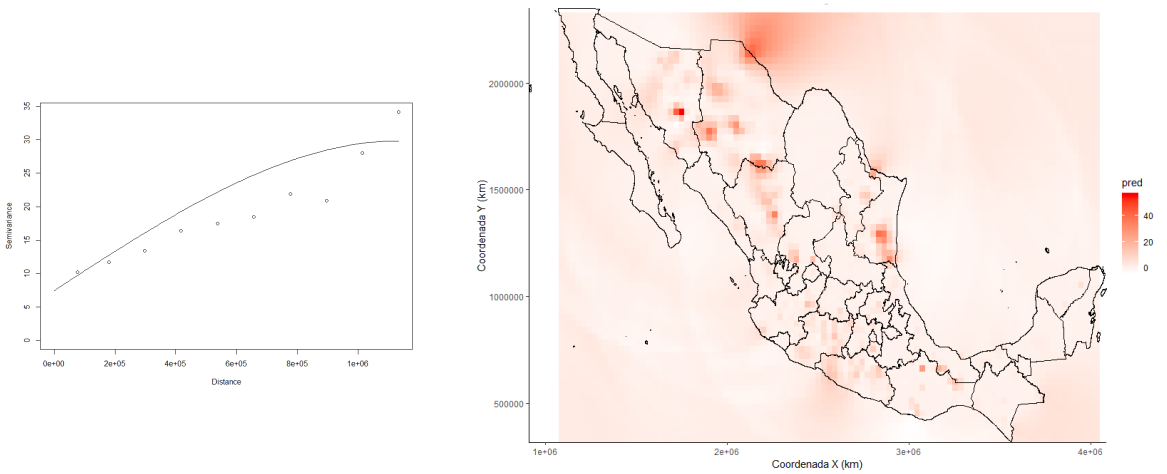


Figura 4.10: Semivariograma y predicción de Kriging Esférico para Enfrentamientos noveno semestre

Por último, en la Figura 4.11 podemos ver que la forma del semivariograma estimado

también es parecida a la del semivariograma empírico. Al igual que en el modelo para Enfrentamientos, las zonas de riesgo se encuentran en la región del noroeste.

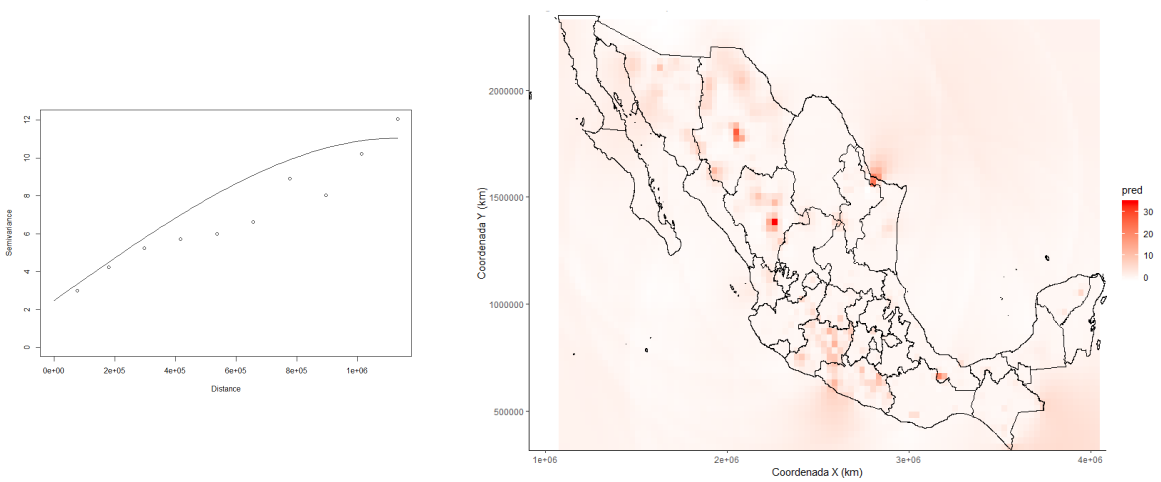


Figura 4.11: Semivariograma y predicción de Kriging Esférico para Agresiones noveno semestre

En la Tabla 4.1 podemos ver los parámetros estimados del modelo de semivariograma exponencial para los tres tipos de eventos violentos. Recordemos que este ajuste se hace minimizando el error cuadrático o lineal que describe el semivariograma empírico y la curva que describe el semivariograma teórico (modelo esférico). Es interesante observar que el efecto nugget y silla en ejecuciones es muy alto en relación a los enfrentamientos y agresiones.

Evento	Efecto-Nugget (c/100,000 hab.)	Rango (km)	Silla (c/100,000 hab.)
Ejecuciones	85.81617	1132806	458.3475
Enfrentamientos	7.44779	1132806	29.82825
Agresiones	2.476736	1132806	11.04786

Tabla 4.1: Cantidades de interés del variograma estimadas

4.4. Modelos jerárquicos preliminares propuestos

Como se menciona en el Capítulo 1, los datos espaciales se definen como realizaciones de un proceso estocástico indexado por el espacio

$$Y(s) \equiv \{y(s), s \in D\}$$

donde D es un subconjunto (fijo) de $\mathbb{R}^d$ ($d = 2$). Los datos observados son representados por una colección de observaciones $\boldsymbol{y} = \{y(s_1), ..., y(s_n)\}$, donde el conjunto $(s_1, ..., s_n)$ indica las unidades espaciales en las que fueron hechas las mediciones. Notemos que D

puede ser una superficie continua o bien un conjunto contable de unidades espaciales $d-$dimensionales.

Por ejemplo, podrían representar la localización puntual en la que se encuentran medidores o bien podrían tratarse de áreas (distritos, municipios o estados) en lugar de puntos. En nuestro ejemplo a nivel estatal $n = 32$, para la i-ésima área, el número de eventos violentos (en un periodo de tiempo) y_i puede ser modelado como

$$y_i \sim \text{Poisson}(\lambda_i),$$

donde la media λ_i se define en término de la tasa ρ_i y el número esperado de eventos E_i como $\lambda_i = \rho_i E_i$.

El modelo se considera

$$O_i \sim \text{Poisson}(\mu_i), i = 1, ..., 32$$

$$\log(\mu_i) = \log(E_i) + \beta_0, i = 1, ..., 32$$

Aquí, O_i es el número de eventos violentos en el estado i, E_i es el número esperado de casos. El número de casos se incluye con un peso (*offset*) para tomar en cuenta la distribución desigual de la población a través de los estados (o municipios). Notemos que está asociado al valor esperado de la variable Y_i y tiene un efecto multiplicativo. Podríamos llamar *offset* neutro al caso particular donde $E_i = 1$ para todo $i = 1, ..., n$. En el caso de los eventos violentos, la población por área puede jugar un papel importante, por lo que debemos fijar los *offset*'s de forma adecuada. Una posibilidad es fijarlos de modo que multipliquemos la tasa total (nacional) de eventos por la población del estado i (P_i), es decir,

$$E_i = P_i \times \left(\frac{\sum_{i=1}^{32} O_i}{\sum_{i=1}^{32} P_i} \right)$$

Igual que como se hace en los modelos lineales generalizados, el predictor lineal se define en la escala logarítmica

$$\mu_i = \log(\rho_i) = \alpha + \upsilon_i + \nu_i,$$

donde α es el intercepto que cuantifica la tasa de eventos promedio en los 32 estados; $\upsilon_i = f_1(i)$ y $\nu_i = f_2(i)$ son dos efectos específicos del área; $i = \{1, ..., n\}$ es el indicador de cada estados (áreas espaciales).

