Universidad Nacional del Sur

Tesis de Doctor en Matemática

Estadística de Procesos Estocásticos aplicados a Redes Sociales de alta volatilidad

Bavio José Manuel

Bahía Blanca Argentina

Agosto, 2013

para Carolina

Índice general

Prefacio

Esta tesis es presentada como parte de los requisitos para optar al grado académico de Doctor en Matemática de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otras. La misma contiene resultados obtenidos en investigaciones llevadas a cabo en el Departamento de Matemática de la Universidad Nacional de Sur durante el período comprendido entre los meses de mayo de 2008 y junio de 2013, bajo la dirección de Gonzalo Perera, Profesor Titular del La Universidad de la República, Uruguay y la supervisión de la Dra. Ana Tablar. Este trabajo ha sido financiado con una beca otorgada por ANPCyT y UNS en el marco del Proyecto de Formación de Doctores en Áreas Tecnológicas Prioritarias (PFDT) del Programa de Recursos Humanos (PRH 2007 - código 37).

Agradecimientos

A lo largo de la elaboración de esta tesis he recibido el apoyo de muchas personas y quisiera con estas palabras expresar mi agradecimiento:

- *A mis padres Marta y Rodolfo*: por apoyarme a lo largo de la vida y enseñarme el valor del esfuerzo y el conocimiento.
- *A mi esposa Carolina*: por ser mi compañera en cada paso de mi vida y de esta tesis y por permitirme ir juntos a la par.
- *A mi Director Gonzalo Perera*: por su generosidad y su capacidad para proponer problemas afines a mis inquietudes científicas, por su calidad humana y hospitalidad.
- *A Ana Tablar, Beatriz Marrón y Melina Guardiola*: por su acompañamiento a cada paso de este doctorado no solo en el aspecto académico sino dando afectuosos consejos y aportando certezas sobre el mejor paso a dar.
- *A Miguel, Agustín y Fernando*: por el aliento, el buen humor y los momentos compartidos en el laboratorio.
- *Al departamento de Matemática*: por el apoyo brindado a lo largo de estos años sin el cual esta tesis hubiera sido mucho más difícil de alcanzar.

Finalmente a todos los que de una forma u otra han aportado para que cada una de estas páginas vea la luz.

Muchas gracias!!!
José Bavio

3 de Julio de 2013

Departamento de Matemática.

Universidad Nacional del Sur.

Resumen

Las redes sociales virtuales como Facebook y Twitter están muy difundidas en nuestras vidas cotidianas y generan un montón de datos de intercambios. Planteamos un modelo estocástico para Twitter que nos permite estudiar la dinámica de la red y el comportamiento de los usuarios sobre su saturación. Para estudiar este modelo estocástico se utiliza la herramienta estadística de cópulas que analiza la dependencia de variables aleatorias.

Este trabajo de tesis proponemos una generalización del estimador por núcleos de cópulas para serie de tiempos presentado por Fermanian y Scaillet en 2002. Dicha generalización se extiende a procesos estocásticos de difusión.

A partir de éste estimador, se puede analizar la probabilidad de saturación de un usuario de Twitter y otras medidas vinculadas con esta saturación.

Abstract

Virtual social networks like Facebook and Twitter are very spread in daily life. Using it generates an incredible amount of exchange information.

In this work we propose an stochastic model for Twitter that allows the study of network dynamics and users behavior specially concern with saturation.

To study this model we use a statistical tool named as copulas that analices dependence between random variables.

In this thesis we propose a generalization of the non-parametric copula estimator presented by Fermanian and Scaillet in 2002. This generalization reaches continuos process as diffusion.

From this estimator we can analize profile saturation probability and other measures related with saturation.

Introducción

El desarrollo de las nuevas tecnologías de comunicación tiene un impacto en nuestra vida cotidiana. No hace mucho tiempo la única forma de comunicarse era a través del correo escrito y en la actualidad volvemos a nuestros hogares si nos olvidamos el celular por si alguien nos quiere llamar, y organizamos cumpleaños por Facebook o convocamos reuniones por Twitter.

Una consecuencia de esta penetración de la tecnología puede ser esa necesidad de estar disponible a cualquier hora y en cualquier lugar, donde la persona presente queda muchas veces en segundo lugar respecto de la persona que llama o envía un mensaje.

Otra consecuencia se podría describir como una duplicación de nuestra realidad donde por un lado las situaciones nos suceden en el mundo "real" y al mismo tiempo, las compartimos y publicamos en nuestras redes sociales virtuales.

Estos nuevos espacios sociales que permiten las nuevas tecnologías, despiertan un gran interés y el enfoque multidisciplinario es el mejor abordaje de este fenómeno complejo. Desde el punto de vista social, es interesante ver como se generan y evolucionan estas redes sociales y desarrollar nuevas herramientas y modelos para estudiarlas y analizarlas. Desde el punto de vista de una red compleja puede resultar interesante conocer distintas dinámicas entre sus elementos por ejemplo, la existencia de subredes.

Esta duplicidad de la realidad, por un lado la "virtual" y por el otro la "real" también plantea interrogantes de cómo se relacionan entre sí. Los vínculos empiezan en una y continúan en otra, o terminan en una dimensión y pueden no terminar en otra.

Respecto de esta duplicidad, existen ejemplos en los que las redes sociales virtuales han traspasado lo virtual generando movimientos sociales reales (con manifestaciones y grandes movilizaciones). Un ejemplo de este paso de lo virtual a lo "real" pueden ser las revoluciones de la denominada primavera árabe.

De cualquier manera las redes sociales como nunca se han transformado en un fenómeno real, que de la mano de la penetración de la tecnología, ha venido para quedarse. Conscientes o no, formamos parte él y lo construímos día a día. Por eso creemos que es muy interesante estudiar este fenómeno o algún aspecto de él y la matemática aporta herramientas de suma utilidad en este sentido.

En esta tesis doctoral abordaremos un caso particular de todo el universo de redes so-

ciales existentes. Centraremos nuestro enfoque en la red social Twitter y plantearemos un modelo simple de su funcionamiento que nos permita estudiar algunos comportamientos de los usuarios.

El modelo elegido nos demandará el desarrollo de herramientas y resultados nuevos para responder algunas de las preguntas planteadas en base al mismo y estas herramientas resultarán de gran interés no sólo para este trabajo sino también en otras ciencias.

Abordaremos esta tarea sin pretender describir la realidad en su totalidad y coincidiendo con la siguiente idea de Box y Draper [4].

"Essentially, all models are wrong, but some are useful"

Creemos que cualquiera que quiera considerar la tarea de modelado debería tener esta idea en la mente [44].

El contenido de esta tesis se encuentra organizado de la siguiente forma:

- En el capítulo 1 realizamos una exposición de las redes sociales en general, con u su historia dentro de las ciencias sociales y la evolución de su concepto. Describimos algunas características que permiten clasificarlas y otros aspectos que revisten relevancia sobre las mismas.

- En el capítulo 2 focalizamos el estudio sobre la red social Twitter. Brindamos algunas informaciones básicas sobre su funcionamiento y objetivos comunicacionales así como también sobre algunas discusiones al respecto que están haciendo.

- En el capítulo 3 hacemos una introducción a la teoría de procesos estocásticos y presentamos clasificaciones y algunas propiedades de los mismos.

- En el capítulo 4 planteamos el modelo de Twitter que proponemos para su estudio.

  A partir de este modelo planteamos algunos resultados sobre un evento relevante dentro de la vida del perfil que denominamos saturación y finalmente presentamos algunas simulaciones de estos resultados.

- En el capítulo 5 presentamos la técnica estadística de cópulas. Hacemos una introducción al tema y se describen algunos resultados que serán de utilidad para el siguiente capítulo.

- En el capítulo 6 presentamos una forma de calcular estimaciones de cópulas para cierta clase de procesos estocásticos y luego generalizamos dicho método a procesos continuos.

- En el capítulo 7 retomamos el método del capítulo anterior y lo utilizamos para sacar más conclusiones sobre el modelo de Twitter ya presentado.

- Finalmente presentamos las conclusiones de este trabajo y planteamos algunas líneas de investigación a partir de los desarrollos de esta tesis.

Con este trabajo hacemos nuestro aporte al análisis de redes sociales así como también a la estadística de procesos estocásticos. Deseamos que resulte interesante para la comunidad académica y sirva como punto de partida para futuras investigaciones, tanto de la estadística de procesos como en sus aplicaciones.

Capítulo 1

Análisis de Redes Sociales

1.1. Introducción

En este capítulo daremos una breve descripción del concepto Redes Sociales, el origen de estas y algunas clasificaciones que serán pertinentes y motivadoras de los modelos a desarrollar en el resto de este trabajo.

El estudio de las redes sociales se puede enmarcar dentro de la sociología. Es una disciplina que intenta encontrar información y explicaciones de distintos objetos sociales según distintas orientaciones disciplinarias o científicas. Gran parte de este capítulo está basado en el libro de Furht "Handbook of Social Network Technologies" [18].

1.2. Breve historia Redes Sociales

Los primeros sociólogos a finales del siglo XIX, incluyendo a Durkheim y Tönnies, son precursores de la teoría de redes sociales. Tönnies argumentaba que los grupos sociales pueden existir tanto como lazos personales y dirigidos que unen individuos que comparten valores y creencias o pueden ser vínculos sociales, instrumentales, formales e impersonales.

Durkheim dió una explicación no individualista de los hechos sociales argumentando que los fenómenos sociales surgen cuando la interacción de individuos crea una realidad que no se puede explicar en términos de las propiedades individuales de los actores [17].

George Simmel, a principios del siglo XX, en [36] fue el primer investigador en pensar directamente en términos de redes sociales. Sus ensayos apuntan a la naturaleza del tamaño de la red sobre la interacción y a la probabilidad de interacción en redes ramificadas [36].

Las redes sociales como un concepto relativamente separado, se generaron entre 1920 y 1930 en el campo de la antropología en Gran Bretaña. El antropólogo Roger Brown fue el primer investigador que usó el término red social, implicando que la estructura es

similar a una red y que la comunicación interpersonal entre los individuos se parece a la relación entre un nodo y otro conectados en la red.[35]

Otra línea de Análisis de Redes Sociales (SNA) se puede remontar al Método Sociométrico creado por el psicólogo social Jacob Levy Moreno en la década del 30. Este método allanó el camino para el análisis cuantitativo en redes sociales. También en 1930, Moreno fue el primero en grabar y analizar sistemáticamente la interacción social en grupos pequeños, especialmente aulas y grupos de trabajo.

De acuerdo con Freeman [17], en el libro de Moreno de 1934, se utilizó el término "red" en el sentido que se usa hoy en día. Además para 1938, el trabajo de Moreno - con la ayuda de Jennings y Lazarsfeld - había establecido las cuatro características prinicipales que definen al análisis de redes sociales (SNA). Estas son:

- la intuición que la estructura social está basada en las conecciones entre los actores.
- recolectar datos empíricos sistemáticamente
- utilizar representaciones gráficas de las redes
- basado en modelos matemáticos o computacionales

Otro aporte de trabajo surge de un grupo de Harvard liderado por W. Lloyd Warner y Elton Mayo. Se enfocaron en el estudio de la estructura social a fines de la década del XX. Pero este esfuerzo según Freeman no despegó porque no planteó un modelo general para estudiar el paradigma estructural. De hecho estos esfuerzos del grupo de Harvard casi nunca son reconocidos en las revisiones históricas de SNA [17].

El período de 1940 a 1960 es llamado por Freeman como la Era Oscura en la historia del desarrollo de SNA. En este período no hay ninguna aproximación a la investigación social que incorpore el paradigma estructural. SNA todavía no era identificable ya sea como una perspectiva teórica o como un abordaje a la colección y análisis de datos.

En el período 1960-1970 surge un grupo centrado en Harrison White y sus alumnos en la universidad de Hardvard: Ivan Chase, Bonnie Erickson, Harriet Friedmann, Mark Granovetter, Nancy Howell, Joel Levine, Nicholas Mullins, John Padgett, Michael Schwartz y Barry Wellman. Freeman llama a este proceso como el renacimiento de SNA en Hardvard. La escuela de Harvard publicó teorías tan importantes en esta línea que los científicos sociales sin importar su campo dejaron de desconocer la idea de SNA. Para fines de 1970 SNA se había vuelto universalmente reconocido entre los científicos sociales.

Existen otros grupos de trabajo que se podrían considerar que han aportado al desarrollo de el análisis de redes sociales en el período 1930-1970 [17].

Una vez que el concepto de Redes Sociales fue reconocido por más investigadores, mayores fueron las investigaciones en la metodología: más métricas sobre redes, mayor recolección de datos y más tecnologías de análisis fueron desarrolladas para entender

construcciones sociales y relaciones entre estas. Todo esto ayudó a desarrollar aun más SNA.

En 1980, los sociólogos empezaron a usar SNA como una herramienta analítica para examinar fenómenos sociales y económicos. A mitad de los 80, Mark Granovetter, propuso el concepto de "embedded-ness" (arraigo, incrustación, inserción), llevando el enfoque de SNA al centro de la investigación de estructuras sociales. Granovetter argumentaba que la operación de la economía está inmersa en la estructura social, aunque el núcleo de la estructura social es la red social de cada individuo [19].

Después de 1990, SNA ha sido asociado con el capital social, llamando la atención de los investigadores de distintos campos como la sociología, política, economía, ciencias de la comunicación, y otras disciplinas. El libro de Ronald Burt "Structural Holes" es representativo de este período. Burt decía que el capital social no tiene relación con la fuerza de los vínculos, sino con la existencia de agujeros en la estructura. Lin Nan, otro sociólogo estudió el SNA desde la perspectiva del capital social.

1.3. Algunas definiciones de SNA

Las redes sociales se pueden definir formalmente como un conjunto de actores sociales, que se modelan por nodos. Estos miembros están conectados entre sí por uno o más tipos de relaciones. Esta visión de red, permite realizar distintas clasificaciones entre los nodos, según diversas métricas.

1.3.1. Clasificación de redes

Existen diferentes tipos de redes. Generalmente se distinguen los siguientes tipos:

- *Redes unimodales o polimodales.* El modo se refiere a la clase que represetan los nodos. Si todos los nodos representan individuos la red es unimodal, en cambio si un conjunto de nodos representa individuos, otro asociaciones con fines de lucro, otro representa instituciones del estado podemos decir que la red es polimodal [20].

- *Redes completas o redes egocéntricas.* Las redes pueden abarcar las relaciones de toda una comunidad o bien pueden estudiar sólo las conexiones de un único individuo y centrarse en él.

El análisis de redes sociales estudia estructuras, está basado en la intuición de que los lazos y vínculos sociales a los que el sujeto está "sujetado" tienen consecuencias importantes para el mismo. Este es el razgo característico del SNA.

El estudio empírico de la redes permite determinar los distintos papeles que pueden cumplir los nodos dentro de ella, uno puede decir quiénes son los que conectan, los que

son líderes, los que son puentes, los que están aislados, etc. También se pueden detectar segmentaciones y otras características que puedan resultar relevantes. Naturalmente la descripción matemática de una red es mediante los grafos. Muchas veces se habla de grafo indistintamente que de red social.

A continuación vamos a explicar conceptos importantes de vínculos, centralidad y densidad.

1.3.2. Lazos o vínculos

Si uno tiene sólo un conjunto de nodos, no podemos hablar de una red. Sin considerar sus vínculos son sólo puntos aislados. Los lazos o vínculos conectan dos o más nodos de la red. Muchos comportamientos humanos, como compartir información o prestar dinero, o ser jurado de una tesis doctoral, generan vínculos en los que se establecen jerarquías (acreedor-deudor, aprendiz-maestro, etc.). Estos vínculos se pueden distinguir de otros, como la amistad o la pertenencia a un club, etc. en los que no surge a partir del vínculo un orden evidente. Los vínculos que establecen orden, se llaman dirigidos mientras que los que no establecen orden se llaman no dirigidos. Cabe destacar que un vínculo dirigido entre dos nodos, puede ser recíproco, pero esto no es necesariamente lo mismo que un vínculo no dirigido entre nodos. Según la naturaleza de los vínculos podemos distinguir entre *redes sociales dirigidas*, o *redes sociales no dirigidas* o *redes mixtas*. Por supuesto que estas redes se modelan con grafos dirigidos o no dirigidos o mixtos respectivamente.

1.3.3. Densidad

Un concepto ampliamente utilizado en SNA es el de "densidad". Este concepto describe el nivel general de conectividad entre los puntos del grafo. Un grafo "completo" es uno en el que todos sus nodos o puntos son adyacentes entre sí. Todo punto está directamente conectado con cualquier otro de la red. El concepto de densidad intenta capturar la medida de cuanto se aleja una red de este estado de completitud [35]. La densidad se define cuantitativamente como el número de aristas dividido el número total de nodos. Es una de las primeras medidas en análisis de redes y una de las nociones más comunes que se utilizan en epidemiología social.

1.3.4. Caminos, longitud y distancia

Los nodos de la red pueden estar conectados directamente o pueden estarlo indirectamente mediante una secuencia de nodos y aristas. Una de estas secuencias se llama camino y se distinguen aquellos caminos en el que todas las aristas y los nodos son distintos. El concepto de camino es también importante. La longitud de un camino (sin vértices ni aristas repetidos) se mide según la cantidad de aristas que lo forman. La distancia entre

dos nodos es la longitud del camino más corto que los conecta. Puede valer 0 (si los dos nodos son en realidad el mismo) o puede valer ∞ si no existe un camino más corto.[35]

1.3.5. Centralidad

Las medidas de centralidad identifican los actores más importantes dentro de una red. Esta importancia viene dada por la cantidad de relaciones que tienen con otros miembros de la red. Las medidas más difundidas para medir esta centralidad son:

- *Centralidad según el grado.* El grado de un nodo es la cantidad de otros nodos que están conectados directamente a él. Se asocia con popularidad.
- *Between-ness.* Otra forma de mirar la centralidad es contar la cantidad de veces que un nodo conecta pares de otros nodos, que de otra forma no se podrían conectar. Mide el potencial de control de ese actor en el flujo dentro de la red.
- *Cercanía.* Esta medida está basada en la noción de distancia. Si un nodo es cercano a todos los demás de la red, entonces no depende de nadie para conectarse con el resto. Esto da una noción de independencia o eficiencia del nodo.

1.4. Redes Sociales Virtuales

Dentro del SNA, ha atraído mucha atención las denominadas Redes Sociales Virtuales. Una definición de estas redes se puede encontrar en el trabajo de Boyd y Ellison [5]. Allí se define a las Redes Sociales Virtuales como servicios basados en internet que permiten a los individuos:

1. Construir un perfil público o semi público dentro de un sistema acotado.
2. Articular una lista de otros usuarios con los que comparten conexiones
3. Ver y atravesar la lista de conexiones y las de sus contactos dentro del sistema.

El origen de las redes sociales virtuales se remonta, al menos, a 1995, cuando Randy Conrads crea el sitio web *classmates.com*. Con esta red social se pretendía que la gente pudiera recuperar el contacto con antiguos compañeros de escuela, liceo, universidad, etc. Sin embargo, Boyd et al[5], teniendo en cuenta su definición, afirman que la primera red reconocible fue lanzada en 1997. Esta fue *SixDegrees.com* que permitía a los usuarios crear perfiles, ordenar a sus contactos y luego en 1998 investigar las listas de amigos de sus contactos.

Cada una de estas posibilidades ya existían en algunos sitios de internet antes de Six-Degrees. Los perfiles ya existían en muchas sitios para formar parejas. Los sitios de chat

AIM y ICQ permitían las listas de amigos, pero no se podían ver entre sí. Classmates.com permitía a las personas afiliarse a su escuela secundaria y recorrer la red de otros que estuvieran afiliados, pero no podían crear perfiles. *SixDegrees* fue el primer sitio en combinar todas estas características.

Cualquier resumen de la historia de un fenómeno tan masivo queda necesariamente incompleto. De cualquier manera podemos asegurar que las redes sociales virtuales han dado forma a un nuevo negocio y han hecho un gran aporte al panorama cultural y de investigación.

1.4.1. Clasificación de Redes Sociales Virtuales. Algunos ejemplos

Hay distintas formas de clasificar las redes sociales: según su tamaño, según su ámbito social específico, según la tecnología que utilizan, etc. No pretendemos dar aquí ninguna clasificación exhaustiva de este fenómeno, por lo que detallaremos algunas de las clasificaciones,a nuestro criterio, más generales, y aquellas que delimitan los temas de estudio que presentamos en este trabajo. Esta clasificación está tomada de la página de Mary White [45].

- Que permiten mantener contacto con amigos y familiares.
  - Facebook.
  - Google +
  - Twitter.
  - MySpace
- Intercambio de contenidos Multimedia
  - YouTube
  - Flickr
  - Picasa
- Profesionales.
  - LinkedIn
  - Classroom 2.0
  - Nurse Conect
  - SQL Monster
- De información

- Super Green Me
- HGTV Discussion Forums.
- Do-It-Yourself.

- Educación
  - The Student Room
  - The Math Forum
  - ePALS School Blog
  - eLearners

- Hobbies
  - Oh My Blom
  - My Place at Scrapbook.com
  - Sport Shouting

- Academicas
  - Academia.edu
  - Connotea Collaborative Research

Otra forma de clasificar a las redes sociales virtuales puede ser a partir de la actividad principal que se desarrolla en ellas. Este criterio permite distinguir [40]:

- **Foros**.

  Los foros son espacios del sitio web que permiten a los visitantes interactuar entre ellos e intercambiar opiniones y comentarios. Los foros están bien identificados y moderados, cuando un comentario no pertenece al tema sobre el que se basa el foro, se elimina o redirecciona. Los foros son muy útiles para encontrar personas con intereses similares.

- **Microblogging**.

  El Microblogging es un nuevo fenómeno, está muy relacionado con el servicio de mensajes de textos cortos (Short Message Service) SMS. Consiste en escribir entradas de un blog, pero a pequeña escala, es decir en no mas de 140 caracteres por publicación. Puede escribirse desde cualquier celular con la tecnología adecuada, esto hace que la generación de textos sea practicamente constante. El ejemplo más difundido de este tipo de red social es Twitter. A fines de junio de 2009, las protestas en Irán fueron cubiertas por Twitter, teniendo a disposición información casi instantánea de lo que ocurría.

- **Bookmarking**.

  Como existe una cantidad muy grande de información dentro de internet, (noticias, datos duros, citas bibliográficas), asimilarlas todas y referenciarlas podría ser una tarea imposible, o que tome mucho tiempo. Para ayudar a esta tarea existen sitios como Digg thrive. Uno puede seleccionar links de páginas y fuentes de internet, que considere útiles para algún tema, y permitir que otros Diggers tengan cierto preprocesamiento de los datos disponibles. Sitios como Digg y StumbleUpon proveen links basados en los artículos y preferencias que un usuario ha leído. Esta red forma una especie de medio de comunicación que permite entre otras cosas promover negocios o temas de discusión.