Suponemos una especificación Besag-York-Mollie (BYM) (Besag et al. 1991), tal que υ_i es el residual estructurado espacialmente modelado usando una estructura autorregresiva condicional intrínseca (iCAR)

$$\upsilon_i | \upsilon_{j \neq i} \sim \text{Normal}(m_i, s_i^2)$$

donde

$$m_i = \frac{\sum_{j \in \mathcal{N}(i)} \upsilon_j}{\#\mathcal{N}(i)} \quad \text{y} \quad s_i^2 = \frac{\sigma_\upsilon^2}{\#\mathcal{N}(i)},$$

y $\#\mathcal{N}(i)$ es el número de áreas que comparten bordes con la i-ésima área. El parámetro ν_i representa el residual no estructurado, modelado usando una previa intercambiable $\nu_i \sim \text{Normal}(0, \sigma_\nu^2)$.

Vale a pena destacar que R-INLA parametriza con $\xi_i = \upsilon_i + \nu_i$ y υ_i. Si no se especifica de otra forma, por defecto se fijan previas mínimo informativas en el logaritmo de la precisión del efecto no estructurado ($\tau_\nu = 1/\sigma_\nu^2$)

$$\log \tau_\nu \sim \log \text{Gamma}(1, 0.0005)$$

y también en el logaritmo de la precisión del efecto estructurado

$$\log \tau_\upsilon \sim \log \text{Gamma}(1, 0.0005).$$

Así, los parámetros estimados por INLA son representados por $\theta = \{\alpha, \xi, \upsilon\}$ y los hiperparámetros están dados por las presiones $\psi = \{\tau_\upsilon, \tau_\nu\}$.

Puede ser de interés evaluar la proporción de la varianza explicada por la componente estructurada espacialmente. La cantidad σ_υ^2 es la varianza de la especificación autorregresiva condicional, mientras que σ_ν^2 es la varianza de la componente marginal no estructurada. Por lo tanto, estas dos variables no son directamente comparables. Sin embrago, es posible obtener un estimado de la varianza posterior marginal para el efecto estructurado empírico a través de

$$s_\upsilon^2 = \frac{\sum_{i=1}^{n} (\upsilon_i - \bar{\upsilon})^2}{n-1}$$

donde $\bar{\upsilon}$ es el promedio de las υ_i y luego se compara con la varianza posterior marginal del efecto no estructurado, dado por σ_ν^2

$$\text{frac}_{\text{espacial}} = \frac{s_\upsilon^2}{s_\upsilon^2 + \sigma_\nu^2}.$$

A continuación se describen los modelos que servirán como análisis exploratorio de los datos de nuestro problema de violencia en México. Estos han sido propuestos en Blangiardo, Cameletti, Baio & Rue, 2013.

Modelo A: La formulación clásica de modelos con componente espacial y temporal asume que el predictor lineal se puede escribir

$$\eta_{it} = \alpha + \upsilon_i + \nu_i + (\beta + \delta_i) \times t.$$

Al igual que en el modelo únicamente espacial, la componente espacial se agrega en $\xi_i = \upsilon_i + \nu_i$; tenemos una tendencia lineal β que representa el efecto temporal global, y una tendencia diferencial δ_i, que identifica la interacción entre tiempo y espacio.

Con el propósito de tener identificabilidad, imponemos una restricción de suma-cero

en $\boldsymbol{\delta}$ y $\boldsymbol{\nu}$, por lo que el término δ_i representa la diferencia entre la tendencia global β y la tendencia del área en cuestión. Si $\delta_i < 0$ entonces el área i tiene menos pendiente que la tendencia media, mientras que $\delta_i > 0$ implica una tendencia más inclinada que la media. Asumimos $\delta_i \sim \text{Normal}(0, \tau_\delta)$ pero también es posible usar una estructura autorregresiva condicional.

Este modelo supone un efecto lineal del tiempo para cada área (δ_i). Los parámetros a estimar son $\boldsymbol{\theta} = \{\alpha, \beta, \boldsymbol{\xi}, \boldsymbol{\upsilon}, \boldsymbol{\delta}\}$ y los hiperparámetros $\boldsymbol{\psi} = \{\tau_\upsilon, \tau_\nu, \tau_\delta\}$.

Modelo B: El supuesto de linealidad en δ_i puede relajarse usando una formulación dinámica no paramétrica para el predictor lineal

$$\nu_{it} = \alpha + \upsilon_i + \nu_i + \gamma_t + \phi_t.$$

El término γ_t representa el efecto temporal estructurado, modelado de forma dinámica (por ejemplo, usando una caminata aleatoria) en una estructura de vecinos.

$$\begin{aligned}
\gamma_t|\gamma_{-t} &\sim \text{Normal}(\gamma_{t+1}, \tau_\gamma) \text{ para } t = 1 \\
\gamma_t|\gamma_{-t} &\sim \text{Normal}\left(\frac{\gamma_{t+1} + \gamma_{t-1}}{2}, \frac{\tau_\gamma}{2}\right) \text{ para } t = 2, ..., T-1 \\
\gamma_t|\gamma_{-t} &\sim \text{Normal}(\gamma_{t-1}, \tau_\gamma) \text{ para } t = T
\end{aligned}$$

Aquí ϕ_t se define como una previa Normal intercambiable

$$\phi_t \sim \text{Normal}(0, \tau_\phi).$$

Así $\boldsymbol{\theta} = \{\alpha, \boldsymbol{\xi}, \boldsymbol{\upsilon}, \boldsymbol{\gamma}, \boldsymbol{\phi}\}$ y los hiperparámetros $\boldsymbol{\psi} = \{\tau_\upsilon, \tau_\nu, \tau_\gamma, \tau_\phi\}$.

Modelo C: Es fácil expandir este modelo para que considere una interacción entre espacio y tiempo. La especificación es

$$\nu_{it} = \alpha + \upsilon_i + \nu_i + \gamma_t + \phi_t + \delta_{it}.$$

Existen varias formas de definir el término de interacción. Aquí, asumimos que los dos efectos no estructurados ν_i y ϕ_t interactúan. Reescribimos la matriz de precisión como el producto escalar τ_ν (o τ_ϕ) y la llamada matriz de estructura F_ν (F_ϕ), que identifica la estructura vecindante; aquí la matriz de estructura F_δ puede factorizarse como el producto de Kronecker de las matrices de estructura para $\boldsymbol{\nu}$ y $\boldsymbol{\phi}$: $F_\delta = F_\nu \otimes F_\phi = I \otimes I = I$ (ya que ambos $\boldsymbol{\nu}$ y $\boldsymbol{\phi}$ son no estructurados). Consecuentemente no asumimos estructura ni espacial ni temporal en la interacción y por lo tanto $\delta_{it} \sim \text{Normal}(0, \tau_\delta)$.

Por lo que en este modelo sus parámetros son $\boldsymbol{\theta} = \{\alpha, \boldsymbol{\xi}, \boldsymbol{\upsilon}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\delta}\}$ y sus hiperparámetros corresponden a $\boldsymbol{\psi} = \{\tau_\upsilon, \tau_\nu, \tau_\gamma, \tau_\phi, \tau_\delta\}$.

Resultados de modelos preliminares propuestos

Implementamos los modelos A, B y C, con lo que podemos hacer la inferencia de los parámetros e hiperparámetros se hizo a través de R-INLA. Podemos obtener las distribuciones marginales de cualquier parámetro de interés y cualquier integral que nos interese. En este caso, vamos a modelar el fenómeno de Ejecuciones.

Con el fin de poder comparar el ajuste de los modelos podemos usar el DIC descrito en la sección de metodología. Como el AIC, preferimos modelos con bajo DIC, y tenemos que es notablemente menor el DIC del modelo C respecto al DIC de los otros dos modelos A y B. Además, tenemos el número de parámetros efectivos calculado, que nos indica que el modelo C es mucho más complejo que los otros dos, sin embargo, el mejor ajuste compensa este aumento en la complejidad, pues el DIC es bastante menor.

Observemos en la Tabla 4.2 que el número de parámetros efectivos estimados (Eff/ P_D), es decir que tantos "grados de libertad" tiene el campo aleatorio latente, es mucho mayor en el modelo C. Esto se debe a que es un modelo más flexible respecto a los efectos aleatorios. También vemos que el valor de la log-verosimilitud es más negativo en los primeros dos modelos. Estas medidas apoyan el uso del modelo C. Por último, en las últimas columnas aparecen los estadísticos de pruebas y p-valores de pruebas de bondad de ajuste aplicada a las PIT's respecto a una Uniforme$(0, 1)$. Los detalles de estas pruebas, que son la prueba χ^2 de Pearson y la prueba de Kolmogorov-Smirnov, se encuentran en el Capítulo 2. Como vemos, todos los p-valores son tan pequeños que las computadora los reporto como ceros numéricos (menores a $2.2e-16$). Sin embargo, podemos comparar los valores de los estadísticos de prueba para ver que tan distantes son de un buen modelo. Tanto el estadístico de la prueba χ^2 de Pearson como estadístico de la prueba de Kolmogorov-Smirnov son mucho más cercanos al que esperaríamos si se cumple la hipótesis de venir de una distribución Uniforme$(0, 1)$. Los modelos no son buenos, pero definitivamente el menos malo es el modelo C.

Evento	DIC	Eff(P_D)	WAIC	logMLIK	D_{KS}	p_v KM
Modelo A Lineal en T	5250.58	62.05	7488.73	-3371.676	0.257	$<$ 2.2e-16
Modelo A Lineal en T	7770.01	34.19	10565.27	-4835.26	0.330	$<$ 2.2e-16
Modelo C RW en T (Int.)	2330	273.60	2275	-1535.68	0.164	6.082e-08

Tabla 4.2: Criterios de selección de modelos para Ejecuciones

Respecto a Enfrentamientos en la Tabla 4.3 vemos que los modelos tipo B tanto con salida Poisson como ZIP tienen un pobre ajuste y sus PIT's distan de ser normales. Los

modelos con tiempo lineal (A) y espacio temporales con interacción (C) son parecidos tanto en ajuste como en bondad de ajuste según los PIT's. En general el modelo Poisson se comporta mejor que el ZIP, lo cuál es razonable ya que la proporción de ceros no es muy alta. De modo que el mejor modeo para este conjunto de datos es el Modelo C con variable de salida Poisson, tanto en ajuste como por prueba de uniformidad de los PIT's.

Evento	DIC	Eff(P_D)	WAIC	logMLIK	D_{KS}	p_v KM
Modelo A Poisson Lineal en T	1638.38	54.45	1715.75	-921.43	0.1024	0.0024
Modelo A ZIP Lineal en T	1764.69	48.24	1838.53	-972.81	0.1142	0.0004
Modelo B Poisson RW en T (Int.)	2149.99	31.19	2258.64	-1126.11	0.1472	1.889e-06
Modelo B ZIP Lineal en T	2182.86	29.93	2284.78	-1140.22	0.1249	9.188e-05
Modelo C Poisson Lineal en T	1485.45	190.97	1463.85	-883.32	0.1188	0.0002
Modelo C ZIP RW en T (Int.)	1606.08	161.73	1586.30	-930.94	0.10093	0.0029

Tabla 4.3: Criterios de selección de modelos para Enfrentamientos

Finalmente, al modelar las Agresiones vemos en la Tabla 4.4 que el mejor ajuste es dado por el modelo C con variable de salida Poisson. En este caso, vemos que los criterios de selección de modelos no coinciden ya que si bien el ajuste es relativamente mejor sus PIT's modificados distan mucho de una distribución uniforme. En el caso de agresiones, los modelos con variable de salida ZIP tienen un ajuste peor, sin embargo los PIT's modificados se acercan mucho más a la uniformidad. Esto último es muy razonable ya que el porcentaje de ceros en Agresiones es considerablemente mayor respecto al de los otros fenómenos; cerca de la mitad de las observaciones son ceros.

Evento	DIC	Eff(P_D)	WAIC	logMLIK	D_{KS}	p_v KM
Modelo A Poisson Lineal en T	1190.76	45.84	1265.30	-673.81	0.1212	0.0001
Modelo A ZIP Lineal en T	1374.78	39.17	1446.81	-751.68	0.0830	0.0245
Modelo B Poisson RW en T (Int.)	1302.78	29.43	1367	-686.27	0.1234	0.0001
Modelo B ZIP Lineal en T	1457.55	27.93	1518.92	-762.95	0.0880	0.0140
Modelo C Poisson Lineal en T	1000.5	128.69	987.8	-587	0.18647	4.33e-10
Modelo C ZIP RW en T (Int.)	1202.35	98.93	1192.68	-677.82	0.1150	0.0004

Tabla 4.4: Criterios de selección de modelos para Agresiones

4.5. Ajuste de modelos jerárquicos espacio-temporal con covariables para datos de violencia en México

En contraste con la Sección 4.4 en ésta exploramos las covariables socio-económicas de las que disponemos y las usaremos en modelos con efectos fijos y aleatorios (espacio temporales) ajustados mediante INLA. Como en la sección anterior, el código usado para generar los distintos análisis pueden encontrarse en el siguiente enlace al repositorio en Github https://github.com/mcarranzba/Tesis_INLA.

Hasta ahora solo habíamos considerado modelos con efectos aleatorios. Al introducir covariables para cada uno de las áreas (estados) para cada una de las divisiones temporales tenemos un modelo con efectos mixtos.

Es interesante para los investigadores dentro del área de políticas públicas el efecto que tienen algunos factores socio-económicos. Se recolectaron datos sobre indicadores socio-económicos en el periodo de interés en repositorios de datos abiertos de instituciones como el INEGI y CONAPO. Las que consideramos de interés incluyen incluyen: Coordenadas espaciales y su interacción; porcentaje de desempleo; porcentaje de ocupados en los giros de: Agricultura, Extracción y electricidad, Manufactura, Construcción, Comercio, Restaurantes y hoteles, Transporte y comunicaciones, Servicios profesionales, Servicios sociales, Servicios diversos, Gobierno; Cohesión Social (Coeficiente de Gini); Pobreza alimentaria, Pobreza de capacidades, Pobreza de patrimonio.

Es útil poder visualizar las covariables para identificar que áreas (estados) se parecen entre si. En la Figura 4.12 empleamos gráficas de estrella (o telas de araña) para visualizar la distribución en los distintos giros de la actividad económica para todos los estados. En general es difícil encontrar estados que tengan la misma distribución en los giros sin embargo existen algunos que se parecen entre si. Por ejemplo, Chiapas y Oaxaca tienen diagramas parecidos.

Como en cualquier modelo de regresión lineal (donde se tiene una componente del modelo que es combinación lineal de las covariables) debemos procurar que las covaribles no presenten multicolinealidad, es decir, que una o más covariables puedan ser expresadas como combinaciones lineales del resto. Incluso si esta relación lineal no es perfecta, estar cerca, digamos con un ruido pequeño, puede dar problemas numéricos (particularmente al invertir la matriz de varianza empírica en la solución de máxima verosimilitud o de mínimos cuadrados para el modelo lineal). Esto puede considerarse como un caso particular de no identificabilidad.