- **Video Sites**.

  No todo lo que se comparte en internet es para leer, también se pueden compartir videos o contenidos de audio. Redes sociales como Youtube permiten compartir videos. Se pueden mirar videos, luego comentarlos y recomendarlos a amigos. Casi cualquier individuo puede "subir" un video. Algunos tipos de video que se pueden ver son: educativos, películas independientes, sketchs de comedia, capítulos de series, webisodes (capítulos pensados para difundir por la web).

- **Motores de Búsqueda (Search Engines)**.

  Este es un tipo de sitio web que promueve una interacción social más "escondida". Antes uno escribía lo que buscaba y aparecían los resultados. Ahora uno puede personalizar esos resultados según su región, según las preferencias de sus contactos, guardar las búsquedas y otras interacciones que pueden o no estar basadas en preferencias de sus contactos o de búsquedas previas. Este almacenamiento de las preferencias del usuario y de sus contactos se puede pensar una interacción social sujeta de análisis.

Otra clasificación importante se basa en la naturaleza de los lazos de la red. Con este criterio pueden ser:

- Dirigidas. Los lazos se establecen según la voluntad de una sóla de las partes. El ejemplo más difundido es *Twitter*, en el que "seguir" a alguien sólo depende de la voluntad del seguidor.

- No dirigidas. Los lazos se establecen según la voluntad de los dos actores, por ejemplo en *facebook* uno envía una "solicitud de amistad" y ésta debe ser aceptada por el otro perfil para que se de la conexión.

Como decíamos antes, existen muchas maneras de caracterizar y clasificar las redes sociales online, muchas categorías se solapan, también puede ocurrir que una misma red

social virtual pertenezca a varias categorías, o que vaya evolucionando y adquiera nuevas características que no encajen en la clasificación preexistente.

1.5. Evolución de una red social

Como todo fenómeno social el de las redes sociales virtuales muestra dinámicas globales, aplicables (en términos generales) a cada una de las redes, y otros comportamientos que son específicas de cada una.

Las redes sociales en internet están creciendo muy rápido en número y tamaño. Esta popularidad masiva hace que llamen la atención de los investigadores. Para lograr conocimiento nuevo a partir de estos objetos sociales, es necesario entender cómo las redes crecen y cambian.

La mayoría de los estudios de las redes sociales de internet involucran la dimensión temporal, ya que una característica de estas redes es el dinamismo.

1.5.1. La "S" tecnológica

Aunque de momento no hemos observado ciclos completos de vida de una red social de éxito, los datos que hasta ahora tenemos nos permiten establecer un crecimiento en forma de curva logística (o en forma de "S") en 3 etapas: inicio, explosión y cima.

En la etapa de inicio los usuarios comienzan a invitar a potenciales usuarios, con una tasa de éxito importante que conlleva un crecimiento exponencial. Aunque es importante contar con una buena fuente de usuarios. En este punto, la tasa de abandono de los usuarios suele ser baja.

En la figura (1.1), vemos un ejemplo gráfico de este crecimiento.

La etapa de explosión viene marcada por una bajada de la tasa de éxito, escondida por el gran número de usuarios ya disponibles. El crecimiento en este período deja de ser exponencial para convertirse en lineal. Su duración depende del nivel de saturación de la red y de la tasa de abandono de los usuarios.

No sabemos aún si se ha llegado a la cima en cuanto al número de usuarios registrados, pero puede ser que sí estemos asistiendo a esta fase en lo que a número de usuarios activos (usuarios que acceden al menos una vez al mes) se refiere [7].

Estas ideas coinciden con principios de marketing y economía, el ciclo vital de un producto tiene cuatro fases:

- introducción/lanzamiento
- crecimiento

Figura 1.1: Evolución de la penetración de una tecnología o una red social

- madurez
- declive

No solo las redes sociales siguen este comportamiento, también lo hacen otros servicios tecnológicos como la telefonía celular, la televisión por cable.

La etapa de declive, se puede dar por infinidad de causas, saturación, competencia de otra red social, razones sociológicas, etc.

Caracterizamos éstas etapas de desarrollo de una red social porque el comportamiento de los usuarios dentro de ella varía en las distintas etapas. En la etapa de lanzamiento, en general solo una pequeña elite tiene conocimiento y acceso a la red social, estos usuarios tienen un uso que puede ser intenso del servicio, pero acotado a su núcleo. Muchos consideran el final de esta etapa como el momento clave que determina la vida de la red social, si logra surgir de este pequeño grupo y aumentar la cantidad de usuarios, entonces entrará en la etapa de crecimiento y se difundirá el uso.

1.6. Comportamientos de usuarios

También podemos mencionar que el comportamiento del usuario tiene una evolución, respecto de sus niveles de actividad. En el trabajo [1] la descripción de las actividades de

los usuarios cuando utilizan un servicio de red social es la siguiente:

De 41 actividades distintas identificadas (que requerían una interacción HTTP), las agruparon de la siguiente forma:

- Búsqueda
- Scrapbook (recortes, collage,etc)
- Mensajes
- Testimonios
- Videos
- Photos
- Perfil & amigos
- Comunidad
- Otras

La siguiente tabla muestra la lista y los porcentajes de actividades realizadas en la red social Orkut.

Category ID	Description of activity	# Users (%)	# Requests (%)	Bytes (MB)
Profile & Friends	17 Browse profiles	19,984 (17.9)	149,402 (19.0)	3,534
	18 Browse homepage	18,868 (16.9)	92,699 (11.8)	3,866
	19 Browse the list of friends	6,364 (5.7)	50,537 (6.4)	1,032
	20 Manage friend invitations	1,656 (1.5)	8,517 (1.1)	144
	21 Browse friend updates	1,601 (1.4)	6,644 (0.8)	200
	22 Browse member communities	1,455 (1.3)	6,963 (0.9)	133
	23 Profile editing	1,293 (1.2)	7,054 (0.9)	369
	24 Browse fans	361 (0.3)	1,103 (0.1)	17
	25 Browse user lists	126 (0.1)	626 (0.1)	9
	26 Manage user events	44 (<0.1)	129 (<0.1)	2
Communities	27 Browse a community	2,109 (1.9)	8,850 (1.1)	164
	28 Browse a topic in a community	926 (0.8)	9,454 (1.2)	143
	29 Join or leave communities	523 (0.5)	3,043 (0.4)	43
	30 Browse members in communities	415 (0.4)	3,639 (0.5)	56
	31 Browse the list community topics	412 (0.4)	2,066 (0.3)	38
	32 Post in a community topic	227 (0.2)	1,680 (0.2)	24
	33 Community management	105 (0.1)	682 (0.1)	12
	34 Accessing polls in communities	99 (0.1)	360 (<0.1)	6
	35 Browse the list of communities	47 (<0.1)	337 (<0.1)	8
	36 Manage community invitations	20 (<0.1)	63 (<0.1)	1
	37 Community events	19 (<0.1)	41 (<0.1)	1
Other	38 Accessing applications	1,092 (1.0)	4,043 (0.5)	61
	39 User settings	403 (0.4)	2,020 (0.3)	32
	40 Spam folder, feeds, captcha	48 (<0.1)	150 (<0.1)	2
	41 Account login and deletion	39 (0.1)	76 (<0.1)	1

Category ID	Description of activity	# Users (%)	# Requests (%)	Bytes (MB)
Search	1 Universal search	2,383 (2.1)	15,409 (2.0)	287
Scrapbook	2 Browse scraps	17,753 (15.9)	147,249 (18.7)	2,740
	3 Write scraps	2,307 (2.1)	7,623 (1.0)	113
Messages	4 Browse messages	931 (0.8)	3,905 (0.5)	64
	5 Write messages	70 (0.1)	289 (<0.1)	5
Testimonials	6 Browse testimonials received	1,085 (1.0)	3,402 (0.4)	57
	7 Write testimonials	911 (0.8)	4,128 (0.5)	65
	8 Browse testimonials written	540 (0.5)	1,633 (0.2)	26
Videos	9 Browse the list of favorite videos	494 (0.4)	2,262 (0.3)	44
	10 Browse a favorite video	390 (0.3)	862 (0.1)	13
Photos	11 Browse a list of albums	8,769 (7.8)	43,743 (5.6)	871
	12 Browse photo albums	8,201 (7.3)	70,329 (8.9)	2,313
	13 Browse photos	8,176 (7.3)	122,152 (15.5)	1,147
	14 Browse photos the user was tagged	1,217 (1.1)	3,004 (0.4)	47
	15 Browse photo comments	355 (0.3)	842 (0.1)	16
	16 Edit and organize photos	82 (0.1)	266 (0.0)	3

Como podemos ver, la cantidad de actividades que puede realizar un usuario es muy variada y hay que tener en cuenta que en esta tabla está estudiada una sóla red social.

El relevamiento de todas las actividades que puede realizar un usuario es una tarea ardua y escapa al objetivo de esta tesis. No queremos dejar de resaltar la complejidad de

fenómeno de las redes sociales y la dificultades que presenta el abordaje sistemático de las mismas.

1.7. Conclusiones

El concepto red social, como vimos al comienzo de este capítulo, tiene un largo historial y una diversidad y amplitud de enfoques muy grande. Con el desarrollo de las telecomunicaciones, han surgido nuevas plataformas virtuales que han modificado la forma en la que muchas personas se comunican y generan vínculos sociales. Sitios como Facebbok y Twitter, conocidos como sitios de redes sociales, han experimentado una penetración muy grande en nuestras sociedades logrando algo muy significativo. Han apropiado para sí la denominación de red social, de manera que si uno realiza una encuesta sobre qué es una red social, entre las primeras respuestas están estas dos palabras: Facebook y Twitter.

En general, a lo largo del trabajo cuando hablemos de red social, nos referiremos a Twitter o a Facebook porque si bien, y lo intentamos reflejar en el comienzo de este capítulo, tomó mucho tiempo pensar lo sociológico en términos estructurales y desarrollar la idea de las redes sociales, no queremos desconocer que en la actualidad la principal asociación con el concepto "red social" es Facebook o Twitter.

En el próximo capítulo vamos a profundizar más sobre la red social Twitter dado que esta tesis aporta un modelo matemático aplicado a la misma. Considerando que estamos situados en el estadío de madurez de la red donde el comportamiento de los usuarios ha decantado la "novedad" y se encuentra en un régimen estacionario.

Capítulo 2

Twitter I: Introducción

2.1. Introducción

En este capítulo, presentaremos en más detalle la red social Twitter. Describiremos algunos aspectos sobresalientes de su estructura y funcionamiento que hacen relevante su estudio, y que también nos servirán para justificar el modelo propuesto más adelante.

También plantearemos algunas polémicas respecto de su objetivos comunicacionales y haremos hincapié en la volatilidad de la generación y distribución de los contenidos que circulan por ella.

2.2. ¿Qué es Twitter?

Twitter es muchas cosas, una empresa de software, una marca, una red social. Los aspectos que nos interesa estudiar de Twitter refieren a su calidad de servicio de red social virtual. Es decir como "social network website". Dentro de las caracterizaciones y clasificaciones que dimos en el capítulo anterior, se lo puede ubicar como una red unilateral, orientada a Microblogging.

Esto es básicamente una red en la que el vínculo no requiere aprobación, el usuario sigue al perfil que quiere, cuando quiere, y lo que hace principalmente es publicar mensajes de texto con una longitud de hasta 140 caracteres. Si bien existen otras opciones de publicación en Twitter, como la tweetcam, en la que el usuario puede publicar un video, la característica por excelencia de Twitter son los textos cortos.

Para poder enviar mensajes a través de Twitter, una persona debe registrarse y el perfil queda asociado a una cuenta de correo electrónico válida. En este registro se requiere un nombre de usuario y una contraseña con el que queda identificado el perfil y restringido el acceso al mismo a aquellos que poseen la contraseña.

Una vez registrado, el usuario puede empezar a emitir mensajes que se denominan

"tweets", buscar y elegir perfiles de otros usuarios para empezar a seguir.

El conjunto de perfiles que sigue un perfil lo denominamos conjunto de líderes. La cantidad de líderes está restringida a 2000 por perfil, aunque estos límites pueden variar de un perfil a otro. Estos límites se deben a cuestiones técnicas y para evitar spam (correos no deseados) entre otras causas [42]. Otro conjunto que también se puede distinguir es el de seguidores o followers, que está formado por todos los seguidores de un perfil.

Es importante destacar que un usuario sólo puede tomar decisiones sobre su conjunto de líderes, ya que no es posible elegir ni restringir quienes serán nuestros seguidores.

"Tweets"

Los tweets son públicos por defecto, sin embargo se puede restringir la visibilidad sólo a los seguidores del emisor o a un único perfil. De todas formas, la característica principal y su mayor uso es para emitir tweets públicos.

Los tweets son textos de hasta 140 caracteres y pueden contener palabras, links, y algunas referencias a otros perfiles o a otros tweets mediante signos especiales como # o @. Permitiendo agrupar los tweets de distintas maneras, bajo un mismo tema, o relacionado a uno o más usuarios.

Los usuarios pueden emitir sus tweets desde el sitio de Twitter con cualquier computadora, o desde teléfonos celulares con la tecnología adecuada (smartphones con acceso a internet) y en algunos países lo pueden hacer hasta por SMS (Short Message Service). De manera que la cantidad de tweets que se emiten es muy grande y permanente[46]. Este caudal de información constante es una de las principales características de esta red. Y es el aspecto que nos inclina a elegir el modelo que vamos a presentar más adelante.

Twitter en números

Algunas cifras que llaman la atención de este fenómeno de comunicación se pueden ver en la tabla (2.1):

2.2.1. La comunicación "por" Twitter

Definitivamente las redes sociales han creado un espacio virtual donde la realidad se duplica, se recrea y/o se amplía. En esta realidad virtual la comunicación también adquiere características propias.

2.2.2. Twitter: ¿Es una red social o plataforma de contenidos?

Podemos definir un servicio de red social como una tecnología que facilita la conexión de lazos sociales ya existentes. Uno comparte fotos, mensajes, videos, etc, con personas que

Número total de usuarios registrados	500,250,000
Número de usuarios diarios nuevos	150,000
Número de visitantes distintos por mes	180 millones
Promedio de tweets diario	55 millones
Número de búsquedas diarias en Twitter	1.600 millones
% de usuarios de Twitter que usan el celular para tweetear	41 %
Número de empleados de Twitter	175
Número de usuarios activos por mes	100 millones
% Twitteros que no tweetean pero miran otros perfiles	40 %
Número de días para contar 1.000.000 de tweets	5 días
Número de tweets por segundo	8.900

Cuadro 2.1: Estadísticas sobre Twitter [6]

ya conoce. Por ejemplo "Facebook" está diseñado para reforzar conexiones con la gente que ya se conoce bien en la vida real. En este punto, Twitter posee algunas características de una red social. Twitter enfatiza conexiones débiles con gente que uno conoce poco de la vida real o directamente que sólo conoce por su actividad online o en otros medios de comunicación. Permite seguir sus actividades o entablar conversaciones.

La respuesta de si Twitter es una red social es que sí, pero a diferencia de otras redes, puede ser considerada desde otros puntos de vista.

Una plataforma de distribución de conocimiento es un sistema que provee a los usuarios de información relevante y links, al mismo tiempo brinda a los que publican contenidos, una forma de llevar el mismo a los que puedan estar interesados.

Esta finalidad de Twitter se puede ver en acción si uno tiene en cuenta a compañías como CNN que tweetean links sobre sus noticias a mas de 1 millón de usuarios.

Tweetear es una herramienta efectiva para difundir contenidos no solo para las cadenas de noticias de todo el mundo, sino también para organizaciones no gubernamentales, organizaciones gubernamentales, y aún también para individuos particulares.

Algunos datos tomados muestran que las dos afirmaciones anteriores son ciertas. De hecho, permiten clasificar en dos grandes grupos la actividad en la "Tweettesfera". Por un lado, muchas cuentas de Twitter no tienen seguidores o nunca han tweeteado nada. Esto indica que estos usuarios solo siguen a sus artistas y celebridades favoritas o esperan fuentes de información para mantenerse actualizados en algún tema.

Por otro lado, más del 30 % de todos los tweets son respuestas. Esto significa que cuando la gente común (no los medios) utilizan Twitter, lo usan para conversar con sus contactos. Estos usuarios "activos" son los que ejemplifican la dimensión de red social de Twitter.[47]

Twitter enfatizó su estrategia de red de noticias en noviembre de 2009 cuando cambió

la pregunta que le hacía a sus usuarios de ¿Qué estas haciendo? a ¿Qué está pasando?.

La revista *Entertainment Weekly* posicionó a Twitter en su ranking de los 10 mejores de la década. Refiriéndose a twitter de la siguiente forma: "limitarse a 140 caracteres (el máximo que se permite por cada mensaje en esta herramienta de conexión social adictiva y diabólica) es fácil "

En noviembre de 2010, Biz Stone, co-fundador de la compañía expresó por primera vez la idea de Twitter como red de noticias, un concepto de una red de servicios de noticias en la que estuvo trabajando por años [46].

2.3. Conclusiones

Hemos presentado en este capítulo a la red social Twitter, mencionado algunas de sus características y de los debates que se han generado a su alrededor.

En el próximo capítulo haremos una introducción a los procesos estocásticos y algunos temas sobre ellos que retomaremos en el Capítulo 4 para plantear un modelo que permita explicar algunas dinámicas de esta red social.

Capítulo 3

Procesos Estocásticos

3.1. Introducción

En este capítulo introduciremos algunos conceptos y definiciones de procesos estocásticos que serán utilizados en los capítulos siguientes para plantear el modelo de Twitter. Las principales referencias utilizadas para este capítulo son [24], [23], [25].

3.2. Definiciones

Definición 3.1 *Un proceso estocástico es una colección de variables aleatorias*

$$X = \{X(t)|t \in T\},$$

definidas en un espacio de probabilidad común $(\Omega, \mathcal{A}, P)$*. La colección está indexada por un parámetro* $t \in T \subset \mathbb{R}$*. El parámetro* t *usualmente se interpreta como el tiempo.*

.

Puede pensarse al proceso como una función $X : T \times \Omega \to R$ tal que $X(t, \cdot)$ es $\mathcal{A}$-medible en $\omega \in \Omega$ para cada $t \in T$.

Podemos establecer una convención y decir que $X(t) = X_t$.

3.2.1. Clasificaciones

Encontramos en [24] una clasificación de procesos estocásticos generales. Los principales elementos que distinguen a los procesos son:

- el espacio de estados,
- el espacio de índices,
- la estructura de dependencia entre las variables X_t.

Espacio de Estados

Es el conjunto de valores que puede tomar las variables X_t. Puede ser discreto o continuo, de cualquier dimensión. Según cómo sea este conjunto nos referimos a un proceso discreto (natural o entero), o a un proceso de valores reales, etc. Si tiene dimensión mayor que uno hablaremos de un proceso multidimensional discreto, continuo, etc.

Parámetro de indexación T

Si T es un conjunto discreto decimos que el proceso es a tiempo discreto. Muchas veces cuando el tiempo sea discreto podremos escribir X_n. En cambio si $T = [0, \infty)$, X_t es un proceso a tiempo continuo. Por último el conjunto T puede ser multidimensional. En este caso puede no tener una interpretación temporal, un ejemplo de esto puede ser un proceso $X_{t=(x,y)}$, donde el vector $t = (x, y)$ es una posición de latitud y longitud y el valor X_t representa la altura de una ola en ese lugar.

Relaciones de dependencia

A continuación vamos a exponer algunas de los distintas estructuras de dependencia que se pueden observar en los procesos estocásticos.

- **Incrementos estacionarios e independientes**

  Si las variables $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$ son independientes para cualquier elección de $t_1, ..., t_n$ ordenados de menor a mayor, entonces el proceso se dice que tiene incrementos independientes. Si además podemos decir que la distribución del incremento $X_{t_1+h} - X_{t_1}$ depende solo de la longitud h del intervalo y no de t_1 entonces el proceso se dice que tiene incrementos estacionarios.

- **Martingalas**

  Ya sea a tiempo continuo o discreto decimos que un proceso estocástico es una martingala si $E(|X_t|) < \infty$ para todo t y cualquiera sean $t_1 < t_2 < ... < t_{n+1}$ vale que $E(X_{t_n+1}|X_{t_1} = a_1, ..., X_{t_n} = a_n) = a_n$.

- **Procesos de Markov**

  La propiedad de Markov dice que la probabilidad de cualquier comportamiento futuro, dado que el que se conoce el estado actual del proceso, no se modifica si conocemos más estados previos del proceso. En términos formales,

$$P(X_t \in B|X_{t_1} = x_1, ..., X_{t_n} = x_n) = P(X_t \in B|X_{t_n} = x_n)$$

Esto ocurre para todo $t_1 < t_2 < ... < t_n < t$. Gracias a esta propiedad se puede definir una función que llamaremos Probabilidad de Transición de la siguiente manera:

$$P(x, s; t, A) = P(X_t \in A | X_s = x); \;\; t > s.$$

Un proceso de Markov con espacio de estados discreto se llama Cadena de Markov. Si las realizaciones o trayectorias del proceso son continuas el proceso se denomina difusión. Algunas propiedades y características de los procesos de Markov se estudiarán en las secciones siguientes, ya que los procesos que utilizamos para modelar la emisión de tweets pertenecen a esta familia.

- **Procesos estacionarios**

  Un proceso X_t se dice estrictamente estacionario si la distribución conjunta de la familia de variables aleatorias $(X_{t_1+h}, ..., X_{t_n+h})$ y $(X_{t_1}, ..., X_{t_n})$ es la misma para todo $h > 0$ y elecciones arbitrarias de $t_1, t_2, ..., t_n$. Esta condición asegura que el proceso se encuentra en un estado de equilibrio probabilístico y que los instantes de tiempo en los que observamos al mismo no tienen relevancia. En particular también implica que X_t tiene la misma distribución para todo t.

- **Procesos Ergódicos**

  El término ergódico proviene del griego ergon que significa trabajo y del vocablo hodos que significa trayectoria. Fue acuñado por L. Boltzmann al estudiar algunos problemas de la mecánica estadística. Si se cumple la hipótesis ergódica en un sistema dinámico, se establece que los promedios temporales son iguales a los promedios espaciales [33].

  Una consecuencia de esta propiedad, es que en estos procesos se pueden hacer estimaciones a partir de una única traza del proceso, y existen muchos casos prácticos, estudios de temperatura, de vientos, etc. en los que, como sólo se puede obtener una única traza, es útil saber si el proceso estudiado es ergódico o no.

  Sea X_t un proceso estocástico, decimos que es ergódico si se cumple que:

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T f(X_t) dt = \int_{-\infty}^{\infty} f(x) \tilde{p}(x) dx, \tag{3.1}$$

  donde $\tilde{p}(x)$ es la distribución estacionaria del proceso.