Sea $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ un modelo estadístico donde el espacio parametrico Θ es finito o infinito dimensional. Decimos que $\mathcal{P}$ es identificable si el mapeo $\theta \mapsto P_\theta$ es inyectiva o uno a uno, es decir,

$$P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2 \text{ para todo } \theta_1, \theta_2 \in \Theta.$$

En este caso, diferentes combinaciones de coeficientes pueden dar predictores similares, y por tanto la variabilidad en la inferencia de estos coeficientes puede ser innecesariamente

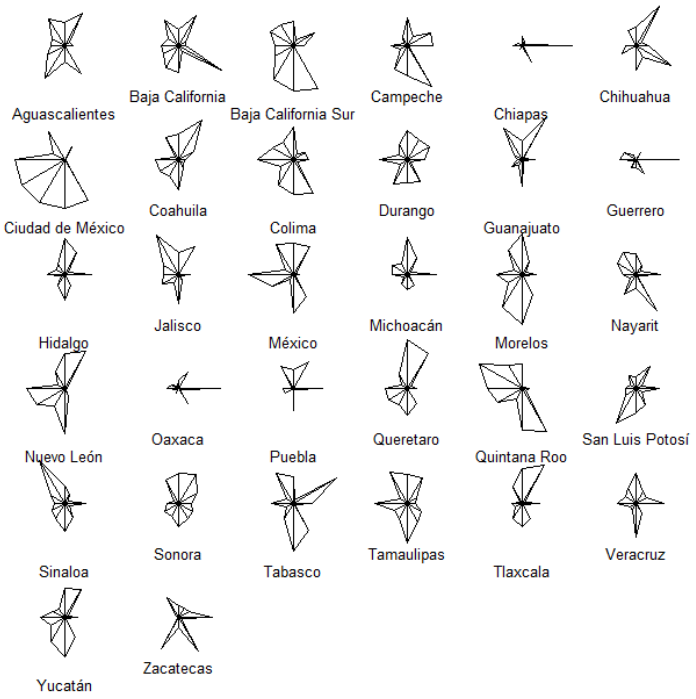


Figura 4.12: Diagrama de estrella para las proporciones de giro industrial en los estados

alta (medida como varianza de los estimadores o bien la varianza de las distribuciones posteriores). Podemos identificar las dependencias lineales dos a dos graficando la matriz de correlación (de Pearson) de las covariables. Sin embargo, esto no detecta cuando una variable puede expresarse como combinación lineal de varias covariables.

En nuestro ejemplo práctico, este problema se soluciona al retirar las variables que están muy correlacionadas entre si en el sentido dos a dos, y respecto a las proporciones de ocupación basta quitar la categoría de no especificado (que esta determinado por el resto de componentes). Adicionalmente, es posible que debido a ligeras variaciones variables que son, linealmente, muy parecidas entre sí (tienen una correlación lineal cercana a 1 o -1) se prefieran unas sobre otras. Sin embargo, estas pequeñas variaciones no son muy importantes para nuestro problema de linealidad múltiple y conviene quedarnos con aquella que tenga mayor sentido de interpretación. Por ejemplo, en nuestro ejercicio solo incluimos la variable de pobreza alimentaria, ya que tanto la pobreza de capacidades y la pobreza patrimonial están muy correlacionadas.

Por otro lado, recordemos que para los eventos de enfrentamientos y agresiones muchos de los conteos resultaron cero incluso a nivel estatal. Por ello conviene considerar modelos (a nivel de observaciones) que tengan mayor probabilidad de ceros, tales como los modelos Poisson Cero Inflados (ZIP) o los modelos Hurdle.

Respecto a la especificación de la componente espacial decidimos apegarnos al modelo de Besag. Mientras que para la parte temporal comparamos los modelos autoregresivos de

orden 1 contra un modelo de caminata aleatoria de orden 1. Por ejemplo, en el caso de un modelo con componente temporal de tipo caminata aleatoria el modelo es de la forma:

Nivel	Nombre	Especificación
I	Poisson Cero Inflado (exceso de ceros)	$\Pr(y_{it}=0) = p_{it} + (1-p_{it})e^{-\lambda_{it}}$ con $\lambda_{it} = \exp(\nu_{it})$ $\Pr(y_{it}=k) = (1-p_{it})\frac{\lambda_{it}^k e^{-\lambda_{it}}}{k!}, \qquad k \geq 1$
II	Campo aleatorio latente (Efectos fijos, Besag y caminata aleatoria)	$\nu_{it} = \alpha + \beta X + \upsilon_i + \nu_i + \gamma_t + \phi_t + \delta_{it}$, $\phi_t \sim \text{Normal}(0, \tau_\phi)$ $\upsilon_i \vert \upsilon_{j\neq i} \sim \text{Normal}(m_i, T_i)$, $m_i = \frac{\sum_{j\in\mathcal{N}(i)}\upsilon_j}{\#\mathcal{N}(i)}$ y $T_i = \frac{\tau_\upsilon}{\#\mathcal{N}(i)}$ $\gamma_t \vert \gamma_{-t} \sim \text{Normal}\left(\frac{\gamma_{t+1}+\gamma_{t-1}}{2}, \frac{\tau_\gamma}{2}\right)$ para $t = 2, ..., T-1$
III	Hiperparámetros	$p(\beta) \sim \text{Normal}(\mu, \Sigma_0), \tau_k \sim \text{Gamma}(\alpha_k, \beta_k)$ para $k = \upsilon, \gamma, \phi, \delta$

4.6. Resultados a nivel estatal

A continuación veremos los resultados de los modelos con covariables aplicados a los tres fenómenos de interés. Respecto a los niveles de pobreza, estos tienen una muy alta correlación por pares, por lo que basta y es mejor tomar sólo la pobreza alimentaria. Deseamos poder comparar estos modelos en términos de indicadores de ajuste como la log-verosimilitud marginal y los criterios basados en distribuciones predictivas de tipo "Leave One Out", tales como indicadores de la uniformidad de los PIT reflejado por el estadístico de Kolmogorov-Smirnov.

Vemos en la Tabla 4.5 que los criterios DIC y WAIC dan en general valores muy parecidos. Adicionalmente, la log verosimilitud marginal también mantiene una relación constante con estos valores. Esto es de esperarse ya que todas estas cantidades tratan de aproximar un mismo valor común. Observamos que en general este valor es alto para los modelos ZIP. Esto se debe a que por la forma en que se implementan los modelos ZIP en INLA consideran el doble de datos en la verosimilitud (uno para la parte Bernoulli y otro para la parte de la Poisson). Esto puede tiene solución clara en la teoría, pero implementarlo en la práctica requiere modificar los algoritmos dentro de INLA debido a complicada estructura del modelo. Una alternativa a estas cantidades es la suma de los logaritmos de los CPO. Ya que tenemos la contribución de cada una de las observaciones por separado fácilmente puede encontrarse el CPO y PIT correspondiente al modelo ZIP.

Podemos usar el DIC, WAIC y logMLIK para comparar los modelos según sus diferencias en cuanto estructura, mientras que la suma de log CPO's (SLCPO) y el estadístico $\mathcal{D}$ de Kolmogorov-Smirnov permite comparar las dos variables de salida. Tras comparar DIC, WAIC y logMLIK, vemos que suelen dar las mismas conclusiones sobre la comparación

	Evento	Estructura	Salida	DIC	Eff_PD	WAIC	logMLIK	SLCPO	D_KS	lpv_KM
1	Ejec	Fijos	Pois	751.06	17.00	739.61	-457.60	-369.83	0.19	-9.25
2	Ejec	Fijos	ZIP	27733.98	18.23	40370.87	-18333.82	-1113.91	0.46	-Inf
3	Ejec	Aleat	Pois	709.03	30.63	682.30	-381.06	-341.18	0.29	-Inf
4	Ejec	Aleat	ZIP	2376.05	271.94	2318.05	-1566.53	-1101.69	0.18	-Inf
5	Ejec	Ambos	Pois	719.04	22.75	700.70	-442.33	-350.38	0.25	-Inf
6	Ejec	Ambos	ZIP	2374.08	269.85	2316.49	-1628.85	-1094.67	0.15	-12.16
7	Enfr	Fijos	Pois	754.22	17.00	747.22	-457.45	-373.68	0.13	-4.24
8	Enfr	Fijos	ZIP	3388.03	18.02	3795.77	-1874.74	-528.05	0.20	-Inf
9	Enfr	Aleat	Pois	665.74	29.03	644.06	-357.58	-322.08	0.23	-14.81
10	Enfr	Aleat	ZIP	1627.57	173.06	1613.45	-950.20	-555.04	0.12	-7.54
11	Enfr	Ambos	Pois	674.09	24.00	657.29	-421.77	-328.69	0.19	-9.64
12	Enfr	Ambos	ZIP	1629.66	169.09	1614.95	-1014.23	-548.72	0.11	-6.88
13	Agre	Fijos	Pois	578.14	17.00	572.52	-365.44	-286.39	0.08	-1.27
14	Agre	Fijos	ZIP	2149.04	18.01	2285.41	-1174.97	-266.29	0.19	-Inf
15	Agre	Aleat	Pois	536.13	19.61	525.21	-285.38	-262.66	0.10	-2.74
16	Agre	Aleat	ZIP	1269.04	127.92	1261.70	-728.30	-293.51	0.24	-Inf
17	Agre	Ambos	Pois	543.80	22.31	532.57	-350.22	-266.39	0.11	-3.29
18	Agre	Ambos	ZIP	1236.04	145.21	1223.97	-794.40	-314.56	0.26	-Inf

Tabla 4.5: Criterios de selección de modelos para Ejecuciones, Enfrentamientos y Agresiones

de los modelos, por lo que nos limitaremos a comentar lo observado con logMLIK que es la mejor aproximación al aprovechar toda la sofisticación de las aproximaciones de INLA.

Vemos en la Figura 4.13 que los modelos solo con efectos aleatorios son generalmente mejores que el resto de modelos, incluso que aquellos que incluyen ambos tipos de efectos. En el caso de variables de salida Poisson esta diferencia es moderada, mientras que en el caso de la variable de salida ZIP la ventaja de los modelos con efectos espacio-temporales respecto de los que sólo tienen efectos fijos es muy grande.

Mientras que en la Figura 4.14, que si admite comparación a través de variables de salida, vemos que el ajuste es considerablemente mejor con el modelo Poisson en ejecuciones. Esto posiblemente debido a casi no hay ceros en los datos de ejecuciones cuando se toman al nivel de agrupamiento considerado en estos análisis. El ajuste es mejor del modelo Poisson en el caso de enfrentamientos y parecido en el caso de agresiones, debido a la mayor presencia de ceros. Aquí el ajuste también es mejor con los modelos que solo tienen efectos aleatorios. Sin embargo, en el caso de agresiones hay un modelo sólo con covariables y salida ZIP que tiene un valor similar a sus contrapartes con efectos aleatorios.

En la Figura 4.15 vemos que en general los PIT modificados están lejos de distribuirse cerca a una Uniforme$(0, 1)$. Estos indicadores no son tanto un indicador del ajuste del modelo sino que tan razonablemente se cumplen los supuestos en cuanto distribución de los modelos. Curiosamente los PIT de las ejecuciones con modelo ZIP tienen una distribución más cercana a la normal, pero debido al pobre ajuste de estos modelos no los hace buenos candidatos. En cuanto a enfrentamientos los que más se acercan con los modelos Poisson con efectos aleatorios y el modelo ZIP sólo con covariables. Finalmente respecto a agresiones, los más cercanos son los modelos Poisson, y en especial aquel que sólo tiene covariables.

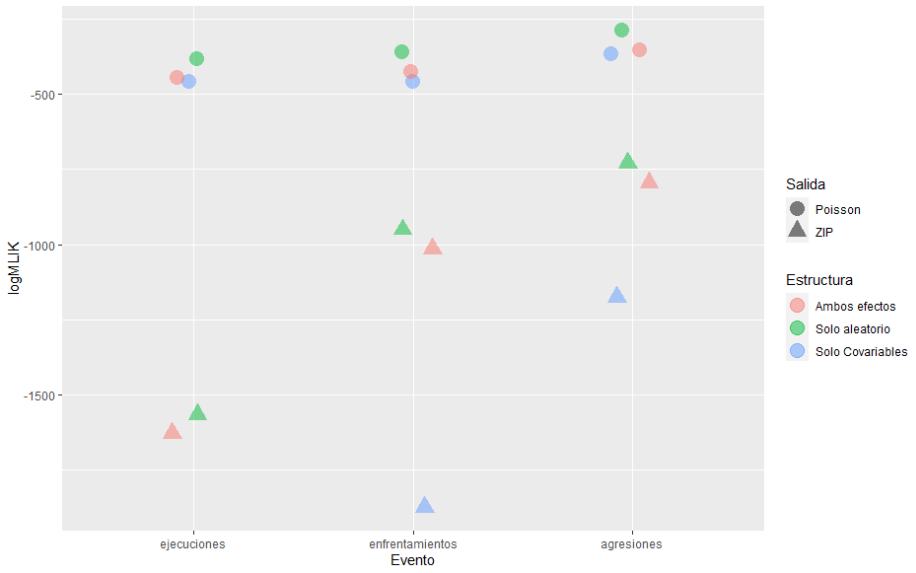


Figura 4.13: Comparación del la log verosimilitud marginal

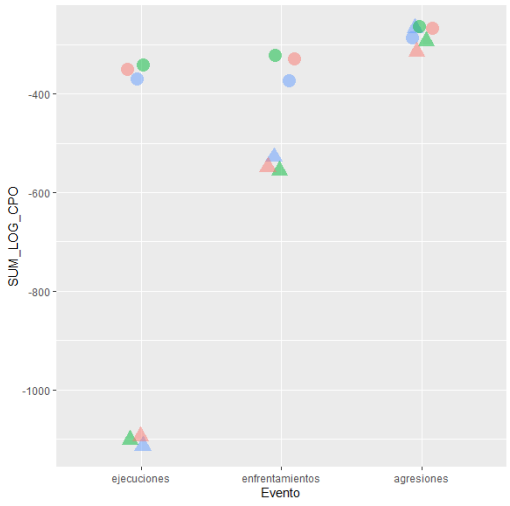


Figura 4.14: Comparación de la suma de log CPO

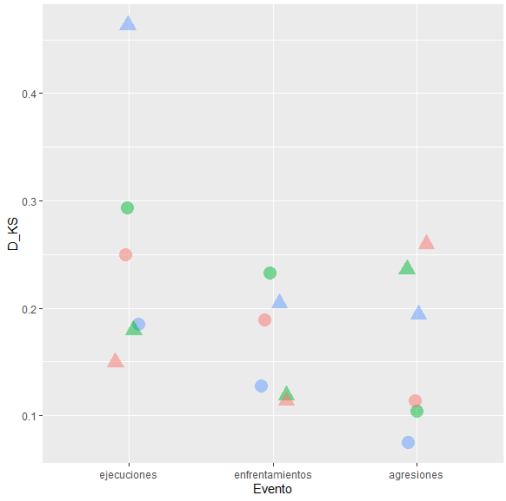


Figura 4.15: Comparación del estadístico de Kolmogorov-Smirnov

Tras comentar los indicadores parece que los modelos que tienen solo efectos aleatorios son mejores para predecir los datos observados y la incorporación de covariables con efectos fijos no aumenta esta poder predictivo. Sin embargo, conviene recordar que en un modelo estadístico otra característica deseable de un ajuste es la interpretabilidad de su modelo.

Una razón de porque la inclusión de covariables no aumenta el poder predictivo es que en realidad no puede predecirse mucho del fenómeno (debido a su complejidad) y lo poco que puede predecirse se puede explicar bien en términos espacio-temporales o de variables demográficas y económicas. Además conviene recordar que estas variables no son independientes y su contribución individual no es fácilmente separable. Esta incapacidad de separar la contribución de las covariables se ve acentuado por el número limitado de

datos con los que contamos.

Por otro lado, es interesante observar que tipo de información nos dan los coeficientes de covariables y tratar de darles una interpretación desde el punto de vista económico-social.

Vemos en la Figura 4.16 que la incertidumbre es mucho mayor cuando se tienen además efectos aleatorios. Esto es debido a que parte de la información que buscan explicar estas covariables es igualmente (o incluso mejor) explicada por los efectos aleatorios. Un ejemplo de esto es la covariable "ycoord" en el modelo ZIP, ya que se explica mejor por el efecto espacial local aunque hay una tendencia espacial a nivel nacional. Notemos que en el modelo Poisson las covariables no cambias mucho según la estructura, pero si lo hacen en el caso del modelo ZIP. En general los servicios (sociales y profesionales) y el coeficiente de Gini tienen un efecto negativo de forma consistente.