- **Procesos Mixing**

  Las condiciones de mixing describen situaciones de dependencia débiles en la que los eventos más cercanos en el tiempo son más dependientes entre sí que los eventos

que se encuentran alejados. Los coeficientes de mixing están basados en coeficientes que describen la dependencia de dos σ-álgebras de un espacio de probabilidades.

Definición 3.2 *Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad, $\mathcal{F}, \mathcal{G}$ sub σ-álgebras de $\mathcal{A}$. Definimos los siguientes coeficientes:*

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup\{P(A \cap B) - P(A)P(B) \ : A \in \mathcal{F},\ B \in \mathcal{G}\} \tag{3.2}$$

Este coeficiente se conoce como α-mixing.

$$\beta(\mathcal{F}, \mathcal{G}) = \sup\left\{\frac{1}{2}\sum_{i\in I}\sum_{j\in J}|P(A_i \cap B_j) - P(Ai)P(Bj)| \ :\ A_i \in \mathcal{F},\ B_j \in \mathcal{G}\right\} \tag{3.3}$$

siendo $(A_i)_{i\in I}$ y $(Bj)_{j\in J}$ particiones de Ω. Este coeficiente se conoce como β-mixing.

$$\phi(\mathcal{F}, \mathcal{G}) = \sup\{|P(B|A) - P(B)| \ :\ A \in \mathcal{F},\ P(A) \neq 0\ ,\ B \in \mathcal{G}\} \tag{3.4}$$

Este coeficiente se conoce como ϕ-mixing.

$$\varphi(\mathcal{F}, \mathcal{G}) = \sup\left\{\frac{|P(A \cap B) - P(A)P(B)|}{P(A)P(B)}\ A \in \mathcal{F},\ P(A) \neq 0\ ,\ B \in \mathcal{G},\ P(B) \neq 0\right\} \tag{3.5}$$

Este coeficiente se conoce como φ-mixing.

$$\rho(\mathcal{F}, \mathcal{G}) = \sup\{|Corr(X, Y)| : X \in L^2(\mathcal{F}), Y \in L^2(\mathcal{G})\} \tag{3.6}$$

donde $L^2(\mathcal{F}) = \{X : X$ es $\mathcal{F} -$ medible y $E(X^2) < \infty\}$ Por último a esta cantidad se la denomina coeficiente de ρ-mixing.

A partir de estos coeficientes definidos para σ-álgebras, podemos definir para un proceso estocástico las respectivas condiciones de mixing de la siguiente forma.

Consideremos un proceso estocástico $X = \{X(t)|t \in T\}$, donde $T \in \mathbb{R}$ *o* $\mathbb{N}$. Para todo $A \subset T$ consideramos la σ-álgebra generada por las variables

$$\{X_t :\ t \in A\} = \sigma^X(A).$$

Definición 3.3 *Definimos para el proceso* $X_t,\ t \in T$ *el coeficiente:*

$$\alpha_m^X = \sup\{\alpha(\sigma^X(A), \sigma^X(B)) :\ A, B \subset T,\ d(A,B) > m,\ t > s + m,\ \forall s \in A,\ \forall t \in B\} \tag{3.7}$$

También se puede definir $\alpha_m^X(a,b)$, para $a, b \in \mathbb{N}$ como en (3.2) y con la condición adicional

$$card(A) \leq a,\ card(B) \leq b.$$

Finalmente un proceso se dice que es ν^X-mixing si $\nu_m^X \to 0$ cuando $m \to \infty$ con $\nu = \alpha,\ \beta,\ \phi,\ \varphi,\ \rho$ y análogamente se define $\nu_{(a,b)}^X$-mixing.

Por último mencionaremos un resultado que se puede ver en [14] que relaciona los coeficientes de mixing de la siguiente manera

$$2\alpha(\mathcal{F}, \mathcal{G}) \leq \beta(\mathcal{F}, \mathcal{G}) \leq \varphi(\mathcal{F}, \mathcal{G}) \leq \frac{1}{2}\phi(\mathcal{F}, \mathcal{G}), \tag{3.8}$$

$$4\alpha(\mathcal{F}, \mathcal{G}) \leq \rho(\mathcal{F}, \mathcal{G}) \leq \varphi^{\frac{1}{2}}(\mathcal{F}, \mathcal{G})\varphi^{\frac{1}{2}}(\mathcal{G}, \mathcal{F}), \tag{3.9}$$

$$\rho(\mathcal{F}, \mathcal{G}) \leq \phi(\mathcal{F}, \mathcal{G}). \tag{3.10}$$

Esta desigualdad nos indica que el coeficiente de α-mixing es el más debil por lo tanto cuando se imponen condiciones de dependencia débil, se lo hace sobre el coeficiente α-mixing porque los demás son demasiado fuertes.

Tamaño del Mixing

La velocidad a la que X_t yX_{t+m} se vuelven independientes a medida que $m \to \infty$ se puede medir con una magnitud llamada "tamaño". Esta definición se encuentra en [22].

Decimos que X_t es strong-mixing de tamaño $\lambda > 0$ si:

$$\alpha_m = o(m^{-\lambda}).$$

Se puede intepretar esta cantidad, como mide la velocidad con la que el proceso pierde la memoria, es decir la velocidad a la que X_t se vuelve independiente de X_{t+m}.

Si decimos que X_t es un proceso mixing fuerte de tamaño dos entonces para todo $\delta \leq 2$ se cumple que:

$$m^{\delta} \times \sup_{t \in \mathbb{Z}} \sup_{G \in \mathcal{F}_{-\infty}^{t} H \in \mathcal{F}_{t+m}^{-\infty}} |P(G \cup H) - P(g)P(h)| \to 0,$$

cuando $m \to \infty$. Entonces la velocidad con la que X_t y X_{t+m} se vuelven independiente es m^2. Obvservemos que cuanto mayor es el tamaño, se independizan a mayor velocidad.

En este sentido definimos un proceso geométricamente strong-mixing si existe un número $\iota \in (0, 1)$ tal que

$$\alpha_m = o(\iota^m).$$

Es fácil ver que la condición de geométricamente mixing fuerte implica mixing fuerte de cualquier tamaño λ. Es una condición más fuerte y representa procesos en los que X_t y X_{t+m} se vuelven independientes comparativamente rápido.

3.3. Procesos estocásticos Markovianos

Un proceso estocástico markoviano puede ser de distinta naturaleza según el tiempo y el espacio de estados. Si el tiempo y el espacio de estados es discreto, el proceso se denomina una cadena de Markov. En este trabajo vamos a estudiar procesos markovianos a tiempo continuo y espacio de estados continuo.

La markovianidad de estos procesos se traduce en que la probabilidad de transición de un estado a otro se puede escribir:

$$p(s, x; t, y) = \int_{-\infty}^{\infty} p(s, x; \tau, z) p(\tau, z; t, y) dz, \tag{3.11}$$

para todo $s \leq \tau \leq t$ y $x, y \in \mathbb{R}$.

Un proceso con estas características se dice homogéneo si estas probabilidades sólo dependen de los estados de salida y llegada y del lapso $t - s$ de tiempo.

3.4. Procesos de Difusión

3.4.1. Definición

Un proceso de Markov con probabilidad de transición $p(s, x; t, y)$ se dice una difusión si los siguientes tres límites existen para todo $\epsilon > 0$, $s \geq 0$, $x \in \mathbb{R}$:

$$\lim_{t \searrow s} \frac{1}{t-s} \int_{|y-x|>\epsilon} p(s,x;t,y)dy = 0, \tag{3.12}$$

$$\lim_{t \searrow s} \frac{1}{t-s} \int_{|y-x|<\epsilon} (y-x)p(s,x;t,y)dy = b(s,x), \tag{3.13}$$

$$\lim_{t \searrow s} \frac{1}{t-s} \int_{|y-x|<\epsilon} (y-x)^2 p(s,x;t,y)dy = \sigma^2(s,x), \tag{3.14}$$

donde b y σ son funciones bien definidas.

La condición (3.12) evita que la difusión tenga saltos instántaneos y la condición (3.13) implica que

$$b(s,x) = \lim_{t \searrow s} \frac{1}{t-s} E(X_t - X_s | X_s = x). \tag{3.15}$$

Este cociente incremental nos da la idea que la cantidad $b(s,x)$ mide la variación instantánea de la media del proceso, dado que $X_s = x$. Esta función se llama drift de la difusión.

La condición (3.14) implica que

$$\sigma^2(s,x) = \lim_{t \searrow s} \frac{1}{t-s} E((X_t - X_s)^2 | X_s = x), \tag{3.16}$$

y mide la tasa de variación instantánea de la fluctuación al cuadrado del proceso dado que $X_s = x$. Esta función se denomina el coeficiente de difusión a tiempo s y posición x.

Cuando el drift b y el coeficiente de difusión σ de un proceso de difusión son funciones suaves, entonces la densidad de transición $p(s,x;t,y)$ también satisface ecuaciones diferenciales parciales. Estas son:

Ecuación "backward" de Kolmogorov

$$\frac{\partial p}{\partial s} + b(s,x)\frac{\partial}{\partial t}p + \frac{1}{2}\sigma^2(s,x)\frac{\partial^2}{\partial x^2}p = 0, \quad (t,y) \text{ fijos.} \tag{3.17}$$

Ecuación "forward" de Kolmogorov también se la conoce como ecuación de Fokker-Planck y es la adjunta formal de la ecuación 3.17

$$\frac{\partial p}{\partial t} + \frac{\partial}{\partial y}\{b(t,y)p\} - \frac{1}{2}\frac{\partial^2}{\partial y^2}\{\sigma^2(t,y)p\} = 0, \quad (s,x) \text{ fijos.} \tag{3.18}$$

3.4.2. Difusiones y ecuaciones diferenciales estocásticas (SDE)

Vamos a explicitar una relación entre los procesos de difusión y aquellos que resuelven ecuaciones diferenciales estocásticas. Hacemos hincapié en esta relación porque algunas propiedades que se estudian más adelante están planteadas en términos de SDE (stochastic differential equations) y otras en términos de procesos de difusión. Creemos que esta dualidad, si bien se podría evitar y unificar la teoría, aporta a la multiplicidad de interpretaciones que tienen estos conceptos ya sean pensados dentro de la matemática, o en el campo de las aplicaciones.

Procesos definidos a partir del Movimiento Browniano

El Movimiento Browniano o proceso de Wiener, es un proceso estocástico muy conocido en la literatura que puede ser tomado como base para definir nuevos procesos. Los conceptos aquí explicados se pueden encontrar en [30]. Una forma de construir procesos a partir del movimiento browniano es definir procesos estocásticos cuyos incrementos dependen de éste mismo.

Definición 3.4 *Movimiento Browniano aritmético (MBA) es un proceso estocástico definido en términos de un proceso de Wiener del siguiente modo:*

$$x_t - x_{t-1} = \Delta x = \mu \Delta t + \sigma \Delta z, \tag{3.19}$$

donde μ y σ son constantes y Δz es el incremento de un proceso de Wiener.

La constante μ representa la tasa esperada de cambio de la variable x por unidad de tiempo. En efecto, si eliminásemos el segundo sumando tendríamos que $x_t = x_{t-1} + \mu \Delta t$, el término $\sigma \Delta z$ "perturba" la tendencia marcada por $\mu \Delta t$. Dicha perturbación es σ veces un proceso de Wiener Δz.

Otro proceso a partir del Movimiento Browniano es el Movimiento Browniano Geométrico (MBG).

Definición 3.5 *Movimiento Browniano Geométrico (MBG) es un proceso estocástico definido a partir de un proceso de Wiener del modo siguiente:*

$$x_t - x_{t-1} = \Delta x = \mu x_{t-1} \Delta t + \sigma x_{t-1} \Delta z, \tag{3.20}$$

donde μ y σ son constantes y Δz es el incremento de un proceso de Wiener.

Definición 3.6 *Un proceso de Itô o proceso de difusión es un proceso de Wiener generalizado en el que los parámetros μ y σ son ahora funciones de la propia variable y del tiempo:*

$$X_t - X_{t-1} = \Delta X = b(X_t, t) \Delta t + \sigma(X_t, t) \Delta z. \tag{3.21}$$

Si hacemos Δt,Δz tender a 0, podemos expresar formalmente:

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dz. \tag{3.22}$$

Un proceso de Itô, se puede pensar al mismo como solución de una SDE adecuada o se puede considerar un proceso de Markov que cumpla con (3.12), (3.13) y (3.14), donde las funciones de drift de la SDE coincide con la de (3.13) y la de difusión coincide con (3.14).

El enfoque de incrementos es muy útil entre otras cosas para realizar simulaciones numéricas de los procesos.

3.4.3. Continuidad de las trayectorias de las difusiones

De la condición (3.12) podemos esperar que las trayectorias del proceso cumplan alguna condición de continuidad. Más aún, se puede probar que las trayectorias son casi seguramente funciones continuas del tiempo, aunque no necesariamente diferenciables.

En general se puede definir la continuidad de un proceso estocástico de diferentes formas según las diversas convergencias que existen para sucesiones de variables aleatorias.

En particular podemos definir:

Definición 3.7 *1. Continuidad con probabilidad 1.*

$$P(\{\omega \in \Omega : \lim_{s \to t} |X(s, \omega) - X(t, \omega)| = 0\} = 1. \tag{3.23}$$

2. Continuidad en media cuadrática: Si $E[(X_t)^2] < \infty$ *y*

$$\lim_{s \to t} E(|X_s - X_t|^2) = 0. \tag{3.24}$$

3. Continuidad en probabilidad:

$$\lim_{s \to t} P(\{\omega \in \Omega : |X(s, \omega) - X(t, \omega)| \geq \epsilon\}) = 0, \ \forall \epsilon > 0. \tag{3.25}$$

4. Continuidad en distribución:

$$\lim_{s \to t} F_s(x) = F_t(x), \text{ para todos los puntos de continuidad de } F_t. \tag{3.26}$$

3.4.4. Difusiones estacionarias y ergódicas

Definición 3.8 *Un proceso X_t de difusión se dice que es homogéneo si las funciones de drift y de difusión sólo dependen del espacio de estados y no del tiempo. Es decir que X_t resuelve la ecuación*

$$dX_t = b(X_t)dt + \sigma(X_t)dz. \tag{3.27}$$

Sea X_t un proceso de difusión homogéneo, como todo proceso estocástico, puede presentar distintas relaciones de dependencia temporales. Queremos encontrar alguna condición sobre las funciones b y σ para que X_t sea estacionario y ergódico. En [26] vemos que para que esto ocurra la función $b(x)$ y $\sigma(x)$ deben cumplir las siguientes condiciones:

$$\int_0^x \frac{b(u)}{\sigma(u)^2}du \to \infty, \text{cuando } |x| \to \infty, \tag{3.28}$$

y además,

$$G(b) \equiv G = \int_{-\infty}^{\infty} \sigma(x)^{-2} \exp 2 \int_0^x \frac{b(u)}{\sigma(u)^2}du\,dx < \infty. \tag{3.29}$$

Las condiciones (3.28) y (3.29) aseguran que el proceso $X_t,\ t \geq 0$ posee una distribución estacionaria

$$F(x) = G(b)^{-1} \int_{-\infty}^{x} \sigma(y)^{-2} \exp 2 \int_0^y \frac{b(u)}{\sigma(u)^2}dudy. \tag{3.30}$$

Este resultado también se puede encontrar en [28].

También podemos observar que bajo éstas condiciones (3.28) y (3.29), se cumple que el proceso de difusión X_t es ergódico, es decir que para cualquier función medible g respecto de la distribución invariante (3.30) que tenga esperanza finita $E_F(g) < \infty$ vale que:

$$P\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T g(X_t)du = E_F(g)\right) = 1. \tag{3.31}$$

3.4.5. Difusiones Mixing

Vamos a describir ahora las condiciones de los coeficientes de drift y de difusión para que el proceso X_t cumpla alguna de las condiciones de mixing definidas en (3.3). Estos resultados los encontramos en [43] y apuntan a que el proceso sea β-mixing,

Teorema 3.4.1 *Sea X_t solución de la ecuación* (3.22)*. Si la función $b(x)$ cumple que*

$$\langle b(x), \frac{x}{|x|}\rangle < -\frac{r}{|x|}, \quad |x| > M_0,\ r > 0,$$

donde $\langle .,. \rangle$ *es el producto escalar de* $\mathbb{R}^d$*, considerando además que* $r > \frac{d}{2}+1$ *para cualquier* $0 < k < r - \frac{d}{2} - 1$ *con* $m \in (2k+2, 2r-2)$ *vale lo siguiente:*

$$\beta(t) \leq C(x)(1+t)^{-(k+1)}, \quad C(x) = C(1+|x|^m), \tag{3.32}$$

k y m no son necesariamente enteros.

Un proceso de difusión que verifica el **Teorema** 3.4.1, por la fórmula (3.8), resulta que también es un proceso α-mixing.

3.5. Tiempo Local

En esta sección vamos a describir un proceso asociado a X_t que nos servirá más adelante.

El proceso que vamos a definir surge de la necesidad de desarrollar alguna herramienta que nos permita calcular la cantidad de tiempo que permanece un proceso estocástico alrededor de un punto dado. Comenzaremos esta presentación con el Movimiento Browniano.

3.5.1. Tiempo local para el Movimiento Browniano

Sea el conjunto de ceros de una trayectoria del Movimiento Browniano como se considera en [23] definido por:

$$\mathcal{Z}_\omega(0) = \{0 \leq t < \infty;\ B_t(\omega) = 0\}. \tag{3.33}$$

Recordemos que el ω es fijo en este caso. Este conjunto contiene todos los instantes t tales que el Movimiento Browniano B_t es igual a cero. Para este conjunto, en [23], se prueba lo siguiente:

Teorema 3.9 *Para casi todo* $\omega \in \Omega$*, el conjunto* $\mathcal{Z}_\omega(0)$*:*

1. *tiene medida de Lebesgue cero,*
2. *es cerrado y no acotado,*
3. *tiene un punto de acumulación en* $t = 0$*,*
4. *no tiene puntos aislados en* $(0, \infty)$*.*

A partir de este resultado podemos ver que para todo $b \in \mathbb{R}$ y para casi todo ω el conjunto de nivel $\mathcal{Z}_\omega(b)$ también es cerrado, no acotado, de medida de Lebesgue cero y denso en si mismo.

Este conjunto no da ninguna información sobre los tiempos de permanencia que estamos buscando.

Trabajando sobre este tema, P. Levy definió un proceso de dos parámetros:

$$L_t(x) = \lim_{\epsilon \searrow 0} \frac{1}{4\epsilon} \lambda(\{0 \leq s \leq t : |W_s - x| \leq \epsilon\}); \quad t \in [0, \infty),\ x \in \mathbb{R}, \tag{3.34}$$

donde λ indica la medida de Lebesgue y mostró que este límite existe y es finito pero no idénticamente cero.

Se puede mostrar que este proceso se puede elegir continuo en (t, x) y para x fijo, no decreciente en t y constante en cada intervalo del complemento del conjunto cerrado $\mathcal{Z}_\omega(x)$.

Con estas condiciones, existe la derivada $\dfrac{dL_t(x)}{dt}$ y es cero para casi todo punto t según la medida de Lebesgue. Este proceso $L_t(x)$ se denomina tiempo local y y mide la cantidad de tiempo que el proceso permanece en la vecindad del punto x.

Trotter [41], en 1958 probó la existencia del proceso tiempo local para el Movimiento Browniano.

3.5.2. Tiempo Local para un proceso estocástico

Ahora vamos a dar una definición del tiempo local para procesos estocásticos de la forma $X_t, t \in [0, +\infty)$.

Definición 3.10 (Medida de Ocupación) *La medida de ocupación del proceso X_t sobre el intervalo $[0, t]$ es:*

$$\mu_t(B) = \lambda\{s \in [0, t] : X_s \in B\}, \ \ \forall\ B \in \mathcal{B}(\mathbb{R}^d). \tag{3.35}$$

Si μ_t es absolutamente continua respecto de la medida de Lebesgue en $\mathbb{R}^d$, decimos que X_t tiene tiempo local en $[0, t]$ y definimos los tiempos locales $L_t(x) = \dfrac{d\mu_t}{d\lambda}(x)$ para todo $x \in \mathbb{R}$.

Es decir el tiempo local es la derivada de Radón de la medida de ocupación respecto de la medida de Lebesgue.

3.5.3. Tiempo local para procesos de difusión

Para procesos de difusión, el tiempo local también mide la cantidad de tiempo que permanece el proceso en el estado x en el intervalo de tiempo $[0, t]$. Sin embargo podemos observar que la interpretación de éste como la derivada de Radón de la medida de ocupación respecto de la medida de Lebesgue se modifica un poco.

Tiempo local para semimartingalas

El tiempo local para una semimartingala continua se define como el siguiente límite:

$$L_X(t, a) = \lim_{\epsilon \to 0} \int_0^t \mathbb{1}_{[a,a+\epsilon)}(X_s) d[X]_s, \ \forall a, s, \tag{3.36}$$

donde $[X]_t$ es la variación cuadrática del proceso semimartingala continuo en tiempo t.

Sea ahora el proceso X_t una difusión de la forma (3.22). Como los procesos de difusión son semimartingalas, podemos pensar en una versión reescalada del tiempo local definida por

$$\overline{L}_X(t, a) = \frac{L_X(t, a)}{\sigma^2(a)}, \tag{3.37}$$

y como para el proceso X_t vale que $d[X]_s = \sigma^2(X_s)ds$, entonces la fórmula (3.37) se puede interpretar como el tiempo cronológico que el proceso permanece en el estado a.

En el proceso $L_X(t, a)$ el tiempo se mide en unidades de variación cuadrática en cambio en el $\overline{L}_X(t, a)$ el tiempo que se mide es el cronológico.

En otras palabras, $\overline{L}_X(t, a)$ almacena la cantidad de tiempo calendario que el proceso permanece en un entorno de a y se puede definir como tiempo local cronológico.

3.5.4. Medidas de ocupación para una difusión

Las medidas de ocupación o procesos de ocupación asociados a una difusión X_t brindan información de las trayectorias que va siguiendo un proceso.

Estos procesos también se pueden pensar como la integral sobre el espacio de estados del tiempo local de la difusión y nos brindan información de cuanto tiempo el proceso ha estado en un conjunto determinado en el intervalo $[0, t)$.

Sea X_t un proceso de difusión unidimensional.

Definición 3.11 *Llamaremos proceso de ocupación de X_t al siguiente proceso:*

$$\rho_t^X(x) = \lambda\{s \in [0, t] : X_t \leq x\} \tag{3.38}$$

El tiempo local cronológico es la densidad de la distribución de la medida de ocupación del proceso X_t, es decir del proceso (3.38). Dicho de otra manera, el tiempo local cronológico es una versión de la derivada de Radón de la medida de ocupación con respecto de la medida de Lebesgue y es una densidad de ocupación. Podemos escribir:

$$\rho_t^X(x) = \lambda\{s \in [0,t] : X_t \leq x\} = \int_0^t \mathbb{1}_{X_s \leq x} ds = \int_{-\infty}^x \overline{L}_X(t,s) ds.$$

Este proceso lo dividiremos por la longitud del intervalo y definiremos el proceso

$$\mu_t^X(x) = \frac{1}{t}\rho_t^X(x).$$

Este proceso se puede pensar como la proporción del tiempo que el proceso permanece en un conjunto determinado en el intervalo $[0,t]$.