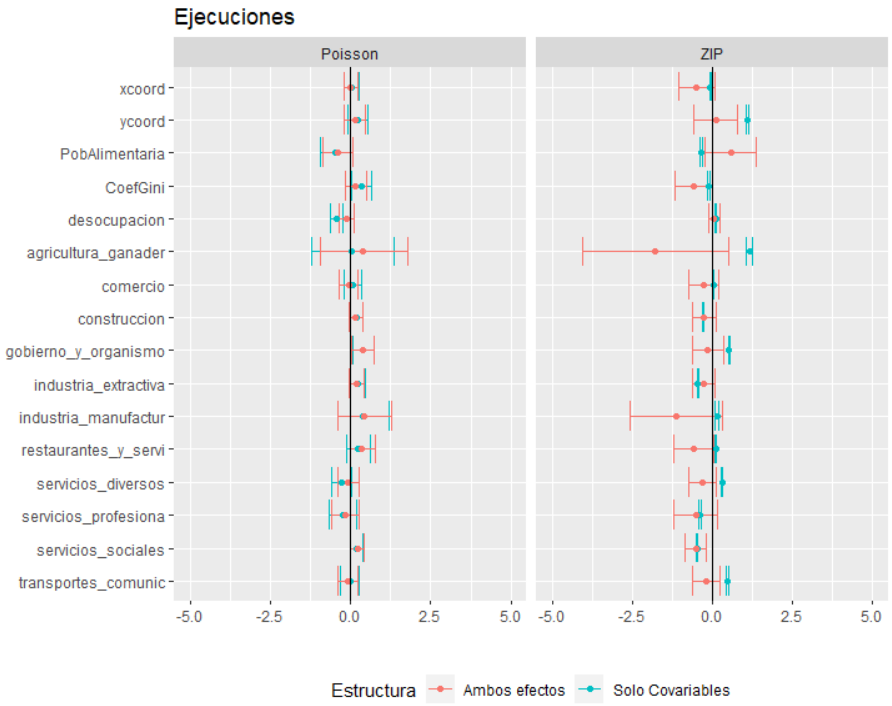


Figura 4.16: Comparación de los coeficientes de efectos fijos estandarizados en Ejecuciones

En cuanto a enfrentamientos, en la Figura 4.17 vemos que también los servicios profesionales y sociales tienen un efecto negativo consistentes. Además, en el caso del modelo Poisson vemos que al incluir los efectos aleatorios la agricultura y ganadería así como la industria y manufactura tienen potencialmente efectos negativos (reducen el número de enfrentamientos) importantes.

Finalmente, respecto a las agresiones en la Figura 4.18 vemos que en el modelo ZIP la pobreza alimentaria tiene un efecto positivo (a mayor pobreza más agresiones) y el coeficiente de Gini un efecto negativo (a mayor desigualdad menos agresiones). Este resultado es algo contra intuitivo pero parece que en este caso concreto se da esta relación.

Vale la pena destacar el caso de la pobreza alimentaria en el caso de los ZIP, cuyo efecto es positivo en ejecuciones y agresiones. Mientras que el Coeficiente de Gini tiene un efecto

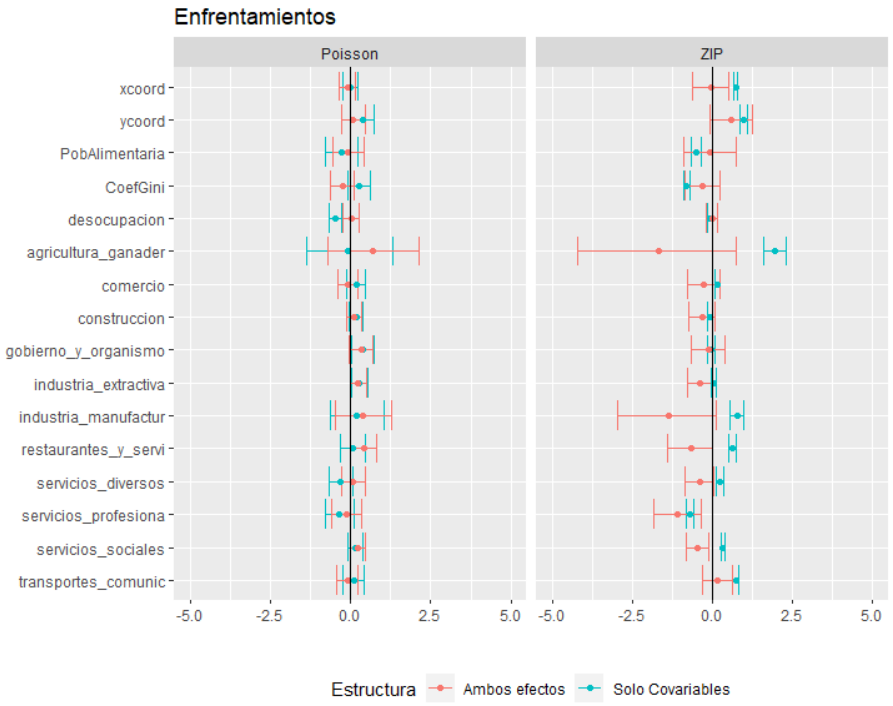


Figura 4.17: Comparación de los coeficientes de efectos fijos estandarizados en Enfrentamientos

negativo, también en el caso de ejecuciones y agresiones. Otra variable con un efecto relevante en los enfrentamientos y ejecuciones son los servicios profesionales y sociales. En la Figura 4.19 vemos la relación de Coeficiente de Gini y pobreza alimentaria, así como su evolución a través del tiempo. En general los estados con mayores eventos violentos (estados del norte) tienen coeficientes de Gini más bajos. La excepción es Guerrero, que a pesar de tener un indice de Gini y pobreza alimentaria altos también registro alto número de eventos violentos. Esta observación puede ayudarnos a darle sentido a la curiosa observación de los modelos para ejecuciones y agresiones.

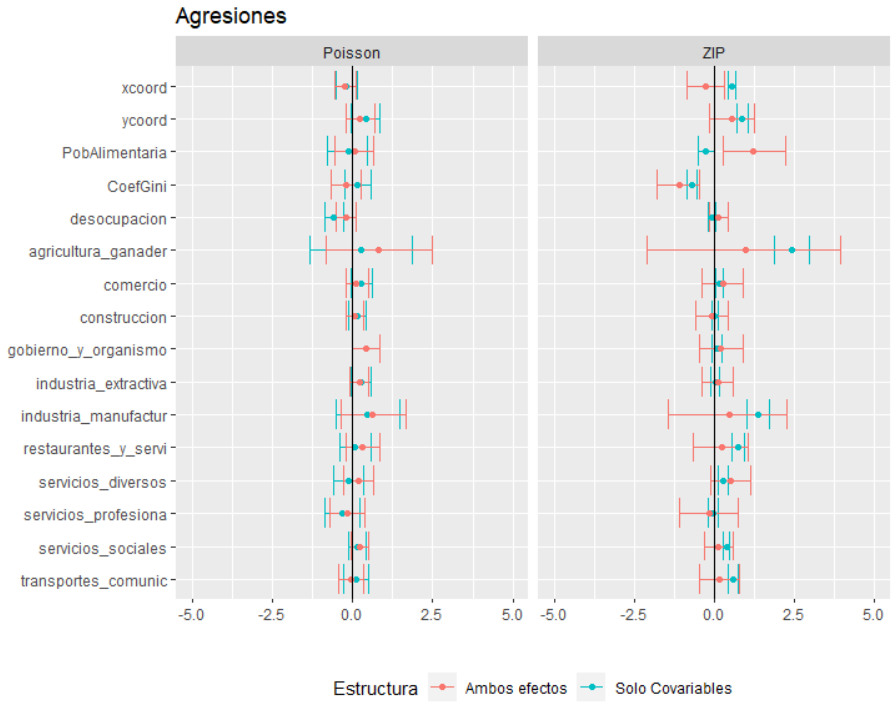


Figura 4.18: Comparación de los coeficientes de efectos fijos estandarizados en Agresiones

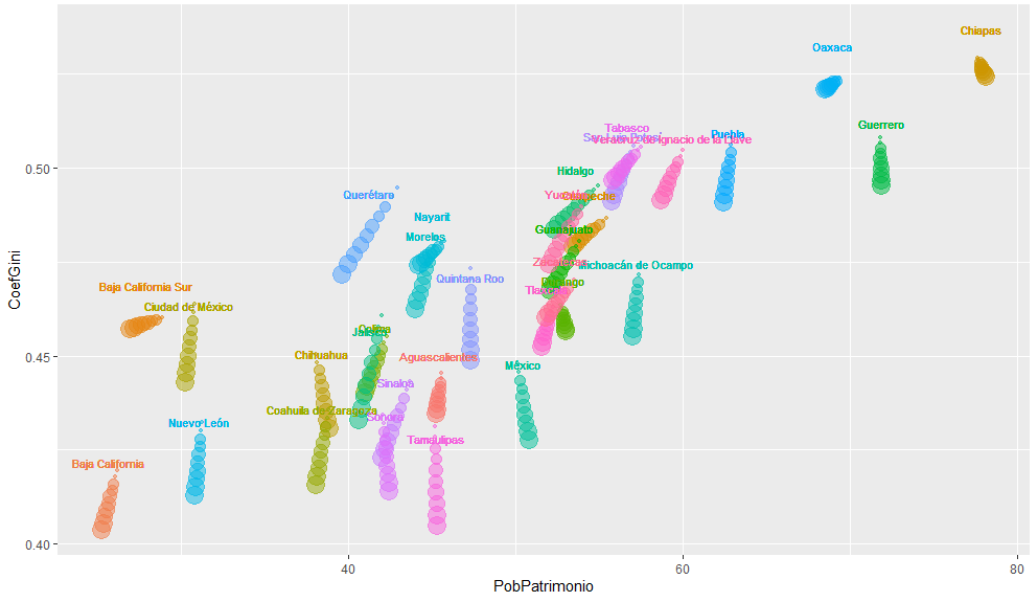


Figura 4.19: Evolución en distintos estados del porcentaje de pobreza alimentaria e índice de Gini

Discusión y trabajo futuro

En este trabajo, con el fin de analizar la base de datos asociada a los eventos violentos en México se proponen diversos modelos que permiten extraer distintas características de los fenómenos que nos interesa estudiar. Para ello desarrollaremos las definiciones y propiedades de estos modelos para así poder interpretarlos de forma adecuada.

De este modo, destacamos los dos grandes objetivos de este trabajo:

- Explicar teóricamente los modelos empleado para así entender las implicaciones de sus parámetros.
- Aplicar estos métodos a los datos de interés para proponer conclusiones a través de la interpretación correcta de sus parámetros y otros indicadores.

Respecto al primer objetivo pondremos especial atención al caso de la Aproximación de Laplace Integrada Anidada, ya que es un esquema relativamente nuevo y su método de inferencia es complejo, aunque el modelo puede describirse de forma relativamente sencilla.

Comentarios generales sobre INLA

En particular mencionamos como la forma rala de la matriz de covarianza (también de la matriz de precisión) permite emplear herramientas algorítmicas que reduzcan el tiempo de cálculo de las densidades marginales. Estos cálculos rápidos en INLA para modelos complejos supone una gran ventaja frente a los métodos MCMC. Sin embargo, en INLA los modelos están restringidos a los Campos Gaussianos Markovianos Latentes (CGML), mientras que los métodos basados en simulación son más flexibles. Esto puede representar un reto si se busca aplicar un modelo particular, pero al ajustar datos el catálogo de modelos disponibles en INLA es suficientemente amplio. Destacamos lo relativamente fácil que se pueden hacer los cálculos de criterios de selección de modelos, con principal interés en la aproximación de INLA de los criterios de tipo Leave One Out (LOO) tales como la

Ordenada Predictiva Condicional (CPO) y la Transformación Integral Predictiva. Vemos algunas ventajas que tienen sobre los otros criterios comúnmente usados, en especial su ventaja respecto al sobre-ajuste.

Una de las grandes ventajas que ofrece INLA es su velocidad de ejecución para modelos sumamente complejos. Su librería de modelos disponibles es muy completa pero de vez en cuando nos saldremos de lo que esta disponible de forma directa. A través de algunas opciones avanzadas como los modelos con funciones de verosimilitud conjunta es posible implementar modelos como los cero-inflados donde la probabilidad de tener ceros también depende del predictor lineal.

Comentarios generales de los datos y modelos

El segundo objetivo de la presente tesis es aplicar los modelos propuestos para analizar los datos de eventos violentos y así hacer posible la comprensión de las características del fenómeno a nivel global. Se propusieron tres enfoque principales, dos de ellos usando herramientas comunes de modelos espaciales y un enfoque que permite agregar una componente temporal y de covariables socio-económicas.

El enfoque de procesos puntuales y Kriging permiten visualizar el desarrollo del fenómeno en el espacio. Usando procesos puntuales podemos usar directamente los datos geolocalizados y mediante Kriging a nivel de conteos por municipio. Calculando los correspondientes mapas de calor vemos como hay una concentración (hacia el fina del sexenio) en el norte (Chihuahua) y suroeste (Guerrero) respecto a ejecuciones. Mientras que respecto a enfrentamientos y agresiones la concentración es principalmente hacia el noreste (Nuevo León y Tamaulipas).

El enfoque de los modelos espacio temporales en INLA se ve limitado a un análisis a nivel de estados por dos factores, la dificultad de obtener todas las covariables a un nivel de desagregación mayor y el hecho de que a pesar de las estrategias de INLA para hacer posibles los cálculos el número de municipios de toda la nación es demasiado grande y las matrices de precisión de los efectos aleatorios no pueden manejarse. En este enfoque ajustamos los datos de eventos violentos por población a través de offsets. Al comparar los modelos con covariables y sin covariables vemos que mucha de la información que logra explicar las covariables socio-económicas también pueden explicarse con los efectos espacio-temporales. Esto no debe extrañarnos ya que la cercanía entre estados claramente afecta sus variables socio-económicas (flujo de personas, bienes y capital). Debemos tener cuidado al interpretar los interceptos de cada uno de los modelos, ya que el coeficientes dado por un modelo sólo con covariables socio-económicas y el dado por otro modelo que además incluye efectos aleatorios espacio temporales pueden ser diferentes para una misma covariable. Esta hecho no indica que exista contradicción, sino que se puede explicar la variabilidad de distintas formas, pero la correlación de las variables sigue estando presente. Este hecho debe hacernos muy cautos al proponer medidas de política económica, ya que se trata del estudio de una caso particular en el y tiempo y por otro la causa real no puede obtenerse con este tipo de datos. Estudios de causalidad requerirían datos experimentales

y debido a la naturaleza del fenómeno difícilmente podremos controlar las variables que tendríamos que fijar para llevar a cabo el experimento, por no mencionar las consideraciones éticas del mismo. Por otra parte la introducción del modelo ZIP no parece mejorar mucho el ajuste de los modelos. En particular sólo se alcanzan ajustes similares en el caso de agresiones, donde la cantidad de ceros es cercana a la mitad de los casos. Esto nos hace pensar que los modelos inflados en cero solo son necesarios cuando la proporción de ceros es considerablemente mayor a la de casos observados. Esto puede ser relevante si se aplican estos modelos a nivel municipal.

Trabajo futuro

Una de las limitantes constantes que surgieron en el desarrollo de este trabajo fue la limitada ventana de tiempo observada, restringida al sexenio de Felipe Calderón. Será importante poder ver la evolución de estos fenómenos de eventos violentos hacia adelante o bien hacia atrás en el tiempo. Respecto a la capacidad de cálculo de INLA para matrices muy grandes de vecindad y disponibilidad de las covariables, otra aproximación que quedó sin desarrollar pero será interesante ver es la aplicación de los modelo con efectos aleatorios (espacio-temporales) y fijos (covariables socio-económicas) a nivel municipal. Aquí la limitación es que al considerar todos los municipios la matriz de precisión e del CGML es tan grande que no puede manejarse dentro de INLA. A este nivel de resolución será de especial interés los modelos cero-inflados ya que la cantidad de ceros (en especial en enfrentamientos y agresiones) será sustancialmente mayor que la de no-ceros.