Para procesos d-dimensionales también definiremos el proceso

$$\begin{aligned} \mu_t^{X^1,...,X^d}(x_1,...,x_d) &= \frac{1}{t}\lambda\{s \in [0,t] : (X_t^1,...,X_t^d) \leq (x_1,...,x_d)\} \\ &= \frac{1}{t}\lambda\{s \in [0,t] : (X_t^1 \leq x_1,...,X_t^d \leq x_d\}. \end{aligned} \tag{3.39}$$

Más adelante encontraremos una manera de expresar la distribución conjunta de (3.39) en función de las marginales (3.38).

Capítulo 4

Twitter II: Modelo y Dinámica

4.1. Introducción

En este capítulo presentaremos el modelo que proponemos para estudiar la red social Twitter.

Elegiremos un tipo de proceso estocástico de los estudiados en (3.2.1), y describiremos con su ayuda algunos aspectos de la dinámica de la red social.

La selección del proceso la haremos buscando que sus características se ajusten al funcionamiento de Twitter.

Hacemos hincapié en la filosofía que mencionamos en la Introducción para abordar esta tarea, es decir no pretendemos describir la realidad en su totalidad ni capturar toda la complejidad que presenta Twitter, pero aún asi podemos obtener modelos útiles.

4.2. Modelado de la emisión de Tweets

4.2.1. Proceso de generación de tweets

La cantidad de tweets que uno puede emitir es muy grande y, salvo los tweets por SMS, no tiene cargos extra. Se podría decir que es gratuita (al menos por ahora). Esta cantidad tan grande se debe principalmente a la naturaleza de los tweets, sólo contienen caracteres y hasta 140 de ellos. Estas características hacen que un tweet se pueda escribir en muy poco tiempo y se transmita por la red a muy alta velocidad siempre que el estado de la red en cuanto a tráfico lo permita.

Aunque no deja de ser cierto lo que dice el sitio web [12] que para escribir un tweet sería interesante pensar qué decir (con la lentitud que esto acarrea), en la mayor cantidad de tweets emitidos la velocidad es muy alta y la cantidad de tweets “lentos” la podemos despreciar.

Existen además fenómenos sociales que aumentan esta velocidad de generación de tweets. Se puede dar algún tema sobre el que todos quieran opinar, haciendo que aumenten los usuarios de Twitter que generan tweets.

Un ejemplo de esto, que trascendió a medios masivos digitales y en otros soportes, se puede ver en [13] donde se habla de la cantidad de tweets que se emitieron sobre el intercambio de opiniones entre un actor argentino llamado Ricardo Darín y la presidente de Argentina, Cristina Fernandez.

Si se considera además, que varios usuarios (personas) pueden enviar desde el mismo perfil (como por ejemplo los perfiles de agencias de noticias), la cantidad de tweets por unidad de tiempo se incrementa enormemente.

Por estas razones, proponemos modelar la generación con un proceso continuo. Éste es uno de los supuestos del modelo y es la base para las conclusiones que se derivan después.

La otra característica importante del modelo se refiere a la estructura de dependencia del proceso. Siguiendo las posturas que aparecen en [1] y [38], proponemos una estructura de dependencia markoviana para el proceso.

El comportamiento de los agentes dentro de la red social depende principalmente del pasado reciente y no del pasado remoto. Es razonable que a una persona le moleste un tweet y deje de seguir a ese perfil inmediatamente a que pase un día y cuando vuelva a ingresar a su Twitter se diga a si misma, "este perfil ayer publicó algo que no me gustó por eso hoy lo voy a dejar de seguir".

Por estas dos condiciones, markovianidad y gran cantidad de tweets por unidad de tiempo, es que modelaremos el proceso de emisión de tweets de un perfil o un conjunto de perfiles con un proceso estocástico markoviano a tiempo continuo y espacio de estados continuo.

Estamos interesados en el comportamiento conjunto de varios de estos procesos de emisión, por este motivo es que antes de presentar el modelo de emisión de tweets vamos a proponer distintas formas de construir procesos multidimensionales a partir de los procesos de emisión de tweets de diferentes perfiles.

Segmentos y afinidades

Para empezar a analizar las interacciones entre los perfiles vamos a considerar una segmentación de estos a partir de un usuario p. Esto quiere decir que el usuario del perfil p realiza una clasificación de todos los perfiles entre su conjunto de líderes.

Más precisamente sea Ω el conjunto de todos los perfiles de Twitter. Una segmentación es una partición $\{S_1, ..., S_d\}$ de Ω. A cada perfil $p \in \Omega$ asignamos una segmentación $\{S_1(p), ..., S_d(p)\}$. Cuando no sea necesario omitiremos explicitar la dependencia en p.

Los criterios para realizar esta clasificación no son en sí mismos relevantes, de hecho se pueden enfocar distintas líneas de trabajo para obtener distintas segmentaciones según

objetivos de optimización buscados. Además cada perfil segmenta el universo de Twitter según sus propias preferencias.

Algunos ejemplos de esta segmentación:

1. Según la profesión del perfil: periodista político, deportivo o de espectáculo; político; médicos; deportista (fútbol, automovilismo, etc); líder religioso; actor; etc.

2. Según el estilo de comentario que realiza el perfil: de información; de reflexión; de humor; de denuncia; de convocatoria; etc.

   Este es un tipo de tweet:

   - **Emisor**: Bloomberg TV
     Contenido: EXCLUSIVE: President of St. Louis Fed James Bullard sits down with @Mckonomy @SRuhle. WATCH LIVE http://bloomberg.com/tv.
   - **Emisor:** Radio Fe y Alegría.
     Contenido: Con el puño cerrado se amenaza, con la mano extendida se ofrece ayuda, Fe y Alegría invade los medios este 19 de febrero #EscuelaparalaVida
   - **Emisor:** 20minutos.es
     Contenido: La actriz que da vida a Arya Stark dice que en 'Juego de tronos' hay demasiado sexo para los niños
   - **Emisor:** Wanda Nara
     Contenido: Nuestra película de ayer MATILDA en casa #cine http://instagram.com/p/kCokEmSLMq/

3. Según grupos etarios de los usuarios o su nacionalidad, etc.

Sea $S_1, ..., S_d$ una segmentación de los perfiles de Twitter de algún tipo, realizada por un perfil p. Cada segmento despierta cierto nivel de afinidad sobre p.

Definimos las cantidades α_{jl} como la valoración que realiza un perfil $p \in S_l$ de los tweets que emite un perfil $p' \in S_j$.

Estas afinidades pueden ser positivas, negativas o nulas. Asumiremos que las afinidades se mantienen constantes para todos los perfiles del segmento, es decir que p valora por segmentos y no por perfiles.

Una observación que podemos hacer sobre estos segmentos es que son relativos al perfil p (se podría prescindir del subíndice que refiere al perfil de p). Si bien es posible trabajar con segmentos generales aplicables a todos los perfiles, es decir, suponiendo que todos los perfiles segmentan igual, por el momento no estamos interesados en éstas.

Tipos de tweets

Como vimos, otra forma de clasificar los tweets es a partir de su contenido. Podemos hablar de tweets de opinión, de crítica, de adhesión, de convocatoria, de información, etc.

Sea $C_1, ..., C_d$ una clasificación en categorías de los distintos tipos de tweets que se pueden emitir realizada por el perfil p, podemos definir también ciertas afinidades de un perfil hacia cada una de estas categorías por medio de los coeficientes α^i Este coeficiente representa la afinidad que tiene el perfil p sobre los tweets de tipo i. Observemos que en la segmentación por perfiles, importa quien es el emisor mientras que en esta segmentación sólo importa el contenido.

Podemos también combinar estas afinidades y distinguir por emisor y por tipo simultáneamente.

4.2.2. Modelo de emisión de tweets

Retomando la clasificación de (3.2.1) y las ideas de (4.2.1) modelaremos la emisión de tweets de un perfil o de un conjunto de perfiles mediante un proceso de difusión de la siguiente forma:

Notación 4.2.1 *Si queremos representar los tweets emitidos por un <u>único</u> perfil p' entonces notaremos:*

$$dX_t(p') = b(X_t)dt + \sigma(X_t)dW_t. \tag{4.1}$$

Donde W_t es un movimiento browniano y las funciones $b(x)$ y $\sigma(x)$ verifican las siguientes condiciones:

1. *Las funciones $b(x)$ y $\sigma(x)$ son continuas para todo $x \in \mathbb{R}$.*
2. *Valen las desigualdades $|b(x)-b(y)| < L|x-y|$ y $|\sigma(x)-\sigma(y)| < L|x-y|$ con $L > 0$ y para todo $x, y \in \mathbb{R}$.*
3. *Además se verifican las desigualdades $|b(x)| < L(1+|x|)$ y $|\sigma(x)| < L(1+|x|)$ con $L > 0$ para todo $x \in \mathbb{R}$.*
4. *Existen $0 < \epsilon < K$ tales que $\epsilon < b(x) < K$ para todo $x \in \mathbb{R}$.*
5. *Existen $0 < \nu < M$ tales que $\epsilon < \sigma(x) < M$ para todo $x \in \mathbb{R}$.*

Observación 4.1 *Las tres primeras hipótesis que se piden a las funciones b y σ son las condiciones de existencia y unicidad de solución fuerte de la SDE. Las últimas dos condiciones surgen del modelo considerado. Estamos pensando que el "drift" de la ecuación es positivo, esto se puede interpretar como que el perfil emisor envia tweets con una tasa de incremento y que estos no se pueden borrar.*

Consideremos ahora una segmentación $S_1, ..., S_d$ de perfiles de Twitter respecto del usuario $p \in S_l$. Sea $\mathcal{L}(p) = \{p'_1, ..., p'_r\}$ el conjunto de líderes de p donde $p'_1, ..., p'_{i_1} \in S_1$, $p'_{i_1+1}, ..., p'_{i_1+i_2} \in S_2$, ... ,$p'_{i_1+i_2+...+i_{d-1}+1}, ..., p'_{i_1+i_2+...+i_d} \in S_d$, donde $i_1 + i_2 + ... + i_d = r$.

Vamos a considerar todas las emisiones de tweets de los perfiles del segmento S_i como una sola fuente y las modelaremos con un sólo proceso de difusión X_t^i de la siguiente manera

$$X_t^i = X_0^i + \int_0^t b^i(X_s^i)ds + \int_0^t \sigma^i(X_s^i)dW_s \tag{4.2}$$

Nuevamente las funciones $b(x)$ y $\sigma(x)$ verifican condiciones de existencia y unicidad y W_s es un Movimiento Browniano.

De esta manera obtenemos un proceso d-dimensional

$$(X_t^1, ..., X_t^d) \tag{4.3}$$

que indica la cantidad de tweets que el perfil p recibe de cada segmento en el instante t.

También podríamos definir un proceso d-dimensional si en vez de agrupar los tweets por segmento, los agrupamos según su tipo de contenido, construyendo un proceso similar al (4.3) que cuenta la cantidad de tweets emitidos de cada tipo en cada instante t.

$$(Y_t^1, ..., Y_t^d) \tag{4.4}$$

Observación: El proceso (4.3) suma en una sóla difusión la emisión de varios perfiles, en cambio el proceso (4.4) reparte los tweets de un perfil en las d clases y junta todos los de una misma clase o tipo en un único proceso de difusión.

4.2.3. Proceso de interés

Para un perfil $p \in S_j$ y $p' \in S_l$, vamos a definir el proceso de interés como:

Definición 4.2

$$I_t(p', p) = I_0(p', p) + \alpha_{jl} X_t(p') \tag{4.5}$$

Este proceso de interés mide "el interés que mantienen en el perfil p los tweets que emite el perfil p'".

Si en la definición (4.2) queremos considerar el interés del perfil p frente a todo el segmento i, notaremos este proceso de interés como I_t^i. En este caso no indicamos a p, el perfil seguidor.

Notaremos con $I_0(p', p)$ al interés inicial, α_{jl} representa el nivel de afinidad de los segmentos involucrados y $X_t(p')$ es el proceso de emisión de tweets del perfil p' o de todo un segmento según el caso que estemos analizando.

Sobre este proceso de interés vamos a estudiar, en particular, la probabilidad que éste baje más allá de un umbral dependiendo del signo de la afinidad del perfil seguidor respecto del perfil p' o del segmento o el tipo de tweet. Como el modelo de emisión es del mismo estilo, es decir un proceso de difusión, para los tres, las conclusiones también serán del mismo tipo para los tres procesos sin importar la forma en la que estamos mirando los tweets que ingresan.

Para estudiar esta probabilidad necesitamos definir un umbral:

Definición 4.3 *Vamos a considerar un número positivo u que llamaremos umbral de saturación de manera que si $I_t(p', p) < u$, el perfil de p se satura de los tweets que emite p'.*

Estamos interesados en estudiar la probabilidad de que esta saturación efectivamente se alcance e intentaremos decir algo al respecto del tiempo en el que ocurre. Para ello además definimos:

Definición 4.4 *Llamaremos tiempo de saturación a la variable $T(p', p) = \inf_{t>0}\{I_t(p', p) < u\}$.*

Este tiempo de saturación es el primer momento en el que el perfil p se satura de los tweets que emite p', o de cierto tipo de tweet, etc.

Observación 4.5 *La cantidad $I_0(p', p)$ debe ser mayor que u. Ésto se interpreta de la siguiente manera: por algún motivo el interés inicial de p sobre p', o sobre el segmento i o el tipo de tweet j fue positivo y comenzó a seguirlo. De otra manera, $T(p', p) = 0$, pues p no empieza a seguir a p' nunca. Éste interés inicial se puede originar a partir de una promoción, o de una acción de marketing, o el comentario de algún conocido, etc.*

Observación 4.6 *La cantidad $I_0(p', p)$ se puede modelar de distintas maneras, sin pérdida de generalidad podemos decir que $I_0(p', p) = au,\ a > 1$. Esta constante a se puede pensar universal para todos los perfiles o constante por segmentos, o dependiendo del tiempo y modo de uso de Tweeter del usuario. Incluso se podría llegar a estudiar y cuantificar de alguna forma en futuros trabajos. Se podría pensar que a es una constante que depende del efecto de una acción de marketing (anuncio, promoción, descuento, etc) de una marca sobre los usuarios de Tweeter. En este trabajo consideraremos que es constante.*

Otra consideración del modelo será que $I_0(p', p)$ es determinístico, de manera que en los cálculos no aporta otra fuente de aleatoriedad, esta es una simplificación del modelo que perfectamente se puede desarrollar en trabajos futuros.

4.3. Dinámica del modelo

4.3.1. Saturación

Tiempo de saturación

Sean $p' \in S_j$ y $p \in S_l$, vamos a estudiar el comportamiento de la variable $T(p', p)$ para cuando el signo de α_{jl} es no positivo. Es decir cuando p y p' tienen afinidad negativa o nula.

- Caso $\alpha_{jl} = 0$:

  Si $\alpha_{jl} = 0$, el interés es constante y por la observación 4.6

  $$I_t(p', p) = I_0(p', p) = au > u \text{ para todo } t \in \mathbb{R},$$

  luego el interés nunca baja del umbral u por lo tanto $T(p', p) = \infty$.

- Caso $\alpha_{jl} < 0$:

  Supongamos que $\alpha_{jl} < 0$, es decir que el segmento de p tiene una afinidad negativa al segmento de p'. Nuevamente:

$$\begin{aligned} P(T(p', p) > x) &= P\left(\inf_{t\in[0,x]} au + \alpha_{jl} X_t(p') \geq u\right) \\ &= P\left(\forall t \in [0, x] : X_t(p') \leq \frac{(1-a)u}{\alpha_{jl}}\right) \\ &\leq P\left(\forall t \in [0, x] : \int_0^t \sigma(X_s) dW_s \leq \frac{(1-a)u}{\alpha_{jl}} - \epsilon t\right) \\ &\leq P\left(\forall t \in [0, x] : \int_0^t \sigma(X_s) dW_s \leq \frac{(1-a)u}{\alpha_{jl}} - \epsilon x\right). \end{aligned} \tag{4.6}$$

  Supongamos que x es tal que $\dfrac{(1-a)u}{\alpha_{jl}} - \epsilon x < 0$ entonces puedo aplicar valor absoluto dentro de la desigualdad en 4.6.

$$P\left(\forall t \in [0, x] : \left|\int_0^t \sigma(X_s) dW_s\right| \geq \left|\frac{(1-a)u}{\alpha_{jl}} - \epsilon x\right|\right),$$

  Consideremos los conjuntos

$$A = \left\{\omega : \ \forall t \in [0, x] : \left|\int_0^t \sigma(X_s(\omega)) dW_s(\omega)\right| \geq \left|\frac{(1-a)u}{\alpha_{jl}} - \epsilon x\right|\right\} \text{ y}$$

$$B = \left\{\omega : \ \left|\int_0^x \sigma(X_s(\omega)) dW_s(\omega)\right| \geq \left|\frac{(1-a)u}{\alpha_{jl}} - \epsilon x\right|\right\}.$$

Si ω es una trayectoria que pertenece a A, vale en particular para x que $\left|\int_0^x \sigma(X_s(\omega))dW_s(\omega)\right| \geq \left|\frac{(1-a)u}{\alpha_{jl}} - \epsilon x\right|$, luego $\omega \in B$. Por lo tanto $A \subset B$, por lo tanto vale:

$$P\left(\forall t \in [0,x] : \left|\int_0^t \sigma(X_s)dW_s\right| \geq \left|\frac{(1-a)u}{\alpha_{jl}} - \epsilon x\right|\right) \leq$$
$$\leq P\left(\left|\int_0^x \sigma(X_s)dW_s\right| \geq \left|\frac{(1-a)u}{\alpha_{jl}} - \epsilon x\right|\right).$$

Aplicando desigualdad de martingalas y considerando $\mathbb{A}$ como en el Caso 1,

$$P\left(\left|\int_0^x \sigma(X_s)dW_s\right| \geq \epsilon x - \mathbb{A}\right) \leq \frac{E(|\int_0^x \sigma(X_s)ds|^2)}{(\epsilon x - \mathbb{A})^2}, \tag{4.7}$$

La expresión de la derecha de 4.7 permite aplicar la isometría de Itô y aplicando Fubbini obtenemos que:

$$\frac{E(|\int_0^x \sigma^2(X_s)ds|^2)}{(\epsilon x - \mathbb{A})^2} = \frac{\int_0^x E(\sigma^2(X_s))ds}{(\epsilon x - \mathbb{A})^2}.$$

Como la función $\sigma(x)$ está acotada superiormente por M vale que:

$$\int_0^x E(\sigma^2(X_s))ds < M^2 x,$$

luego tenemos que,

$$\frac{E(|\int_0^x \sigma^2(X_s)ds|^2)}{(\epsilon x - \mathbb{A})^2} \leq \frac{M^2 x}{(\epsilon x - \mathbb{A})^2} \tag{4.8}$$

Analizando los exponentes de x en la expresión $\frac{M^2 x}{(\epsilon x - \mathbb{A})^2}$ vemos que es del orden de $\frac{\mathbb{K}_2}{x}$.
Finalmente hemos probado:

$$P(T(p', p) > x) \leq \frac{\mathbb{K}_2}{x}. \tag{4.9}$$

Como el evento $\{T(p',p) = +\infty\} = \bigcap_{n=1}^{\infty} \{T(p',p) > n\}$ entonces por 4.9 vale que:

$$P\left(T(p',p) = +\infty\right) = P\left(\bigcap_{n=1}^{\infty} T(p',p) > n\right) = \tag{4.10}$$

$$= \lim_{n\to+\infty} P\left(T(p',p) > n\right) \leq \lim_{n\to+\infty} \frac{\mathbb{K}_2}{n} = 0. \tag{4.11}$$

Tomando el complemento de $T(p',p) = +\infty$ nos lleva a que:

$$T(p',p) < +\infty,\ c.s. \tag{4.12}$$

Las conclusiones de los casos negativo y neutro, nos dan las primeras pistas que permiten empezar a describir algunos comportamientos de los perfiles de Twitter. El $\alpha_{jl} = 0$ nos dice que el perfil p no se satura "nunca" de aquellos perfiles, segmentos o tipos de tweets con los que tiene neutralidad. En cambio la conclusión cuando $\alpha_{jl} < 0$ nos dice que si el perfil tiene afinidad negativa, el tiempo de saturación es finito, es decir que se alcanza en algún momento.

Sería interesante analizar el comportamiento que sigue un perfil cuando alcanza su tiempo de saturación. Esto determinaría de alguna manera, cuál es la tendencia de toda la red.

Se puede decir que la saturación es más "intensa" si sólo proviene de un perfil que si el proceso de difusión modela a todo un segmento, ya que no es lo mismo que un periodista deportivo me sature, a que lo haga todo el segmento de periodistas que estoy siguiendo. De la misma manera que no tiene el mismo efecto la saturación de todo un tipo de tweets.

Podríamos asumir tres comportamientos alternativos frente a la saturación.

- No realiza ninguna opción
- Se borra de twitter
- Elimina al perfil que lo saturó de su conjunto de líderes.

Se pueden analizar cada una de las opciones anteriores y su estudio daría origen a distintos comportamientos asintóticos de toda la red. Más adelante discutiremos sobre algunas de ellas.

4.3.2. Cotas para el tiempo de saturación

En la sección anterior probamos que para un par de perfiles p y p' de segmentos no afines (con afinidad negativa), el tiempo de saturación es finito. A continuación vamos a

pedir algunas condiciones para el proceso de emisión de twits de p', para obtener probabilidades más precisas de este tiempo de saturación, concretamente acotaremos inferiormente la $P(T(p', p) < x)$.

En lo que sigue supondremos que las funciones b y σ del proceso de generación de tweets $X_t(p')$ son constantes y verifican las condiciones de solución de una SDE.