APÉNDICE A

Notas adicionales

A.1. Notación O grande

La notación $\mathcal{O}$, O grande, es una notación matemática usada para describir el comportamiento límite de una función cuando el argumento tiende a un valor particular o al infinito. Es comúnmente usada para clasificar algoritmos ya que describe el número de operaciones que requiere y por tanto su tiempo de ejecución o espacio en memoria requerida.

Definición Sea f una función que toma valores en $\mathbb{R}$ o $\mathbb{C}$ y g una función en los reales. Decimos que f es $\mathcal{O}(g)$ o bien $f(x) = \mathcal{O}(g(x))$ conforme $x \to \infty$ si existe un número real positivo M y un número real x_0 tales que

$$|f(x)| \leq Mg(x) \text{ para todo } x \geq x_0.$$

Esta notación también puede ser usado para describir el comportamiento de f cerca de un número real a (usualmente $a = 0$) de modo que escribimos $f(x) = \mathcal{O}(g(x))$ conforme $x \to a$ si existen un número real positivo M y un número real positivo δ tales que

$$|f(x)| \leq Mg(x) \text{ para todo } x \text{ tal que } 0 < |x - a| < \delta.$$

Una definición alternativa para funciones $g(X)$ que no toma el valor 0 es $f(x) = \mathcal{O}(g(x))$ conforme $x \to a$ si

$$\limsup_{x \to a} \left| \frac{f(x)}{g(x)} \right| < \infty.$$

También, la definición de otra notación relacionada , o chica, para funciones $g(X)$ que no toma el valor 0 es $f(x) = (g(x))$ conforme $x \to \infty$ si

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0.$$

La notación O grande tiene dos principales aplicaciones

- Resultados asintóticos infinitos: Análisis de algoritmos.
- Resultados asintóticos infinitesimales: Análisis de aproximación de series truncadas, en especial series de Taylor.

En el caso del análisis infinito y en particular para análisis de algoritmos en lugar de $T(n) = \mathcal{O}(n^p)$, una notación más precisa sería $T(n) \in \mathcal{O}(n^p)$ pues $O(n^p)$ denota una clase de equivalencia de algoritmos.

Algunas propiedades útiles que justifican su uso para simplificar las cuentas son

- Producto

$$f_1 = \mathcal{O}(g_1) \text{ y } f_2 = \mathcal{O}(g_2) \Rightarrow f_1 f_2 = \mathcal{O}(g_1 g_2)$$
$$f \cdot \mathcal{O}(g) = \mathcal{O}(fg) f \cdot \mathcal{O}(g) = \mathcal{O}(fg)$$

- Suma

$$f_1 = \mathcal{O}(g_1) \text{ y } f_2 = \mathcal{O}(g_2) \Rightarrow f_1 + f_2 = \mathcal{O}(\text{máx}(g_1, g_2))$$

  Esto implica que si $f_1 = \mathcal{O}(g)$ y $f_2 = \mathcal{O}(g) \Rightarrow f_1 + f_2 \in \mathcal{O}(g)$ por lo que $\mathcal{O}(g)$ es un cono convexo.

- Multiplicación por una constante

  Sea k una constante, entonces

$$\mathcal{O}(|k|g) = \mathcal{O}(g) \text{ si } k \text{ es distinto de } 0.$$

$$f = \mathcal{O}(g) \Rightarrow kf = \mathcal{O}(g).$$

Podemos decir que en el contexto infinitesimal, $x \to 0$ potencias grandes de $\mathcal{O}(x^n)$ implican mejores aproximaciones (el error es menor), mientras que en el contexto infinito potencias pequeñas de $\mathcal{O}(x^n)$ implican mejores algoritmos pues el número de operaciones es asintóticamente menor.

APÉNDICE B

Lista de funciones en R-INLA

B.1. Manejo de marginales

Función	Descripción
inla.emarginal()	Calcular la esperanza de una función.
inla.dmarginal()	Calcular la densidad.
inla.pmarginal()	Calcular una probabilidad.
inla.qmarginal()	Calcular un cuantil.
inla.rmarginal()	Simular de la marginal.
inla.hpdmarginal()	Calcular un intervalo de probabilidad de alta densidad (HPD).
inla.smarginal()	Interpolar la marginal posterior.
inla.mmarginal()	Calcular la moda.
inla.tmarginal()	Tranformar la marginal.
inla.zmarginal()	Calcular cantidades de resumen.
inla.models()	

B.2. Elementos de un modelo

Valor	Descripción
latent	Modelos latentes disponibles.
group	Modelos disponibles al agrupar observaciones.
link	Funciones liga disponibles.
hazard	Modelos para base de función de riesgo en modelos de superviviencia.
likelihood	Lista de verosmilitudes disponibles.
prior	Lista de previas disponibles.

B.3. Libreria de efectos aleatorios

Efecto	Descripción
exchangeable	Efecto intercambiable.
exchangeablepos	Efecto intercambiable.
ar1	Modelo autoregresivo de orden 1.
ar	Modelo autoregresivo de orden p.
rw1	Caminata aleatoria de orden 1.
rw2	Caminata aleatoria de orden 2.
besag	Modelo espacial de Besag.
iid	Efectos aleatorios independientes e identicamente distribuidos.

Referencias

Akaike, H. (1973). Information theory and the maximum likelihood principle in 2nd international symposium on information theory (bn petrov and f. cs ä ki, eds.). *Akademiai Ki à do, Budapest*.

Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International journal of environmental research and public health*, *12*(9), 10536–10548.

Atuesta, L. H., Siordia, O. S., & Lajous, A. M. (2018). The "war on drugs" in mexico:(official) database of events between december 2006 and november 2011. *Journal of Conflict Resolution*, 0022002718817093.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with r-inla. *Spatial and spatio-temporal epidemiology*, *4*, 33–49.

Box, G. E. & Draper, N. R. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Claeskens, G. & Hjort, N. L. (2008). Model selection and model averaging. Technical report, Cambridge University Press.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, *4*(5), 613–617.

DeGroot, M. H. (2005). *Optimal statistical decisions*, volume 82. John Wiley & Sons.

Dow, M. (2002). Explicit inverses of toeplitz and associated matrices. *ANZIAM Journal*, *44*, 185–215.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, *24*(6), 997–1016.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Gutiérrez-Peña, E. (1997). Métodos computacionales en la inferencia bayesiana. *Monografía IIMAS*.

Gutiérrez-Peña, E. (1998). Análisis bayesiano de modelos jerárquicos lineales. *Monográfías*, *7*(16).

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.

Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: a comparison of mcmc and inla. In *Statistical modelling and regression structures* (pp. 91–110). Springer.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, *50*(2), 120–126.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Konishi, S. & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Lindgren, F., Rue, H., et al. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, *63*(19), 1–25.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

Nakayama, M. K. (2011). Asymptotic properties of kernel density estimators when applying importance sampling. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, (pp. 556–568). IEEE.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065–1076.

Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(1), 175–184.

PNUD (2013). Informe regional de desarrollo humano 2013-2014: Seguridad ciudadana con rostro humano: diagnóstico y propuestas para américa latina.

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, *25*(4), 1145–1165.

Robert, C. & Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Rue, H. & Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of statistical planning and inference*, *137*(10), 3177–3192.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, *71*(2), 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, *4*, 395–421.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Snow, G. (2016). *TeachingDemos: Demonstrations for Teaching and Learning*. R package

version 2.10.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

Takahashi, K. (1973). Formation of sparse bus impedance matrix and its application to short circuit study. In *Proc. PICA Conference, June, 1973*.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(Dec), 3571–3594.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.