Es decir que

$$X_t(p') = bt + \sigma W_t, \quad t > 0.$$

Seguimos considerando $\alpha_{jl} < 0$, $I_0(p', p)$ y todas las constantes previas como en la sección anterior. Nuevamente,

$$\begin{aligned} P(T(p', p) > x) &= P\left(\inf_{t\in[0,x]} au + \alpha_{jl} X_t(p') \geq u\right) \\ &= P\left(\forall t \in [0, x] : X_t(p') \leq \frac{(1-a)u}{\alpha_{jl}}\right) \\ &\leq P\left(\forall t \in [0, x] : W_t \leq \frac{(1-a)u}{\sigma\alpha_{jl}} - \frac{bt}{\sigma}\right) \\ &= P\left(\forall t \in [0, x] : W_t \leq \mathbb{A} - \frac{bt}{\sigma}\right). \end{aligned}$$

Vamos a aplicar una transformación decreciente dentro de la probabilidad en ambos miembros de la desigualdad,

$$P\left(\forall t \in [0, x] : W_t \leq \mathbb{A} - \frac{bt}{\sigma}\right) = P\left(\forall t \in [0, x] : e^{-W_t} > e^{\frac{bt}{\sigma} - \mathbb{A}}\right).$$

Como la función es decreciente vale

$$P\left(\forall t \in [0, x] : e^{-W_t} > e^{\frac{bt}{\sigma} - \mathbb{A}}\right) \leq P\left(\sup_{t\in[0,x]} e^{-W_t} > e^{\frac{bt}{\sigma} - \mathbb{A}}\right).$$

Por la propiedad del Movimiento Browniano, el proceso $-W_t = B_t$ se puede considerar también un Movimiento Browniano [15]. Con un ligero abuso de notación, vamos a notar W_t para el proceso B_t. Con esta aclaración continuamos,

$$P\left(\sup_{t\in[0,x]} e^{B_t} > e^{\frac{bt}{\sigma} - \mathbb{A}}\right) \overset{not}{=} P\left(\sup_{t\in[0,x]} e^{W_t} > e^{\frac{bt}{\sigma} - \mathbb{A}}\right) \leq P\left(\sup_{t\in[0,x]} e^{W_t} > e^{\frac{bx}{\sigma} - \mathbb{A}}\right).$$

La función e^x es una función convexa y como W_t es un martingala, entonces e^{W_t} es una submartingala.

Como e^{W_t} es submartingala puedo aplicar la desigualdad de Doobs para submartingalas [23] (pag. 13) a la última expresión y obtenemos,

$$P\left(\sup_{t\in[0,x]} e^{W_t} > e^{\frac{bx}{\sigma}-\mathbb{A}}\right) \leq \frac{E(e^{W_x})}{e^{\frac{bx}{\sigma}-\mathbb{A}}}, \tag{4.13}$$

y dado que W_x es el browniano en $t = x$, tiene distribución $N(0, x)$ luego

$$E(e^{W_x}) = e^{\frac{x}{2}}.$$

Finalmente reemplazando en (4.13) vale que:

$$\frac{E(e^{W_x})}{e^{\frac{bx}{\sigma}-\mathbb{A}}} = e^{\frac{x}{2}-\frac{bx}{\sigma}+\mathbb{A}} = e^{(\frac{1}{2}-\frac{b}{\sigma})x}e^{\mathbb{A}}. \tag{4.14}$$

Si además se cumple que

$$\frac{1}{2} - \frac{b}{\sigma} < 0, \tag{4.15}$$

vale que $P(T(p', p) > x) \leq \mathbb{K}_2 e^{(\frac{1}{2}-\frac{b}{\sigma})x}$, y por lo tanto

$$P(T(p', p) \leq x) \geq 1 - \mathbb{K}_2 e^{(\frac{1}{2}-\frac{b}{\sigma})x}. \tag{4.16}$$

La interpretación de esto es que si el coeficiente de difusión σ no es tan "grande" como el de drift para que valga la acotación (4.15), la probabilidad de que p' sature a p antes del tiempo x está acotada por debajo. Intuitivamente si la variabilidad del proceso de generación de tweets, no es tan alta en comparación al drift, la cantidad de tweets emitidos en algún momento aumentará mucho al punto de bajar el interés y alcanzar el umbral de saturación del perfil p.

Generalización

Vamos a estudiar ahora el caso general.

Consideremos que el proceso de generación de tweets $X_t(p')$ está definido como antes por

$$X_t(p') = \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s,$$

donde $0 < \epsilon < b(x) < M$ y $0 < \eta < \sigma(x) < k$.

Calculemos nuevamente la probabilidad de que $P(T(p', p) > x)$,

$$\begin{aligned} P(T(p',p)>x) &= P(\forall t\in[0,x]:\int_0^t b(X_s)ds+\int_0^t \sigma(X_s)dW_s<\mathbb{A}) \\ &\leq P(\forall t\in[0,x]:t\epsilon+\int_0^t \sigma(X_s)dW_s<\mathbb{A}) \\ &= P(\forall t\in[0,x]:\int_0^t \sigma(X_s)dW_s<\mathbb{A}-t\epsilon) \\ &= P(\forall t\in[0,x]:-\int_0^t \sigma(X_s)dW_s>t\epsilon-\mathbb{A}). \end{aligned}$$

Como $\int_0^t \sigma(X_s)dW_s$ es una integral estocástica, si consideramos $-W_t=B_t$, se cumple la siguiente igualdaden distribución,

$$-\int_0^t \sigma(X_s)dW_s=\int_0^t \sigma(X_s)dB_s.$$

Nuevamente cometiendo un abuso de notación $\int_0^t \sigma(X_s)dB_s \overset{not}{=} \int_0^t \sigma(X_s)dW_s$.
Luego si $\alpha>0$,

$$P\left(\forall t\in[0,x]:\int_0^t \sigma(X_s)dW_s>t\epsilon-\mathbb{A}\right) = P\left(\forall t\in[0,x]:e^{\alpha\int_0^t \sigma(X_s)dW_s}>e^{\alpha(t\epsilon-\mathbb{A})}\right)$$

Consideremos los conjuntos:

$$\begin{aligned} A&=\{\omega:\ \forall t\in[0,x]\ e^{\alpha\int_0^t \sigma(X_s(\omega))dW_s(\omega)}>e^{\alpha(t\epsilon-\mathbb{A})}\} \\ B&=\{\omega:\ \ e^{\alpha\int_0^x \sigma(X_s(\omega))dW_s(\omega)}>e^{\alpha(x\epsilon-\mathbb{A})}\}. \end{aligned}$$

Si ω es un elemento de A, entonces en particular se cumple que cuando $t=x$, $e^{\alpha\int_0^x \sigma(X_s(\omega))dW_s(\omega)}>e^{\alpha(x\epsilon-\mathbb{A})}$, luego $A\subseteq B$. Con esta inclusión podemos decir que,

$$\begin{aligned} P\left(\forall t\in[0,x]:e^{\alpha\int_0^t \sigma(X_s)dW_s}>e^{\alpha(t\epsilon-\mathbb{A})}\right) &\leq P\left(e^{\alpha\int_0^x \sigma(X_s)dW_s}>e^{\alpha(x\epsilon-\mathbb{A})}\right) \\ &\leq \frac{E(e^{\alpha\int_0^x \sigma(X_s)dW_s})}{e^{\alpha(x\epsilon-\mathbb{A})}}. \end{aligned} \tag{4.17}$$

A continuación vamos a buscar alguna cota superior para $E(e^{\alpha\int_0^x \sigma(X_s)dW_s})$. Llamaremos $M_u=\int_0^u \sigma(X_s)dW_s$, $h'(u)=E(e^{\alpha M_u})$ y $f(M_u)=e^{\alpha M_u}$, aplicando a $f(M_u)$ la fórmula de Itô obtenemos,

$$df(M_u)=\alpha e^{\alpha M_u}dM_u+\frac{1}{2}\alpha^2 e^{\alpha M_u}\sigma^2(M_u)du, \tag{4.18}$$

integrando ambos miembros de (4.18) tenemos,

$$f(M_u) - f(M_0) = \int_0^u \alpha e^{\alpha M_s} dM_s + \frac{1}{2}\alpha^2 \int_0^u e^{\alpha M_s}\sigma^2(M_s)ds. \tag{4.19}$$

Calculamos la esperanza en (4.19) y tenemos que

$$E\left(f(M_u) - f(M_0)\right) = E\left(\int_0^u \alpha e^{\alpha M_s} dM_s + \frac{1}{2}\alpha^2 \int_0^u e^{\alpha M_s}\sigma^2(M_s)ds\right). \tag{4.20}$$

El término $E\left(\int_0^u \alpha e^{\alpha M_s} dM_s\right)$ es cero porque es una integral de un proceso martingala centrado, luego

$$E\left(f(M_u)\right) - E(f(M_0)) = E\left(\frac{1}{2}\alpha^2 \int_0^u e^{\alpha M_s}\sigma^2(M_s)ds\right).$$

Finalmente, como $M_0 = \int_0^0 \sigma(X_s)dB_s = 0$ concluimos:

$$h'(M_u) = E(f(M_u)) = \frac{1}{2}\alpha^2 E\left(\int_0^u e^{\alpha M_s}\sigma^2(M_s)ds\right) + 1.$$

Aplicando Fubini podemos intercambiar la integral con la esperanza, se tiene que

$$\begin{aligned}
\frac{1}{2}\alpha^2 E\left(\int_0^u e^{\alpha M_s}\sigma^2(M_s)ds\right) &= \frac{1}{2}\alpha^2 \int_0^u E(e^{\alpha M_s}\sigma^2(M_s))ds \\
&\leq \frac{1}{2}k^2\alpha^2 \int_0^u E(e^{\alpha M_s})ds \\
&= \frac{1}{2}k^2\alpha^2 \int_0^u h'(M_s)ds \\
&= \frac{1}{2}k^2\alpha^2 (h(M_u) - h(0))ds
\end{aligned}$$

Hemos obtenido la siguiente desigualdad

$$h'(M_u) \leq \frac{1}{2}k^2\alpha^2(h(M_u) - h(0)) + 1.$$

Si reemplazamos $h(0)$ por el valor $\dfrac{2}{\alpha^2 k^2}$ tenemos que

$$h'(M_u) \leq \frac{1}{2}k^2\alpha^2 h(M_u). \tag{4.21}$$

Como $h(M_u) > 0$ para cualquier valor de $t > 0$,

$$\frac{h'(M_u)}{h(M_u)} \leq \frac{1}{2}\alpha^2 k^2,$$

integrando ambos miembros, operando y teniendo en cuenta la acotación de la función $\sigma(x)$

$$\begin{aligned} \int_0^u \frac{h'(M_s)}{h(M_s)} ds &\leq \frac{1}{2}\alpha^2 k^2 u \\ \log \frac{h(M_u)}{h(M_0)} &\leq \frac{1}{2}\alpha^2 k^2 u \\ h(M_u) &\leq h(0) e^{\frac{1}{2}\alpha^2 k^2 u}. \end{aligned}$$

Reemplazando en (4.21) obtenemos,

$$h'(M_u) \leq e^{\frac{1}{2}\alpha^2 k^2 u}. \tag{4.22}$$

Aplicando este resultado en (4.17), usando la desigualdad (4.22) resulta y tomando $u = x$,

$$\begin{aligned} \frac{E(e^{\alpha \int_0^x \sigma(X_s) dW_s})}{e^{\alpha(x\epsilon - A)}} &\leq \frac{e^{\frac{1}{2}\alpha^2 k^2 x}}{e^{\alpha(x\epsilon - A)}} \\ &= e^{\alpha A} e^{(\frac{1}{2}\alpha^2 k^2 - \alpha\epsilon)x}. \end{aligned}$$

Hemos obtenido así una cota para la probabilidad del tiempo de saturación que depende de α,

$$P(T(p', p) > x) \leq e^{\alpha A} e^{(\frac{1}{2}\alpha^2 k^2 - \alpha\epsilon)x}. \tag{4.23}$$

Si $\alpha < \dfrac{2\epsilon}{k^2}$, el coeficiente que multiplica a x en la exponencial es negativo, luego considerando $\alpha = \dfrac{\epsilon}{k^2}$ en (4.23) obtenemos la siguiente cota

$$P(T(p', p) \leq x) \geq 1 - e^{\frac{\epsilon A}{k^2}} e^{-\frac{\epsilon^2}{2k^2} x} \tag{4.24}$$

Esta cota (4.24) nos dice que la probabilidad de que p se sature antes de x tiempo, es mayor a la exponencial evaluada en x.

En la sección 4.5 haremos simulaciones para ver como se comportan las difusiones respecto a esta cota exponencial que hemos encontrado.

4.3.3. Comportamiento post-saturación

Habíamos señalado previamente que el comportamiento ante un evento de saturación podía ser diverso. Desde sólo dejar de seguir al perfil que provocó la saturación, hasta abandonar la red social.

Vamos a desarrollar un escenario para analizar un poco más esta situación. Supongamos que para cada usuario p existe un número $k(p)$ tal que si los saturan $k(p)$ o más perfiles, p abandonará Twitter.

Sean $\{p'_1, ..., p'_N\}$ los perfiles que sigue p tales que sus respectivos segmentos tienen afinidad negativa. Supongamos que los Movimientos Brownianos que gobiernan los procesos $X_t(p'_i)$ son independientes entre sí. Podemos suponer que los parámetros iniciales de interés, y que los mínimos y máximos de las funciones b y σ son todos iguales. Para las cotas del drift y la difusión basta considerar el mínimo de las primeras b y el máximo para las cotas de las funciones σ.

Bajo estas condiciones, calculemos la probabilidad de que a tiempo x, el perfil p abandone Twitter, es decir que tenga por lo menos $k(p)$ saturaciones.

Llamemos $q_j = P(T(p'_j, p) \leq x)$ y sea $q^*_j = 1 - K_4 e^{-\gamma x}$ la cota exponencial de la saturación obtenida en (4.24).

Definamos las variables aleatorias "y^x_i = el perfil p'_i saturó a p antes del tiempo x", para $i = 1, .., N$.

La variable y^x_i tiene distribución $Ber(q_j)$, además y^x_k es independiente de y^x_j si $i \neq k$.

Luego la variable $Q^x = \sum_{i=1}^{N} y^x_i$ tendrá una distribución $Bin(N, q_j)$.

Por el TCL, si N es suficientemente grande, $Q^x \sim N(Nq_j, Nq_j(1-q_j))$, luego,

$$\begin{aligned} P(Q^x > k) &= P\left(\frac{Q^x - Nq}{\sqrt{Nq(1-q)}} > \frac{k - Nq}{\sqrt{Nq(1-q)}}\right) \\ &= P\left(z > \frac{k - Nq}{\sqrt{Nq(1-q)}}\right). \end{aligned}$$

Asumimos que N es suficientemente grande como para que, $k - Nq_j < 0$ y se cumpla $k - Nq_j < k - Nq^*_j$. Si además x es un tiempo suficiente como para que se cumpla que $q^*_j = 1 - \mathbb{K}_4 e^{-\gamma x} > \frac{1}{2}$, entonces vale que,

$$
\begin{aligned}
P\left(z > \frac{k - Nq_j}{\sqrt{Nq_j(1-q_j)}}\right) &\geq P\left(z > \frac{k - Nq_j^*}{\sqrt{Nq_j(1-q_j)}}\right) \\
&\geq P\left(z > \frac{k - Nq_j^*}{\sqrt{Nq_j^*(1-q_j^*)}}\right) \\
&= 1 - \Phi\left(\frac{k - Nq_j^*}{\sqrt{Nq_j^*(1-q_j^*)}}\right),
\end{aligned}
$$

donde Φ es la función de distribución Normal estándar.

Luego la probabilidad de que el perfil p abandone Twitter en un tiempo x, dado que tiene N líderes con afinidad negativa es mayor que

$$
1 - \Phi\left(\frac{k - Nq_j}{\sqrt{Nq_j(1-q_J)}}\right).
$$

Si el comportamiento del perfil p luego de un evento de saturación es eliminar de su lista de líderes al perfil p_i' que lo saturó, con un cálculo del mismo estilo que el anterior podemos decir que el conjunto de líderes con afinidad negativa quedará vacío.

4.4. Caracterizaciones del modelo

Este comportamiento descripto, determina la siguiente dinámica para el perfil p.

- Si p tiene una cantidad de líderes con afinidad negativa suficientemente grande, entonces p se borra de Twitter, abandona su uso,
- Si p no tiene "demasiados" líderes negativos, los irá eliminando y terminará siguiendo solo a aquellos con los que tiene afinidad positiva.

De esta manera, los usuarios de Twitter solo terminarán siguiendo a aquellos con los que son afines, de manera que la segmentación será cada vez mas fuerte.

Este comportamiento que podríamos describir como "sectario" debilita un poco los argumentos de que Twitter fortalece los vínculos sociales transversalmente. Sólo fortalece los vínculos que ya son afines, y tiende a eliminar los que no lo son. Un comportamiento social igual al de las redes sociales que no virtuales.

4.5. Simulaciones

4.5.1. Introducción

En esta sección vamos a mostrar, con simulaciones numéricas, algunos de los resultados que desarrollamos en la sección anterior sobre los tiempos de saturación. También abordaremos brevemente el tema de la simulación de soluciones de SDE.

Las ecuaciones y resultados que se obtienen al trabajar con procesos estocásticos y ecuaciones diferenciales estocásticas son de naturaleza probabilística y muchas veces no tienen fórmulas explícitas o el esfuerzo requerido para encontrarlas no es recompensado con información nueva suficientemente valiosa. Esta dificultad y la necesidad de estudiar estos procesos han llevado a desarrollar métodos para calcular aproximaciones numéricas de las soluciones buscadas.

A veces podemos buscar una buena aproximación de las trayectorias del proceso, otras veces por ejemplo podemos estar interesados en aproximar la esperanza de algún funcional aplicado al proceso, etc.

Algunas de las herramientas utilizadas en esta sección se pueden encontrar en [25].

4.5.2. Aproximaciones discretas

Los métodos más aplicados son los de aproximaciones de tiempo discreto o método de diferencias. En estos métodos la ecuación o proceso se reemplaza por una ecuación a tiempo discreto que genera los valores $y_1, ..., y_n, ...$ para aproximar los valores intermedios de la solución a tiempos $t_1, ..., t_n, ...$. Los valores se calculan sólo en los tiempos de la partición, y si es necesario, los valores intermedios se pueden obtener con algún método de interpolación.

Se espera que estas aproximaciones mejoren su precisión si el incremento de tiempo $\Delta t = t_{n+1} - t_n$, con $n = 0, 1, 2, ...$ es suficientemente pequeño. El método más simple de aproximación es el de Euler-Maruyama.

Sea X_t un proceso difusión que satisface la ecuación

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t,$$

en el intervalo $t_0 \leq t \leq T$ con valor inicial $X_{t_0} = X_0$.

Para una partición dada $t_0 < t_1 < ... < t_N$ del intervalo de tiempo $[t_0, T]$, una aproximación de Euler es un proceso estocástico continuo $Y_t,\ t_0 < t < T$ que satisface el siguiente esquema iterativo

$$Y_{n+1} = Y_n + b(t_n, Y_n)(t_{n+1} - t_n) + \sigma(t_n, Y_n)(W_{t_n+1} - W_n). \tag{4.25}$$

Consideramos las siguientes notaciones

- $Y_n = Y_{t_n} = Y(t_n)$
- $\Delta_n = t_{n+1} - t_n$
- $\delta = \max_n \Delta_n$, máximo de las longitudes de los intervalos de tiempo.

En general se considerarán particiones equidistantes.

Errores de aproximación

Si conocemos la solución exacta de la ecuación podemos calcular el error de aproximación utilizando el criterio de error absoluto. El error absoluto es la esperanza del valor absoluto de la diferencia entre la solución exacta y la aproximación a tiempo T

$$e = E(|X_T - Y_T|),$$

que mide la cercanía de la trayectoria aproximada con la trayectoria real al final del intervalo de simulación.

Este error se puede estimar por el método Montecarlo que consiste en repitir muchas simulaciones y obtener un $\hat{e}$ junto con sus estimadores de media y varianza para encontrar intervalos de confianza asintóticos para el error. Este enfoque permite, por ejemplo, calcular la cantidad de simulaciones requeridas para obtener un error absoluto acotado por algún valor determinado.

Otras fuentes de error son mencionadas en [21], por ejemplo:

- **Error de muestreo**: este error surge cada vez que se estima una esperanza con la media muestral.
- **Sesgo del generador aleatorio**: este error es inherente al motor aleatorio que uno utiliza para calcular el incremento del Browniano. Las secuencias son pseudo aleatorias y si se repiten muchas simulaciones, se pueden obtener resultados dependientes.
- **Error de redondeo**: las computadoras redondean los números que operan y esto también aporta error.

Estas fuentes de error a veces no son consideradas y se asume que los errores de redondeo o por sesgo aleatorio son despreciables.

A continuación introduciremos algunas nociones sobre los procesos de aproximación que describen cuanto se acerca una solución aproximada a la real.

Convergencia fuerte

Decimos que una aproximación discreta Y^δ con el máximo de los intervalos $\Delta_n = \delta$, converge fuertemente a X si

$$\lim_{\delta \searrow 0} E(|X_T - Y_T^\delta|) = 0. \tag{4.26}$$

Es posible encontrar una relación entre el error absoluto y la longitud del incremento de tiempo de la partición δ.

Decimos que una aproximación converge fuertemente con orden $\gamma > 0$ a tiempo T si existe una constante C que no depende de δ y δ_0 tal que

$$e(\delta) = E(|X_T - Y_T^\delta|) \leq C\delta^\gamma, \tag{4.27}$$

para cada $\delta \in (0, \delta_0)$.

Consistencia

Diremos que un método de aproximación discreta con norma δ es fuertemente consistente si existe una función no negativa $c = c(\delta)$ que

$$\lim_{\delta \searrow 0} c(\delta) = 0, \tag{4.28}$$

tal que

$$E\left(\left|E\left(\frac{Y_{n+1}^\delta - Y_n^\delta}{\Delta_n}|\mathcal{A}_{t_n}\right) - b(t_n, Y_n^\delta)\right|^2\right) \leq c(\delta), \tag{4.29}$$

y

$$E\left(\frac{1}{\Delta_n}\left|Y_{n+1}^\delta - Y_n^\delta - E(Y_{n+1}^\delta - Y_n^\delta|\mathcal{A}_{t_n}) - \sigma(t_n, Y_n^\delta)\Delta W_n\right|^2\right) \leq c(\delta), \tag{4.30}$$

donde $\mathcal{A}_{t_n}$ es una familia de σ-álgebras crecientes preasignada, asociada al Movimiento Browniano de la difusión. Significa que la información para calcular la esperanza es la que está disponible hasta el tiempo t_n.

La condición (4.29) pide que la media del incremento de la aproximación converja a la media del proceso. La condición (4.30) pide que la varianza de la diferencia entre el proceso y la aproximación también converja a 0.

Estas dos condiciones indican que las trayectorias serán cercanas a la real, más aún implican que el método es fuertemente convergente.

Teorema 4.7 *Un método de aproximación Y^δ con norma δ que aproxima un proceso de difusión unidimensional X con $Y_0^\delta = X_0$ converge fuertemente a X, si se satisfacen las condiciones de existencia y unicidad de las ecuaciones diferenciales estocásticas.*

Estabilidad numérica

La propagación de errores iniciales y de redondeo en un método de aproximación debe mantenerse controlada, el concepto de estabilidad numérica para métodos estocásticos hace referencia a estas propagaciones.

Sea Y^{δ} un método de aproximación numérica, con norma $\delta > 0$, empezando a tiempo t_0 en Y_0^{δ}, sea además $\bar{Y}^{\delta}$ la correspondiente aproximación en $\bar{Y}_0^{\delta}$.

Diremos que un método numérico Y^{δ} es estocástica y numéricamente estable para una SDE determinada, si para todo intervalo finito $[t_0, T]$ existe una constante $\Delta_0 > 0$ tal que para cualquier $\epsilon > 0$ y para cada $\delta \in (0, \Delta_0)$

$$\lim_{|Y_0^{\delta} - \bar{Y}_0^{\delta}| \to 0} \sup_{t_0 \leq t \leq T} P(|Y_{n_t}^{\delta} - \bar{Y}_{n_t}^{\delta}| \geq \epsilon) = 0. \tag{4.31}$$

Usualmente los métodos que son estocásticos y numérico-estables se denominan procesos estables numéricamente.

La propagación del error inicial de un método estable permanece acotada cuando uno se maneja en un intervalo acotado. Vale la pena enfatizar que los criterios de estabilidad tienen validez para tamaños de paso $\delta > 0$ que sean menores que un tamaño del paso crítico Δ_0, que usualmente depende del horizonte de tiempo de simulación y de la SDE que se intente resolver.

Este valor crítico puede ser muy pequeño, o en otros casos puede ocurrir que sea muy grande.

Para intervalos de simulación infinitos o variables, se define la estabilidad asintótica.

Se define un método numérico asintóticamente estable si es numéricamente estable y existe una constante $\Delta_0 > 0$ tal que cualquiera sea $\epsilon > 0$ y para cada $\delta \in (0, \Delta_0)$ se cumpla:

$$\lim_{|Y_0^{\delta} - \bar{Y}_0^{\delta}| \to 0} \lim_{T \to \infty} P\left(\sup_{t_0 \leq t \leq T} |Y_{n_t}^{\delta} - \bar{Y}_{n_t}^{\delta}| \geq \epsilon \right) = 0. \tag{4.32}$$

4.5.3. Tiempos de saturación simulados

Programamos en MATLAB 7.12.0.635 (R2011a) simulaciones del proceso de generación de tweets $X_t(p')$ y calculamos para un perfil p tal que la afinidad de sus segmentos es negativa, el primer tiempo en el que supera el valor A. Esto es equivalente a que el proceso de interés $I_t(p', p) > u$. Vamos a utilizar el método de Euler-Maruyama para simular. En [25] (cap.9) se estudian las condiciones que deben cumplir las funciones que $a(x)$ y $\sigma(x)$ para que el método de sea convergente y se prueba que es consistente y converge fuertemente a la solución cuando las funciones b y σ verifican las 3 primeras condiciones que se detallan en 4.2.1.

Para simular debemos definir los valores de todas las constantes utilizadas en las secciones anteriores, y aquellas referidas al método utilizado para simular. Esta información se puede ver en el siguiente cuadro 4.1.

Descripción	Const.	Valor
Umbral para el interés	u	1
Afinidad de los segmentos involucrados	α_j	-350
Constante proporcional interés inicial	a	1,05
Función de drift	$b(x)$	$\epsilon + 10e^{-(1/2)*(x-1)^2}$
Mínimo función de drift	$b(x)$	1600,1
Función coeficiente de difusión	$\sigma(x)$	$0,4 - \frac{1}{\sqrt{\pi * 20}, 2} e^{-(1/2*0,2)*(x-1)^2}$
Máx. función coeficiente de difusión	k	0,4
Parámetro de la exponencial (positivo)	α	$\frac{\epsilon}{k^2}$
Paso de la simulación (stepsize)	Δ_t	10^{-8}

Cuadro 4.1: Parámetros utilizados para la simulación

Las función $b(x) = \epsilon + 10e^{-(1/2)*(x-1)^2}$ satisface las condiciones 4.2.1 lo mismo ocurre con la función $\sigma(x) = 0,4 - \frac{1}{\sqrt{\pi * 20}, 2} e^{-(1/2*0,2)*(x-1)^2}$.

En la figura 4.1 vemos los resultados, en ellos comparamos la distribución acumulada de los tiempos de saturación con las cotas exponenciales obtenidas en la sección anterior.

Podemos ver que la curva que minora la probabilidad de saturación a tiempo x, está a la derecha de los datos. Efectivamente la probabilidad de la exponencial es menor.

Cabe destacar que la cota exponencial es la trivial, es decir 0, mientras la misma toma valores negativos.

El gráfico se realizó con el siguiente código:

```
for ind1=1:1
clearvars -except cont tiemp_satura_stepsize ind1
muestras=1000;
tiemp_sat=zeros(1,muestras);
N = 10^8; dt = 1/N;
u=1; %umbral de interés

hold all
alfa_jl=-350;      %nivel de afinidad de los segmentos
a=1.05;            %constante para que el interes inicial
 %supere el umbral u
```

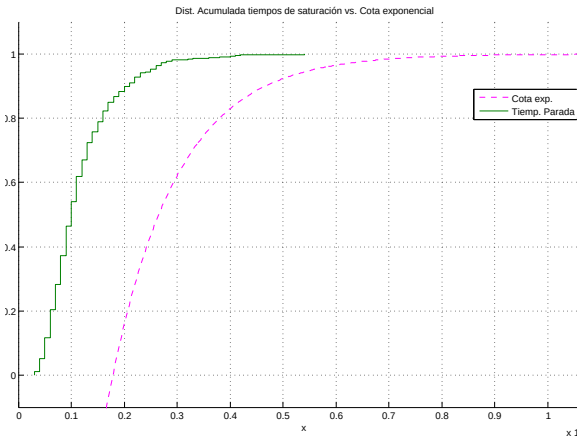

Figura 4.1: Gráfico de distribuciones acumuladas de la simulación vs la cota exponencial

```
 epsilon=1600.1;  eme1=10; %valores para definir b
 nu= 0.2; ka1=.2;       %valores para definir sigma

 A=(1-a)*u/alfa_jl;
 eme=epsilon+eme1; %cota sup de b
 ka=nu+ka1;        %cota sup de sigma


funcionf=@(x) epsilon+eme1*exp(-(1/2)*(x-1)^2) ;
funciong=@(x) ka-1/sqrt(pi*2)*nu*exp(-(1/2*nu)*(x-1)^2);


for ind=1:muestras
    ind
    clear i

eval(['Y',num2str(ind),'(1) = 0;'])           %inicio de la difusion
```

```
i=2;
while eval(['Y',num2str(ind),'(',num2str(i-1),')'])< A

 eval(['aux=Y',num2str(ind),'(',num2str(i-1),');'])

 clear tiempo

 dW = sqrt(dt)*randn(1,1)   ;
gevaluada=funciong(eval(['Y',num2str(ind),'...
                        ...(',num2str(i-1),');']));
eval(['Y',num2str(ind),'(',num2str(i),')=aux+...
                 ...funcionf(aux)*dt+gevaluada*dW;'])%...
%...+0.5*gevaluada*(gevaluada-funciong(Y(ind,i-1)-h))...
%...*(Y(ind,i-1)*(dW)^2-dt)/h;

clear aux

i=i+1;
if i==50000
    tiempo(1,ind)=0;
    fprintf('tiempo infinito alcanzado')
    break
end

tiempo=(i-1)*dt;

tiemp_sat(1,ind)=tiempo;
tamano(ind)=i;

end
 end
 figure(1)
 hold all
 end
 alfa=(epsilon)/(ka^2);
 const=exp(alfa*(A));
 minor=@(x) 1-exp(-(.5*alfa*epsilon)*(x)+alfa*A);
 h=ezplot( minor,[0,2*max(tiemp_sat)]);
```

```
 set(h, 'Color', 'm','Linestyle','--');

  hold all
figure(1)
cdfplot(tiemp_sat)

 xlim([-10,2*max(tiemp_sat)])
 ylim([-0.1,1.1])
```

4.6. Conclusiones

Hemos planteado un modelo sencillo para la red social Twitter que se basa en su alta volatilidad. A partir de él hemos podido determinar algunas probabilidades de saturación para distintas descripciones del modelo de emisión de tweets basados en modelos markovianos.

Hemos verificado estas estimaciones con simulaciones numéricas que permiten validar el modelo.

En el próximo capítulo presentaremos una técnica estadística conocida como cópulas que nos permitirá estudiar más adelante, el comportamiento de la probabilidad de saturación para varias fuentes de emisión de tweets dependientes.

Capítulo 5

Cópulas

5.1. Introducción

En este capítulo presentamos la técnica de cópulas que nos servirá para explorar el procedimiento de "mezcla" que se produce cuando se genera el (3.39). Ésta herramienta nos permitirá el estudio de los procesos de emisión de tweets en simultáneo sin hacer suposiciones de independencia, que no necesariamente son razonables.

5.2. Historia

La noción de cópula fue introducida por Abe Sklar in 1959 [37], intentando responder a una pregunta de Fréchet sobre la relación entre la función de densidad de una variable multidimensional y sus distribuciones marginales, [31]. De alguna manera trata de aportar luz sobre el problema de la dependencia de dos o más variables aleatorias trascendiendo la dicotomía independencia vs. no-independencia. Describiremos los conceptos de cópulas en dimensión dos para no cargar las notaciones, pero todos los resultados se pueden adaptar a d dimensiones.

Las cópulas se utilizan mucho en las finanzas, donde la dependencia de las variables es clave para estudiar las probabilidades de evolución de las variables analizadas, también se las utiliza en hidrología para estudiar la dependencia de las variables del clima (lluvia, vientos, marea) con eventos por ejemplo de inundación o crecidas.

En el contexto de las finanzas, Stefanova [39] se refiere a los riesgos de usar la correlación lineal para medir la dependencia. Estos riesgos surgen principalmente porque la correlación lineal sólo describe completamente los patrones de dependencia en la clase de distribuciones que son caracterizadas por la simetría. También es una herramienta inadecuada para distinguir la eventos extremos de baja probabilidad y alto costo. Otra de sus deficiencias proviene del hecho que los segundos momentos necesarios para calcularlas

deben ser finitos, sin mencionar que la correlación no es invariante bajo transformaciones no lineales estrictamente crecientes, que se sabe que no modifican la estructura de dependencia.

Como contraste de esto, se podrá ver que las medidas de dependencia se mantienen y sólo varían con la cópula. El principal concepto detrás de las cópulas es el de separar la estructura de la distribución conjunta de las marginales univariadas.

5.3. Definición

Definición 5.1 *Una cópula 2-dimensional es una función C que cumple las siguientes propiedades:*

1. $C : [0,1]^2 \to [0,1]$.
2. $C(0,v) = C(u,0) = 0 \, \forall u, v \in [0,1]$.
3. $C(1,v) = v$, $C(u,1) = u \, \forall u, v \in [0,1]$.
4. $\nabla^2 C([a,b] \times [c,d]) = C(b,d) - C(a,d) - C(b,c) + C(a,c) > 0$, *para todo* $a < b$, $c < d$.

Observación 5.2 *La condición 4 se cumple por ejemplo si C es $\mathcal{C}^2$ y* $\dfrac{\partial^2 C}{\partial u \partial v} \geq 0$.

La definición se puede extender fácilmente a d dimensiones. Ésta función cópula, indica como se mezclan las d distribuciones marginales para formar la distribución d-variada de un vector aleatorio de d dimensiones. En la cópula queda capturada toda la estructura de dependencia de las variables.

5.3.1. Existencia y construcción de Sklar

A continuación vamos a ver el primer resultado importante sobre cópulas perteneciente a Sklar [37]:

Teorema 5.3 (Teorema de Sklar) *Sea F una función de distribución d-dimensional con marginales continuas, $F_1, ..., F_d$. Entonces existe una única cópula C tal que:*

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)) \tag{5.1}$$

La hipótesis de continuidad de las distribuciones sólo es necesaria para la unicidad.

Inversa generalizada

Sea F es una distribución de probabilidad cualquiera, su inversa generalizada se define como:

$$F^{-1}(u) = inf\{t \in \mathbb{R} : F(t) \geq u\}, \text{ para todo } u \in (0,1). \tag{5.2}$$

Si F es una distribución estrictamente creciente, la inversa generalizada coincide con la función inversa habitual.

Con esta inversa generalizada podemos construir la cópula de la distribución bivariada de la siguiente manera:

$$C(u,v) = F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v)), \text{ cualquiera sean } u, v \in (0,1). \tag{5.3}$$

Una explicación de la función (5.3) es que para calcular la cópula en un punto (u,v), debo invertir cada una de esas coordenadas a partir de las marginales y encontrar las cantidades $F_X^{-1}(u)$ y $F_Y^{-1}(v)$ que luego se reemplazan en la distribución conjunta de las variables X e Y.

Este procedimiento de Sklar, permite que se puedan estimar las cópulas estimando por un lado las marginales y por otro la distribución conjunta.

Ejemplo 5.4 *El primer ejemplo corresponde a traducir al lenguaje de cópulas la independencia de las variables.*

1. *X e Y son independientes si y sólo si su cópula es $C(u,v) = C^{\perp}(u,v) = uv$, $\forall u, v \in [0,1]$.*

2. *Si $Y = X$ entonces su cópula es $C(u,v) = C^U(u,v) = min(u,v)$ para todo $u, v \in [0,1]$. Esta es la máxima "dependencia positiva".*

3. *Si $Y = -X$ y la distribución de X es simétrica respecto al 0 es decir $F(-x) = 1 - F(x)$. Entonces la cópula es $C(u,v) = C^L(u,v) = max(u+v-1, 0)$. Ésta es la máxima "dependencia negativa".*

En relación a los ejemplos anteriores, se puede probar el siguiente resultado:

Teorema 5.5 (Cotas de Fréchet-Hoefding) *Toda cópula verifica que*

$$C^L(u,v) = max(u+v-1, 0) \leq C(u,v) \leq min(u,v) = C^U(u,v). \tag{5.4}$$

Estas desigualdades se podrían explicar en los siguientes términos, la dependencia de dos variables X e Y, no pueden ser mayor que si fueran iguales, ni menor que si fueran opuestas.

5.4. Ejemplos y propiedades

5.4.1. Indicadores clásicos de dependencia

Las principales medidas conocidas de dependencia de dos variables tienen su expresión en términos de cópulas. Estos resultados los podemos encontrar por ejemplo en [29] y en [9].

τ de Kendall

El indicador τ de Kendall se puede definir como una resta de probabilidades, la probabilidad de concordancia menos la de discordancia. Sean (X_1, Y_1) y (X_2, Y_2) dos vectores aleatorios con funciones de distribución conjuntas $H1$ y $H2$ y posibles cópulas $C1$ y $C2$,

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{5.5}$$

Se demuestra que puede expresarse en términos de las cópulas como

$$\tau = 1 - 4 \int_0^1 \int_0^1 \frac{\partial C}{\partial u}(u, v) \frac{\partial C}{\partial v}(u, v) du dv. \tag{5.6}$$

ρ de Spearman

Sea R_i el rango de x_i entre las x y S_i el rango de y_i entre las y. El coeficiente de correlación de rangos de Spearman es:

$$\rho_S = \frac{\sum_{i=1}^n (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^n (R_i - \overline{R})^2} \sqrt{\sum_{i=1}^n (S_i - \overline{S})}}, \tag{5.7}$$

que en términos de cópulas queda como:

$$\rho = 12 \int_0^1 \int_0^1 C(u, v) du dv - 3. \tag{5.8}$$

Coeficiente de Gini

Otra medida de asociación fue introducida por Conrado Gini cerca de 1910, la llamó indice de cograduación simple.

Si p_i y q_i representan los rangos de una muestra de tamaño n de dos variables aleatorias X e Y respectivamente, entonces

$$g = \frac{1}{\lfloor n^2/2 \rfloor} \left(\sum_{i=1}^{n} |p_i + q_i - n - 1| - \sum_{i=0}^{n} |p_i - q_i| \right) \tag{5.9}$$

donde $\lfloor x \rfloor$ es la parte entera de x.

Si llamamos $U = f_X(x)$ y $V = f_Y(y)$ esta cantidad g es un estimador del parámetro $\gamma = 2E(|U + V - 1| - |U - V|)$, y en términos de la cópula C de X e Y se puede escribir:

$$\gamma = 4 \int_0^1 [C(u, 1-u) + C(u, u)] du$$

Coeficiente de correlación de medianas

Otra medida de asociación se puede construir si en la medida de concordancia de Kendall, en vez de utilizar 2 vectores, utilizamos un vector (X_1, Y_1) y un punto fijo (x_0, y_0), es decir:

$$\tau = P[(X_1 - x_0)(Y_1 - y_0) > 0] - P[(X_1 - x_0)(Y_1 - y_0) < 0]. \tag{5.10}$$

Blomqvist en 1950 estudió esta medida utilizando las medianas para x_0 e y_0. Esta medida la llamaremos β y está dada por

$$\beta = P[(X_1 - \tilde{x})(Y_1 - \tilde{y}) > 0] - P[(X_1 - \tilde{x})(Y_1 - \tilde{y}) < 0], \tag{5.11}$$

donde $\tilde{x}$ y $\tilde{y}$ son las medianas de X e Y respectivamente.

En función de las cópulas el coeficiente β se puede escribir:

$$\beta = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1. \tag{5.12}$$

5.4.2. Familia de Cópulas

Vamos a presentar aquí algunas familias paramétricas de cópulas cuyo uso es bastante difundido. La ventaja de estas familias es que a partir de cualquiera de ellas, podemos construir distribuciones multivariadas con las distribuciones marginales que deseemos. Si bien se pueden construir muchas familias de cópulas , sólo presentaremos las cópulas arquimedianas y la familia de cópulas elípticas.

En el capítulo 3 de [29] se puede profundizar más en este tema.

Cópulas arquimedeanas

Las cópulas arquimedeanas tienen aplicaciones muy variadas porque se pueden construir con facilidad. Originalmente estas cópulas no aparecieron en estadística, sino en el estudio de espacios métricos de probabilidad. Donde fueron estudiadas como parte del desarrollo de una versión probabilística de la desigualdad triangular. Se puede encontrar esta presentación en [34] y las referencias allí citadas.

Sean X e Y variables aleatorias continuas con distribución conjunta H y marginales F y G respectivamente.

Cuando X e Y son independientes, $H(x,y) = F(x)G(y)$, ésta es el único caso en el que la distribución conjunta se puede factorizar en el producto F y G.

Pero sí existen casos en los que una función H se factoriza como producto de funciones de F y de G, por ejemplo las distribuciones conjuntas y marginales de la familia de cópulas Ali-Mikhail-Haq satisfacen la siguiente relación

$$\frac{1-H(x,y)}{H(x,y)} = \frac{1-F(x)}{F(x)} + \frac{1-G(x)}{G(x)} + (1-\theta)\frac{1-F(x)}{F(x)}\frac{1-G(x)}{G(x)},$$

que se puede reescribir como:

$$1 + (1-\theta)\frac{1-H(x,y)}{H(x,y)} = \left[1 + (1-\theta)\frac{1-F(x)}{F(x)}\right]\left[1 + (1-\theta)\frac{1-G(y)}{G(y)}\right].$$

Esto se puede representar como $\lambda(H(x,y)) = \lambda(F(x))\lambda(G(y))$. Cada vez que podemos escribir $\lambda(H(x,y)) = \lambda(F(x))\lambda(G(y))$ para alguna función λ positiva en el intervalo $(0,1)$, si definimos $\Psi(t) = -\log(\lambda(t))$ también podemos escribir a H como una suma de función de las marginales F y G. Es decir que podemos escribir $\Psi(H(x,y)) = \Psi(F(x)) + \Psi(G(y))$, y si pensamos en cópulas,

$$\Psi(C(u,v)) = \Psi(u) + \Psi(v). \tag{5.13}$$

Las cópulas arquimedeanas surgen de buscar resolver (5.13) para alguna función Ψ y considerando eventualmente una inversa más general, es decir

$$C(u,v)) = \Psi^{[-1]}(\Psi(u) + \Psi(v)), \tag{5.14}$$

donde la notación $[-1]$ indica que la función inversa puede no ser la usual.

La función Ψ se denomina generador de la cópula. En el capítulo 4 de [29] se prueba lo siguiente

Proposición 5.6 *Si Ψ es convexa, positiva, estrictamente creciente y cumple $\Psi(1) = 0$ vale que (5.14) es una cópula si y sólo si Ψ es convexa.*

Además se prueban las siguientes propiedades de esta familia de cópulas

Teorema 5.7 *Sea C una cópula arquimedeana con generador Ψ entonces:*

1. *C es simétrica, es decir, $C(u,v) = C(v,u)$ para todo $u, v \in (0,1) = I$.*
2. *C es asociativa, es decir $C(C(u,v),w) = C(u,C(v,w))$, para todo $u, v, w \in I$.*
3. *Si $c > 0$ es constante, entonces $c\Psi$ también es un generador de de C*

Como vemos en la siguiente tabla según la función generadora Ψ, las cópulas arquimedeanas obtienen distintos nombres.

Nombre	$\Psi(t)$
Clayton	$(1+t)^{\frac{1}{\theta}}$
Frank	$-\theta^{-1}\log[1-(1-exp(-\theta))]exp(-t)$
Gumbel	$exp(-t^{\frac{1}{\theta}})$
Independencia	$exp(-t)$
Fue	$1-(1-exp(-t))^{\frac{1}{\theta}}$

Finalmente terminamos esta sección con la explicación del nombre que recibe esta familia de cópulas. Recordemos el axioma de Arquímides para números reales positivos: Si a y b son números reales positivos, existe un número entero n tal que $na > b$. Una cópula arquimedeana se comporta como una operación binaria en el intervalo $I = (0,1)$, donde a cada par (u,v) le asocia un número $C(u,v) \in I$. Del teorema (5.7) podemos decir que esta operación que define la cópula es conmutativa y asociativa y preserva el orden en cada coordenada, y así el par (I, C) es un semigrupo abeliano.

Si definimos para cualquier $u \in I$ las C-potencias de u recursivamente $u_C^1 = u$ y $u_C^{n+1} = C(u, u_c^n)$, se puede probar que el axioma de arquimides vale para esta operación. Es decir, para dos números cualquiera $u, v \in (0,1)$, existe un entero positivo n tal que $u_c^n < v$. Este es el motivo por el que esta familia de cópulas se denomina de esta forma. El término arquimediano para estas cópulas fue introducido por Ling en 1965, [27].

Familia de cópulas elípticas

Un vector aleatorio, $X = (X_1, ..., X_d)$ tiene una distribución elíptica con media $\mu \in \mathbb{R}^d$, matriz de covarianza $\Sigma = \sigma_{ij}$ y generador g, que notamos $X \sim \epsilon(\mu, \Sigma, g)$, si X puede ser expresado en la forma

$$X = \mu + RAU,$$

donde $AA^t = \Sigma$ es la descomposición de Cholesky de Σ, U es un vector aleatorio d-dimensional distribuido uniformemente en la esfera $S^{d-1} = \{u \in R^d : u_1^2 + ... + u_d^2 = 1\}$, y R es un vector aleatorio positivo, independiente de U cuya densidad está dada para todo $r > 0$ por

$$f_g(r) = \frac{\pi^{d/2}}{\Gamma(d/2)} r^{d-1} g(r^2).$$

Si existe la densidad de esta función está definida de la siguiente forma para todo $x \in \mathbb{R}^d$

$$h_g(x) = |\Sigma|^{1/2} g\left((x-\mu)^t \Sigma^{-1}(x-\mu)\right).$$

Para este tipo de distribuciones podemos definir cópula elíptica como sigue

Definición 5.8 *Sea X un vector aleatorio elíptico $\epsilon_d((\mu, \Sigma, g)$. Supongamos que para todo $i = 1, 2, ..., d$, $\frac{X_i}{\sqrt{\sigma_{ii}}} \sim F_g$. Llamamos cópula elíptica a la función de distribución del vector:*

$$\left(F_g\left(\frac{X_1}{\sqrt{\sigma_{11}}}\right), ..., F_g\left(\frac{X_d}{\sqrt{\sigma_{dd}}}\right)\right)$$

Veamos el ejemplo más importante de este tipo de cópulas:

Ejemplo 5.9 (Cópula Gaussiana) *La distrbución normal multivariada, pertenece a la familia de las distribuciones elípticas, y su expresión es:*

$$C(u,v) = \int_{-\infty}^{u}\int_{-\infty}^{v} \frac{1}{det(\Sigma)} exp\left\{\begin{pmatrix}\Phi^{-1}(x)\\ \Phi^{-1}(y)\end{pmatrix}^t \Sigma^{-1} \begin{pmatrix}\Phi^{-1}(x)\\ \Phi^{-1}(y)\end{pmatrix}\right\},$$

para una matriz de covarianza Σ.

El superíndice ^{t} indica la operación transposición de matrices.

5.4.3. Algunas propiedades de cópulas

Transformaciones monótonas

Una propiedad que resulta importante de las cópulas es que éstas son invariantes por transformaciones monótonas de las variables. Esta propiedad resulta de la construcción de Sklar y permite asegurar que los cambios de escala habituales en los análisis de datos no alteran la cópula. Más precisamente:

- Si f, g estrictamente crecientes $\Rightarrow C_{f(X),g(Y)} = C_{X,Y}$
- Si f estrictamente creciente, g estrictamente decreciente $\Rightarrow C_{f(X),g(Y)}(u,v) = C_{X,Y}(u, 1-v)$

- Si g estrictamente creciente, f estrictamente decreciente $\Rightarrow C_{f(X),g(Y)}(u,v) = C_{X,Y}(1-u,v)$
- Si f, g estrictamente decrecientes $\Rightarrow C_{f(X),g(Y)}(u,v) = C_{X,Y}(1-u,1-v)$

Relación entre la cópula y la densidad conjunta

Otra propiedad que también vale la pena mencionar es que si X, Y tiene densidad conjunta $f_{X,Y}$ entonces

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial u \partial v} C(F_X(x), F_Y(y)) f_X(x) f_Y(y).$$

5.5. Estimación de cópulas

A continuación vamos a presentar algunas técnicas de estimación de cópulas.

Las cópulas relacionan distintas funciones, por un lado las distribuciones marginales y por otro la distribución conjunta. Según [8], el primer tema es especificar como se estimarán por separado estas distribuciones. Incluso puede ser que alguna de estas funciones sea totalmente conocida.

Dependiendo de las hipótesis que impongamos sobre cada una de estas, algunas distribuciones podrán ser determinadas paramétricamente, semi-paramétricamente o no paramétricamente.

En el caso no paramétrico se debe elegir entre las metodologías usuales aplicadas a las funciones de distribución empíricas. Estos métodos de suavizado pueden ser por núcleos, wavelets, polinomios ortogonales, vecinos cercanos, etc.

La precisión de la estimación y de los resultados gráficos dependen de todas estas decisiones. Una distribución marginal correctamente asumida puede mejorar notablemente las estimaciones, pero un pequeño error en la hipótesis puede implicar grandes errores en los resultados.

En general, si no se dispone de información previa valiosa, se deben favorecer los métodos no paramétricos.

La siguiente situación nos permitirá mostrar parte de la complejidad que implica la estimación de cópulas.

Supongamos que tenemos un vector aleatorio (X_1, X_2), y que lo hemos observado en T oportunidades. Es decir que tenemos el conjunto de datos $((X_{11}, X_{21}), ..., (X_{1T}, X_{2T}))$. La distribución marginal de X y de Y puede ser estimada empíricamente por :

$$F_k^{(1)}(x) = \sum_{i=1}^{T} \mathbb{1}(X_{ki} \leq x), \tag{5.15}$$

en este caso, $k = 1, 2$.

Otra opción que uno puede tomar para estimar, es suavizar estas distribuciones empíricas utilizando el método de núcleos. (En la próxima sección hablaremos más de este método). A partir de este procedimiento obtenemos otra estimación para las distribuciones marginales $F_k(x)$, estas son

$$F_k^{(2)}(x) = \frac{1}{T}\sum_{i=1}^{T} \mathbb{K}\left(\frac{x - X_{ki}}{h}\right), \tag{5.16}$$

con $k = 1, 2$.

Por último, puede darse la situación en la que podemos asumir un modelo paramétrico subyacente dependiente de θ y que fue previamente ajustado a los datos de una o varias marginales. En este caso una estimación de las mismas sería para $k = 1, 2$

$$F_k^{(3)}(x) = F_k^{(3)}(x, \hat{\theta}_k). \tag{5.17}$$

Esta estimación depende del ajuste paramétrico previo, y más profundamente de que la hipótesis sobre la familia paramétrica sea correcta.

Hasta ahora tenemos 3 estimaciones posibles y diferentes de las marginales

$$F_k(x) = F_k^{(j)}(x),$$

para $j = 1, 2, 3$. El súper índice 1, indica la estimación empírica, el 2 la empírica suavizada por núcleos y el 3 la estimación paramétrica.

La misma situación se da para las diferentes estimaciones de la distribución conjunta:

1. Estimación empírica: $G^{(1)}(x_1, x_2) = \frac{1}{T}\sum_{i=1}^{T} \mathbb{1}((X_1, X_2) \leq (x_1, x_2))$
2. Estimación por núcleos: $G^{(2)}(x_1, x_2)\frac{1}{T}\sum_{i=1}^{T} \mathbb{K}\left(\frac{(x_1, x_2) - (X_1, X_2)}{h}\right)$
3. Estimación con una distribución paramétrica ajustada: $G^{(3)}(x_1, x_2) = G(., \hat{\tau})$

Debido a la construcción de sklar 5.3 una cópula bivariada se puede estimar por:

$$\hat{C}(u_1, u_2) = G^{(j)}\left([F_1^{(j_1)}(u_1)]^{-1}, [F_2^{(j_2)}(u_2)]^{-1}\right), \tag{5.18}$$

donde los índice j, j_1, j_2 pertenecen al conjunto $\{0, 1, 2\}$.

Cada combinación de índices determina corresponde a una forma de estimar, por ejemplo si todos los j_i son iguales a 1, quiere decir que se toman las versiones empíricas de las estimaciones.

La cantidad de decisiones que hay que tomar antes de empezar a estimar, hace que ésta no sea una tarea menor a la que hay que dedicarle tiempo, más aún si se tiene en cuenta las ventajas y desventajas que tiene cada forma de estimar.

Estos estimadores verifican propiedades usuales de los estimadores, básicamente son consistentes y asintóticamente normales, [8].

Ejemplo 5.10 *Sean las funciones empíricas definidas por las siguientes fórmulas:*

$$F_X^n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}},$$

y

$$F_Y^n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}},$$

si tomamos

$$C_n(u,v) = F_{X,Y}^n((F_X^n)^{-1}(u), (F_Y^n)^{-1}(v)),$$

entonces C_n es un estimador consistente de la cópula C de (X,Y). La demostración de este resultado se puede ver en [11].

5.6. Conclusiones

Hemos presentado las cópulas como una herramienta muy interesante para estudiar la dependencia de variables aleatorias. Presentamos además algunas propiedades y ejemplos y distintas maneras de estimarlas.

En el próximo capítulo analizaremos una manera de estudiar los procesos estocásticos desde el enfoque de cópulas a partir de una operación definida sobre cópulas o a partir de procesos empíricos.

Capítulo 6

Cópula para procesos estocásticos

6.1. Introducción

En el capítulo anterior se han presentado algunos resultados de la teoría de cópulas. A continuación presentaremos una relación entre las cópulas y los procesos estocásticos.

Plantearemos una manera de definir cópulas sobre procesos estocásticos y presentaremos algunos resultados sobre estimación de cópulas para procesos discretos.

Como en el modelo que hemos propuesto para Twitter los procesos son continuos, necesitamos extender de alguna manera la estimación para este tipo de procesos.

Para ello, es que generalizamos el estimador de cópulas por núcleos. Este estimador es el principal aporte original de esta tesis, y nos permitirá analizar en conjunto el comportamiento de procesos estocásticos que no son independientes.

6.2. ¿Cómo definir una cópula para procesos estocásticos?

Vamos a presentar una forma de introducir cópulas relacionadas con procesos estocásticos.

Dado un proceso $(X_t)_{t\in T}$ 1-dimensional, para todo $t_1 < t_2 < ... < t_k$ y cualquier $k \in \mathbb{N}$, podemos definir:

$$C_{t_1,...,t_k} = C_{(X_{t_1},...,X_{t_k})}.$$

Esta cópula describe la estructura de dependencia de las variables $X_{t_1}, ..., X_{t_k}$ también se la conoce como autocópula [32].

Para procesos d-dimensionales ($d \geq 2$), de la forma $(X_t^1, ..., X_t^d)$ las posibilidades para definir y construir cópulas se amplían. Podemos definirlas asociando a cada instante t la cópula correspondiente a la distribución del proceso, es decir:

$$C_t(u_1, ..., u_d) = F_{X_t^1,...,X_t^d}\left(F_{X_t^1}^{-1}(u_1), ..., F_{X_t^d}^{-1}(u_d)\right). \tag{6.1}$$

Esta cópula describe el comportamiento conjunto de las variables $(X_t^1, ..., X_t^d)$.

También, podríamos estudiar las relaciones en distintos instantes y distintas coordenadas por ejemplo

$$C_{t_1,...,t_d}(u_1, ..., u_d) = F_{X_{t_1}^1,...,X_{t_d}^d}\left(F_{X_{t_1}^1}^{-1}(u_1), ..., F_{X_{t_d}^d}^{-1}(u_d)\right). \tag{6.2}$$

Esta cópula describe la distribución conjunta del vector $(X_{t_1}^1, ..., X_{t_d}^d)$.

Para la siguiente cópula propuesta introduzcamos la siguiente notación. Al conjunto de variables aleatorias $(X_{t_1}^1, X_{t_1}^2, X_{t_1}^3)$ lo notamos con $(1, 2, 3; t_1)$. Más general, $(1, ..., i; t_1, ..., t_j)$ representa el conjunto de variables $X_{t_1}^1, ..., X_{t_1}^i, X_{t_2}^1, ..., X_{t_2}^i, ..., X_{t_j}^1, ..., X_{t_j}^i$. Es decir que los primeros números representan en que dimensión se miran las variables y los segundos en qué instantes de tiempo.

Ejemplo 6.1 *El conjunto* $(1, 2; t_1, t_5, t_6)(1; t_4)$ *está formado por las variables*

$$(X_{t_1}^1, X_{t_5}^1, X_{t_6}^1, X_{t_1}^2, X_{t_5}^2, X_{t_6}^2, X_{t_4}^1).$$

Utilizando la notación anterior, podríamos también definir cópulas asociadas a un proceso de la siguiente manera:

$$C_{(i_1,...,i_{l_1};t_{j_1},...,t_{j_{k_1}})...(i_{m_1},...,i_{m_l};t_{j_{n_1}},...,t_{j_{n_k}})} \tag{6.3}$$

Esta cópula describe la distribución conjunta de las α variables, donde $\alpha = l_1k_1 + l_2k_2 + ... + m_ln_k$ que representa el conjunto de índices.

Observación 6.2 *Las cópulas que se pueden construir dependen de los tiempos observados del proceso y de las dimensiones en las que se mira.*

Observación 6.3 *Para un proceso d-dimensional* $X = (X_t^1, ..., X_t^d)$ *un caso de particular interés de estudio es:*

$$C_{(\cdot,t)}(u_1, ..., u_d), \tag{6.4}$$

que describe la distribución conjunta del vector X *en cada instante* t*.*

Las cópulas permiten analizar la estructura de dependencia de los procesos estocásticos para distintas configuraciones de espacio de estados y tiempo. A continuación vamos a presentar una operación entre cópulas que aplicada a las autocópulas permite una interesante traducción de la estadística de cierta clase de procesos en términos de cópulas.

6.2.1. Producto de Cópulas

En [10] se estudia el tipo de estructura de dependencia denominada independencia condicional que satisfacen las variables aleatorias en un proceso de Markov. Para analizar esta dependencia introducen una operación binaria entre cópulas y analizan el comportamiento y las propiedades de ésta.

Definición 6.4 *Dadas 2 cópulas bivariadas C_1 y C_2, definimos su producto $C_1 * C_2$ del modo siguiente:*

$$C_1 * C_2(u, v) = \int_0^1 \frac{\partial}{\partial v} C_1(u, z) \frac{\partial}{\partial u} C_2(z, v) dz$$

Propiedades del producto

Si C_1, C_2, C son cópulas bivariadas cualesquiera. Vale lo siguiente:

- $C_1 * C_2$ es una cópula bivariada.
- $C^{\perp} * C = C * C^{\perp} = C^{\perp}$ ($C^{\perp}$ es el elemento neutro de esta operación).
- $C^{U} * C = C * C^{U} = C$ (C^{U} es el elemento identidad de esta operación).
- $C^{L} * C(x, y) = y - C(1 - x, y)\ \forall x, y$
  $C * C^{L}(x, y) = x - C(x, 1 - y)\ \forall x, y$
- En particular, el producto $*$ no es conmutativo.

Las cópulas C^L, C^U y $C^{\perp}$ son las mismas que en (5.4), que en términos de esta operación funcionan como elementos distinguidos.

También se prueban las siguientes propiedades [10]:

- Como operación binaria, el producto de cópulas es continuo a derecha e izquierda sobre combinaciones convexas.
- Si $A_n, B \in \mathcal{C}$ tal que $A_n \to A$ entonces $A_n * B \to A * B$ y $B * A_n \to B * A$.
- El producto $*$ es asociativo.

6.2.2. Cópulas para procesos de Markov

A partir del producto de cópulas se puede probar y demostrar en el siguiente teorema que se encuentra en [10] que se puede expresar con cópulas la condición de markovianidad de un proceso.

Teorema 6.5 *Un proceso estocástico $(X_t)_{t\in T}$ es un proceso de Markov si y sólo si para todos los enteros n todos los números reales $t_1, ..., t_n \in T$ que satisfacen $t_k < t_{k+1}$, para $k = 1, ..., n-1$,*

$$C_{t_l,...,t_n} = C_{t_l t_2} * C_{t_2 t_3} * ... * C_{t_{n-1} t_n}, \tag{6.5}$$

donde $C_{t_l...t_n}$ es la cópula de $X_{t_1}, ..., X_{t_n}$ y $C_{t_k t_{k+1}}$ es la cópula de X_{t_k} y $X_{t_{k+1}}$.

Esta ecuación es la equivalente a la ecuación de Chapman-Kolmogorov en términos de cópulas.

Naturalmente de aquí se visualiza que la estadística de procesos de Markov puede realizarse en base a sus cópulas.

6.3. Estimaciones no paramétricas de cópulas de procesos

En esta sección vamos a estimar las cópulas de procesos estocásticos. Primero presentaremos resultados pensados para series de tiempo y luego siguiendo estos pasos demostraremos que es posible generalizar a procesos de difusión con ciertas características.

6.3.1. Procesos empíricos suavizados por núcleos

Cuando trabajamos en estimaciones no paramétricas, los procesos empíricos asociados al proceso representan una de las herramientas principales para su estudio. A continuación introducimos los procesos empíricos en el contexto de series de tiempo para luego analizar algunas de sus propiedades.

Sea $(X_1, Y_1), ..., (X_t, Y_t), t \in \mathbb{N}$ una trayectoria de una serie de tiempo. Si interpretamos esta trayectoria como una muestra de un vector (X, Y), la distribución empírica bivariada es la que presentamos en (5.15):

$$F^n_{X,Y}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x, Y_i \leq y\}}.$$

Pero esta distribución marginal poco tiene que ver con lo que queremos estudiar de la serie de tiempo. Sólo en el caso que la serie fuese independiente e identicamente distribuída podríamos sacar alguna conclusión relevante. Lo mismo ocurre con las otras funciones de distribución empíricas vistas en (5.5).

Para construir nuestros estimadores debemos utilizar funciones llamadas núcleos.

Definición 6.6 *Un núcleo es una función real acotada $k(x)$ que verifica las siguientes propiedades:*

1. $k(x) \geq 0$
2. $\int_{-\infty}^{\infty} k(x) = 1$
3. $k(x) = k(-x)$
4. $k(x) = 0 \, si \, |x| > M$
5. $k \in \mathcal{C}^{\infty}$

Observemos que las condiciones 1 a 3 se pueden resumir diciendo que $k(x)$ es una densidad simétrica de soporte compacto.

En función de los núcleos k_1 y k_2 construimos un núcleo 2-dimensional

$$k(x, y) = k_1(x)k_2(y),$$

y notemos a su primitiva como $K(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} k(s, t)dsdt = K_1(x)K_2(y)$. Donde $K_i(x) = \int_{-\infty}^{x} k_i(s)ds$, $i = 1, 2$.

Necesitamos introducir un parámetro que controle la ventana de la estimación. Para ello definimos las funciones $h = (h_1^N, h_2^N)$, con $h_i^N > 0$ $i = 1, 2$ que son positivas y tales que $h_i^N \to 0$ cuando $N \to \infty$. Y notamos h_*^N el máximo de los h_i^N.

Teniendo en cuenta estas notaciones, llamaremos $k(x, y; h_1, h_2) = k_1\left(\frac{x}{h_1}\right) k_2\left(\frac{y}{h_2}\right)$

$K(x, y; h_1, h_2) = K\left(\frac{x}{h_1}, \frac{y}{h_2}\right)$.

Con éstos núcleos y estas notaciones, las estimaciones empíricas de las distribuciones marginales y conjunta de la serie de tiempo son,

$$\hat{f}_X(x) = (Nh_1^N)^{-1} \sum_{i=1}^{N} k_1\left(\frac{x - X_i}{h_1^N}\right), \tag{6.6}$$

$$\hat{f}_Y(y) = (Nh_2^N)^{-1} \sum_{i=1}^{N} k_2\left(\frac{y - Y_i}{h_2^N}\right), \tag{6.7}$$

$$\hat{f}_{X,Y}(x, y) = (Nh_1^N h_2^N)^{-1} \sum_{i=1}^{N} k(x - X_i, y - Y_i; h_1^N, h_2^N). \tag{6.8}$$

Para hallar las funciones de distribución estimadas, integramos las densidades (6.6),(6.7),(6.8) y obtenemos:

$$\hat{F}_X(x) = \int_{-\infty}^{x} \hat{f}_X(s)ds,$$

$$\hat{F}_Y(y) = \int_{-\infty}^{y} \hat{f}_Y(s)ds,$$

$$\hat{F}_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \hat{f}_{X,Y}(s,t)dsdt.$$

Luego, siguiendo la construcción de Sklar, el estimador para la cópula es:

$$\hat{C}(u,v) = \hat{F}_{X,Y}(\hat{F}_X^{-1}(u), \hat{F}_Y^{-1}(v)). \tag{6.9}$$

El siguiente teorema, describe el comportamiento asintótico de este estimador bajo ciertas hipótesis sobre la serie de tiempo estudiada, el núcleo utilizado y sobre el comportamiento asintótico del tamaño de la ventana de estimación h_j^N.

Este teorema se encuentra en [16] y vamos a seguir la manera de exponerlo allí.

Comenzamos con las hipótesis referidas al núcleo y a las funciones h_j

Hipótesis 6.7 (Núcleo y ventana)

1. *El tamaño de la ventana es tal que* $N(h_*^N)^2 \to 0$,

1'. *El tamaño de la ventana verifica que* $N(h_*^N)^4 \to 0$ *y el núcleo* k *es par,*

2. *El núcleo* k *tiene soporte compacto.*

Hipótesis 6.8 (Proceso)

1. *El proceso* (X_n, Y_n) *verifica la condición de mixing fuerte* (3.2) *con* α_n *tal que* $\alpha_T = o(T^{-a})$ *para algún* $a > 1$, *cuando* $T \to \infty$.

2. *Las funciones de distribución marginal* F_j, $j = 1, 2$ *son continuamente diferenciables sobre los intervalos* $[F_j^{-1}(a) - \epsilon, F_j^{-1}(b) + \epsilon]$ *para todo* $0 < a < b < 1$ *y algún* $\epsilon > 0$, *con derivada positiva* f_j. *Más aún, la primera derivada parcial de* F *existe y es continua Lipschitz en el producto de estos intervalos.*

Teorema 6.9 (Comportamiento asintótico de la cópula) *Si se cumplen las condiciones (6.7) y (6.8), el proceso*

$$\sqrt{N}(\hat{C} - c), \tag{6.10}$$

tiende débilmente a un proceso gausiano centrado $\phi'(\mathrm{G})$ *en* $l^\infty([0,1]^n)$ *con la norma del supremo, donde el proceso límite está dado por:*

$$\phi'(\mathrm{G})(u_1,u_2) = \mathrm{G}(F_1^{-1}(u_1), F_2^{-1}(u_n)) \\ - \sum_{i=1}^{2} \frac{\partial C}{\partial u_i}(F_1^{-1}(u_1), F_2^{-1}(u_n))\mathrm{G}(+\infty, F_j^{-1}(u_j)). \tag{6.11}$$

Algunos corolarios de este teorema que también se mencionan en [10], aseguran la distribución conjunta normal de los estimadores de la cópula. Y con alguna hipótesis extra sobre las distribuciones bivariadas del proceso también se puede asegurar la distribución normal conjunta de los estimadores de las derivadas primeras de la cópula.

6.4. Cópula para difusiones d-dimensionales

En esta sección vamos a demostrar un resultado similar al teorema (6.9) para procesos estocásticos de difusión. Necesitamos un resultado de esta naturaleza para poder analizar el comportamiento del proceso d-dimensional que modela la generación de tweets que percibe un perfil. Para poder hacer análisis a partir de la versatilidad que tienen las cópulas para estudiar la dependencia del proceso.

Un resultado de esta naturaleza todavía no se ha desarrollado por eso es que aquí vamos a probarlo para luego plantear alguna forma de utilizarlo. Éste es el principal aporte de esta tesis a la estadística de procesos estocásticos.

Sea $(X_t^1, ...X_t^d)$ un proceso estocástico de difusión en d dimensiones. Vamos a considerar los procesos de tiempo de ocupación

$$\mu_t^{X^1,...,X^d}(x_1,...,x_d) = \frac{1}{t}\lambda\left\{s \in [0,t] :\ X_s^i \leq x_i\ , i \in \{1,...,d\}\right\}. \tag{6.12}$$

Si reemplazamos por ∞ alguna o todas las coordenadas de (6.12) vemos que

$$\mu_t^{X^1,...,X^d}(+\infty,...,x_i,...,+\infty) = \frac{1}{t}\lambda\{s \in [0,t] :\ X_s^i \leq x_i\ X_s^j \leq +\infty, \forall j \neq i\}.$$

Pero como $X_s^i \leq +\infty$ para todo s, resulta que

$$\mu_t^{X^1,...,X^d}(+\infty,...,x_i,...,+\infty) = \mu_t^{X^i}(x_i). \tag{6.13}$$

Verifiquemos las propiedades que cumple (6.12)

- Es no negativa, ya que está definida a partir de una medida.

- Es no decreciente ya que $\{X_s^i \leq x_i, s \in [0,t]\} \subseteq \{X_s^i \leq x_i', s \in [0,t]\}$ siempre que $x_i \leq x_i'$.
- Es continua a derecha.
- Para todo $i = 1, ..., d$

$$\lim_{x_i \to +\infty} \mu_t^{X^1,...,X^d}(x_1, ..., x_i, .., x_d) = 1.$$

- Para todo $i = 1, ..., d$

$$\lim_{x_i \to -\infty} \mu_t^{X^1,...,X^d}(x_1, ..., x_i, .., x_d) = 0.$$

Entonces el tiempo de ocupación (6.12) se puede pensar como una función de distribución conjunta para las variables $\mu_t^{X^i}(x_i)$.

Luego, por el teorema de Sklar (5.3) existe una cópula $\mathcal{C}^t$ tal que

$$\mu_t^{X^1,...,X^d}(x_1, ..., x_i, .., x_d) = \mathcal{C}^t(\mu_t^{X^1}(x_1), ..., \mu_t^{X^d}(x_d)). \tag{6.14}$$

Estimaremos esta cópula a partir de las trayectorias de $(X_t^1, ...X_t^d)$ y describiremos la distribución conjunta del proceso $(X_t^1, ..., X_t^d)$.

Es decir, a partir de la cópula $\mathcal{C}^t$ obtenida a partir de una trayectoria del proceso, aportaremos información de la función de distribución del mismo y de su respectiva cópula. En otras palabras, describiremos:

$$F_{X_t^1,...,X_t^d}(x_1, ..., x_d) = \mathcal{V}^t(F_{X_t^1}(x_1), ..., F_{X_t^d}(x_d)). \tag{6.15}$$

Empecemos calculando la esperanza de $\mu_t^{X^1,...,X^d}(x_1, ..., x_d)$

$$\begin{aligned} E(\mu_t^{X^1,...,X^d}(x_1, ..., x_d)) =& \frac{1}{t} E\left(\lambda\{s \in [0,t] : \ X_s^i \leq x_i \ , i \in \{1, ..., d\}\}\right) \\ =& \frac{1}{t} E(\int_0^t \mathbb{1}_{\{X_s^i \leq x_i, \ i=1,...,d\}} ds). \end{aligned}$$

Puedo aplicar el teorema de Fubini e intercambiar la esperanza con la integral:

$$\begin{aligned} \frac{1}{t}\int_0^t E(\mathbb{1}_{\{X_s^i \leq x_i, \ i=1,...,d\}}) ds &= \frac{1}{t}\int_0^t P(X_s^1 \leq x_1, ..., X_s^d \leq x_d) ds \\ &= \frac{1}{t}\int_0^t F_{X_s^1,...,X_s^d}(x_1, ..., x_d) ds \\ &= \frac{1}{t}\int_0^t \mathcal{V}^s(F_{X_s^1}(x_i), ..., F_{X_s^d}(x_d)) ds. \end{aligned}$$

Esto ocurre para cualquier difusión y su proceso de tiempo de ocupación. Es decir que la esperanza del tiempo de ocupación $\mu_t^{X^1,...,X^d}(x_1,...,x_d)$ es igual a la integral de la distribución conjunta $F_{X_t^1,...,X_t^d}(x_1,...,x_d)$ del proceso original.

Si le pedimos al proceso $(X_s^1,...,X_s^d)$ que sea estacionario, entonces la distribución $F_{X_s^1,...,X_s^d}(x_1,...,x_d)$ no depende de s y tampoco las $F_{X_s^i}(x_i)$. Una forma de interpretar esto es que la distribución del proceso mezcla siempre las mismas $F_{X_s^i}(x_i)$, de la misma manera según $F_{X_t^1,...,X_t^d}(x_1,...,x_d)$,

$$E(\mu_t^{X^1,...,X^d}(x_1,...,x_d)) = \frac{1}{t}\int_0^t \mathcal{V}^s(F_{X_s^1}(x_i),...,F_{X_s^d}(x_d))ds \tag{6.16}$$

Como

$$\begin{aligned}\mathcal{V}^s(F_{X_s^1}(x),...,F_{X_s^d}(x_d)) &= F_{X_s^1,...,X_s^d}(x_1,...,x_d)\\ &= F_{X_{s+h}^1,...,X_{s+h}^d}(x_1,...,x_d)\\ &= \mathcal{V}^{s+h}(F_{X_s^1}(x_1),...,F_{X_s^d}(x_d))\\ &= \mathcal{V}^{s+h}(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d)),\end{aligned} \tag{6.17}$$

resulta que $\mathcal{V}^s = \mathcal{V}^{s+h}$ para todo $h > 0$. Luego $\mathcal{V}^s = \mathcal{V}^0$.

Finalmente retomando en (6.16) tenemos que

$$\begin{aligned}\frac{1}{t}\int_0^t \mathcal{V}^s(F_{X_s^1}(x_1),...,F_{X_s^d}(x_d))ds &= \frac{1}{t}\int_0^t \mathcal{V}^0(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d))ds\\ &= \mathcal{V}^0(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d))\end{aligned} \tag{6.18}$$

Resumiendo obtenemos que

$$E(\mu_t^{X^1,...,X^d}(x_1,...,x_d)) = \mathcal{V}^0(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d)). \tag{6.19}$$

Si tenemos en cuenta (6.14) queda probado que:

$$\begin{aligned}E(\mu_t^{X^1,...,X^d}(x_1,...,x_d)) &= E(\mathcal{C}^t(\mu_t^{X^1}(x_1),...,\mu_t^{X^d}(x_d)))\\ &= \mathcal{V}^0(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d)).\end{aligned} \tag{6.20}$$

Vamos a analizar ahora el comportamiento de $\mu_t^{X^1,...,X^d}(x_1,...,x_d)$ cuando $t \to \infty$, más específicamente vamos a estudiar el comportamiento del límite $\lim_{t\to\infty} \mathcal{C}^t$.

$$\begin{aligned}\lim_{t\to\infty} \mu_t^{X^1,...,X^d}(x_1,...,x_d) &= \lim_{t\to\infty} \frac{1}{t}\lambda\{s\in[0,t] : X_s^i \le x_i,\ \ i=1,...,d\}\\ &= \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}_{\{(X_s^1,...,X_s^d)\in(-\infty,x_1]\times...\times(-\infty,x_d]\}}(s)ds\end{aligned} \tag{6.21}$$

Notaremos $A_s(x_1,...,x_d) = \left\{(X_s^1,...,X_s^d) \in (-\infty,x_1] \times ... \times (-\infty,x_d]\right\}$, de esta manera la expresión (6.21) se transforma en:

$$\lim_{t\to\infty} \mu_t^{X^1,...,X^d}(x_1,...,x_d) = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}_{A_s(x_1,...,x_d)} ds. \tag{6.22}$$

Si llamamos $f_s(x_1,...,x_d) = \mathbb{1}_{A_s}(x_1,...,x_d)$, la función f_s es Borel medible y acotada.

Lo que estamos tratando de realizar es expresar el límite (6.22) en los términos de un teorema ergódico lo que resulta:

$$\lim_{t\to\infty} \mu_t^{X^1,...,X^d}(x_1,...,x_d) = \lim_{t\to\infty} \frac{1}{t}\int_0^t f_s(x_1,...,x_d) ds. \tag{6.23}$$

Si además tenemos un proceso ergódico, entonces vale que

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f_s(x_1,...,x_d) ds \underset{t\to+\infty}{\longrightarrow} E(f(u_1,...,u_d)), \quad c.s. \tag{6.24}$$

Calculamos esta esperanza,

$$\begin{aligned} E(f(u_1,...,u_d)) &= E\left(\mathbb{1}_{\{(X_s^1,...,X_s^d)\in(-\infty,x_1]\times...\times(-\infty,x_d]\}}(s)\right) \\ &= F_{X_0^1,...,X_0^d}(x_1,...,x_d) \\ &= \mathcal{V}^0(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d). \end{aligned} \tag{6.25}$$

Finalmente concluimos que:

$$\lim_{t\to\infty} C^t\left(\mu_t^{X^1}(x_1),...,\mu_t^{X^d}(x_d)\right) = \mathcal{V}^0\left(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d)\right). \tag{6.26}$$

Si además el proceso $(X_t^1,...,X_t^d)$ cumple condiciones de mixing, entonces $(\mu_t^{X^1,...,X^d})$ también cumple condiciones de mixing y entonces vale un teorema central del límite

$$\lim_{t\to\infty} \sqrt{t}\left(\mathcal{C}^t(x_1,...,x_d) - \mathcal{V}^0(x_1,...,x_d)\right) \to \mathcal{N}(0,1) \tag{6.27}$$

Podemos resumir el contenido de esta sección en los teoremas siguientes:

Teorema 6.10 *Sea $(X_t^1,...,X_t^d)$ un proceso de difusiones d-dimensional. Llamamando*

$$\nu^t(F_{X_t^1}(x_1),...,F_{X_t^d}(x_d)) = F_{X_t^1,...,X_t^d}(x_1,...,x_d)$$

a la cópula que describe la función de distribución del proceso para cada tiempo. Sea $\mu_t^{X^1,...,X^d}(x_1,...,x_d)$ *el proceso de tiempos de ocupación de la difusión y llamemos*

$$C^t(\mu_{X_t^1}(x_1),...,\mu_{X_t^d}(x_d)) = \mu_t^{X^1,...,X^d}(x_1,...,x_d)$$

a la cópula que mezcla las marginales del proceso μ*. Si el proceso cumple las siguientes hipótesis:*

H1) Si $(X_t^1,...,X_t^d)$ *es un proceso estacionario.*

H2) Si $(X_t^1,...,X_t^d)$ *es un proceso ergódico. Es decir que para toda función* f *que sea medible, se verifica el siguiente límite:*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(X_s)ds = E(f(X))$$

Vale que

$$\lim_{t\to\infty} C^t\left(\mu_t^{X^1}(x_1),...,\mu_t^{X^d}(x_d)\right) = \mathcal{V}^0\left(F_{X_0^1}(x_1),...,F_{X_0^d}(x_d)\right). \tag{6.28}$$

Si además se cumple la hipótesis

H3) El proceso es α*-mixing.*

Tenemos que:

$$\lim_{t\to\infty} \sqrt{t}\left(\mathcal{C}^t(x_1,...,x_d) - \mathcal{V}^0(x_1,...,x_d)\right) \to \mathcal{N}(0,1) \tag{6.29}$$

6.5. Conclusiones

Hemos presentado en este capítulo una nueva forma de estimar la cópula para una proceso d-dimensional markoviano continuo. Si consideramos el proceso $(X_t^1,...,X_t^d)$, la cópula del proceso es $\nu^t(F_{X_t^1}(x_1),...,F_{X_t^d}(x_d))$ que nos da permite relacionar las distribuciones marginales del proceso con su distribución conjunta en cada tiempo.

El estimador para esta cópula es el proceso empírico de ocupación $\mu_t^{X^1,...,X^d}(x_1,...,x_d)$.

Si el proceso cumple las hipótesis de los teoremas 6.10 entonces el estimador efectivamente estima $\nu^t(F_{X_t^1}(x_1),...,F_{X_t^d}(x_d))$, que en este caso no tiene dependencia con el tiempo.

Retomaremos esta herramienta en el capítulo siguiente para obtener una estimación de la probabilidad de saturación que involucre la cópula y la estructura de dependencia del proceso.

Capítulo 7

Twitter III: Evolución del Interés Global

7.1. Introducción

En este capítulo vamos a utilizar el resultado principal (6.27) para profundizar en la descripción que hemos hecho de Twitter.

La importancia de encontrar la cópula que regula la dependencia del proceso estocástico $(X_t^1, ..., X_t^d)$ radica en que nos permitirá obtener conclusiones sobre la evolución asintótica del interés del usuario.

Además plantearemos algunos otros interrogantes que se pueden responder a partir de conocer la cópula que rige el proceso.

7.2. Interés global

Consideremos un estado estable de la red social Twitter, en el que el comportamiento de todos los emisores se encuentra en régimen ergódico y estacionario. Además asumimos que la dependencia del proceso es mixing fuerte. Esta hipótesis es razonable ya que lo que un perfil twiteó hace mucho tiempo pierde influencia en las emisiones actuales y además el perfil que lee también va perdiendo memoria sobre publicaciones antiguas.

Consideremos una segmentación $S_1, ..., S_d$ de perfiles de twitter y un usuario $p \in S_l$ $1 \leq l \leq d$. Sea $\mathcal{L}(p) = \{p'_1, ..., p'_r\}$ el conjunto de líderes de p donde $p'_1, ..., p'_{i_1} \in S_1$, $p'_{i_1+1}, ..., p'_{i_1+i_2} \in S_2$, ... ,$p'_{i_1+i_2+...+i_{d-1}}, ..., p'_{i_1+i_2+...+i_d} \in S_d$, donde $i_1 + i_2 + ... + i_d = r$.

Como vimos en el (4) vamos a modelar las emisiones de tweets de los perfiles del segmento S_i por un proceso de difusión estacionario ergódico mixing fuerte X_t^i, de manera que el proceso d-dimensional $(X_t^1, ..., X_t^d)$ indica la cantidad de tweets que p recibe de cada segmento en cada instante t.

Llamemos $F_{X_t^i}(x_i)$ a la función de distribución de la variable X_t^i y $\mathcal{V}_{F_{X_t^1},...,F_{X_t^d}}(x_1,...,x_d)$ a la cópula que modela la dependencia del proceso $(X_t^1,...,X_t^d)$ que estimamos por el teorema (6.27).

Sean $\alpha_1,...,\alpha_d$ los valores de afinidad que tiene p respecto de los distintos segmentos. Y supongamos que los primeros k coeficientes $\alpha_1,...,\alpha_k$ son positivos y los restantes $\alpha_{(k+1)},...,\alpha_d$ son negativos. Conseguir esto siempre es posible mediante un reordenamiento apropiado de $\mathcal{L}(p)$. Si hubiera un $\alpha_h = 0$ descartamos ese proceso porque no aporta al interés. Vamos a definir el proceso de interés global de la siguiente forma,

$$I_t^g(p) = I_0^g + (X_t^1,...,X_t^d) \times \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix}. \tag{7.1}$$

Otra manera de expresar a I_t^g es

$$I_t^g(p) = I_0^g + \sum_{i=1}^{d} \alpha_i X_t^i, \tag{7.2}$$

dado un umbral de interés u estudiaremos la probabilidad de que el interés tome un valor menor que el del umbral

$$P(I_t^g < u).$$

Vamos a mostrar como se puede aprovechar el conocimiento de la cópula de $(X_t^1,...,X_t^d)$ para analizar el proceso de interés I_t^g.

Observemos que por la propiedad de cópulas vista en (5.4.3) vale que:

$$\begin{aligned}\mathcal{V}_{F_{\alpha_1 X_t^1},...,F_{\alpha_k X_t^k},F_{\alpha_{(k+1)} X_t^{k+1}},...,F_{\alpha_d X_t^d}}(x_1,...,x_k,x_{k+1},...,x_d) = \\ = \mathcal{V}_{F_{X_t^1},...,F_{X_t^d}}(x_1,...,x_k,1-x_{k+1},...,1-x_d)\end{aligned} \tag{7.3}$$

Como yo conozco la cópula para el proceso por el resultado del capítulo anterior, puedo sacar como función aleatoria de un vector la distribución de la sumatoria $\sum_{i=1}^{d} \alpha_i X_t^i$.

7.2.1. Probabilidad global de saturación

Supongamos que la cópula estimada por (6.27) del proceso $(X_t^1,...,X_t^d)$ es $\hat{C}(u_1,...,u_d)$ y que ésta cópula es diferenciable dos veces. La función de distribución del proceso se puede expresar en términos de la cópula y las distribuciones marginales de cada X^i es decir:

$$F_{X^1,...,X^d}(x_1,...,x_d) = \hat{C}(\hat{F}_{X^1}(x_1),...,\hat{F}_{X^{k+1}}(x_{k+1}),...,\hat{F}_{X^d}(x_d)).$$

Para las variables $\tilde{X}^i = \alpha_i X^i$ sus distribuciones son $\hat{F}_{\tilde{X}^i}(x_i) = \hat{F}_{X^i}(\frac{x_i}{\alpha_i})$ y luego si $\alpha_i > 0$ para $i = 1, ..., k$ y $\alpha_j < 0$ para $j = k+1, ..., d$, la función de distribución de todo el proceso es:

$$\hat{F}_{\tilde{X}^1,...,\tilde{X}^d}(x_1, ..., x_d) = \hat{C}\left(\hat{F}_{X^1}(\frac{x_1}{\alpha_1}), ..., \hat{F}_{X^k}(\frac{x_k}{\alpha_k}), 1 - \hat{F}_{X^{k+1}}(\frac{x_{k+1}}{\alpha_{k+1}}), ..., 1 - \hat{F}_{X^d}(\frac{x_d}{\alpha_d})\right). \tag{7.4}$$

Además si derivamos podemos encontrar la función de densidad de esta distribución:

$$\begin{aligned}
&\hat{f}_{\tilde{X}^1,...,\tilde{X}^d}(x_1, ..., x_d) = \\
&= \frac{\partial \hat{C}}{\partial x_1 ... \partial x_d}\left(\hat{F}_{X^1}\left(\frac{x_1}{\alpha_1}\right), ..., 1 - \hat{F}_{X^d}\left(\frac{x_d}{\alpha_d}\right)\right) \times \hat{f}_1\left(\frac{x_1}{\alpha_1}\right) ... \left(-\hat{f}_d\left(\frac{x_d}{\alpha_d}\right)\right) \frac{1}{\alpha_1 ... \alpha_d} \\
&= \frac{\partial \hat{C}}{\partial x_1 ... \partial x_d}\left(\hat{F}_{X^1}\left(\frac{x_1}{\alpha_1}\right), ..., 1 - \hat{F}_{X^d}\left(\frac{x_d}{\alpha_d}\right)\right) \times \hat{f}_1\left(\frac{x_1}{\alpha_1}\right) ... \left(\hat{f}_d\left(\frac{x_d}{\alpha_d}\right)\right) \frac{1}{\alpha_1 .. |\alpha_{k+1}| ... |\alpha_d|}.
\end{aligned} \tag{7.5}$$

Supongamos que la cantidad de interés inicial I_0 es determinística de manera que no interviene en el cálculo de la probabilidad de saturación.

Finalmente la expresión que queda para calcular la probabilidad de saturación global es:

$$P(I_g < u) = \int_{\tilde{X}^1 + ... + \tilde{X}^d < u} \cdots \int \hat{f}_{\tilde{X}^1,...,\tilde{X}^d}(x_1, ..., x_d) dx_1 ... dx_d. \tag{7.6}$$

7.2.2. Afinidades mínimas para saturación

A partir del conocimiento que nos da el teorema del capítulo anterior (6.27), podemos calcular la distribución de algunos procesos relacionados con el interés.

Para simplificar los cálculos consideremos el proceso $\tilde{I}^g_t$ de la siguiente forma. Sean $\alpha_+ = \text{mín}\{\alpha_1, ..., \alpha_k\}$ y $\alpha_- = \text{máx}\{|\alpha_{(k+1)}|, ..., |\alpha_d|\}$. Reemplazaremos los α_i con estos números según su signo.

Llamaremos

$$\tilde{I}^g_t = I^g_0 + \alpha_+ \sum_{i=0}^{k} X^i_t + \alpha_- \sum_{i=k+1}^{d} (-X^i_t).$$

Claramente $I^g_t \geq \tilde{I}^g_t$ y si escribimos

$$\tilde{T}^g = \text{ínf}\{t > 0 : \tilde{I}^g_t < u\},$$

entonces $\tilde{T}^g < T^g$.

Con la información de la cópula y las distribuciones marginales, puedo encontrar la distribución de la variable

$$\Sigma = \sum_{i=1}^{d} \alpha_i X_t^i,$$

o de las variables

$$P = \alpha_+ \sum_{i=0}^{k} X_t^i, \tag{7.7}$$

y

$$N = \alpha_- \sum_{i=k+1}^{d} X_t^i, \tag{7.8}$$

involucradas en el $\tilde{I}_t^g$.

El objetivo es, a partir de esto, encontrar alguna condición para las cantidades α_+ y el α_- de manera que hagan mínima la probabilidad de saturación global.

Si consideramos conocidas a las variables (7.7) y (7.8) podemos preguntarnos qué relación tienen que cumplir los pesos de la valoración positiva y negativa para no saturación.

Vemos que

$$\tilde{I}^g = \alpha_+ P + \alpha_- N,$$

donde el signo del proceso negativo se lo asignamos al proceso N en vez de al coeficiente.

Supongamos que

$$\alpha_+ + \alpha_- = 1, \tag{7.9}$$

Esta condición implica que podemos interpretar las valoraciones de interés como pesos relativos, y que las afinidades que tiene el perfil por los distintos segmentos se reparten por completo. También se pueden interpretar a los coeficientes α_i como la atención posotiva o negativa, de acuerdo al signo, sobre un perfil. Además la condición (7.9) indica que se reparte toda la atención entre lo positivo y lo negativo.

Con esta restricción el proceso de interés queda

$$\tilde{I}^g = \alpha P + (1 - \alpha) N.$$

Quitaremos el subíndice + del coeficiente α para no cargar la notación sin necesidad y omitiremos el superíndice g del proceso $\tilde{I}^g$ por la misma razón.

Dada una probabilidad $0 < \beta < 1$ y un umbral $u > 0$, vamos a estudiar la probabilidad de que se produzca la saturación

$$P(\tilde{I} < u) \leq \beta. \tag{7.10}$$

Estaremos interesados en ver como se relacionan el β con el α.

Ejemplo 7.1 *Comenzaremos con un ejemplo simplificado.*

La variable P tiene distribución $N(\mu_P, \sigma^2)$ y N tiene una distribución $N(\mu_N, \sigma^2)$. Recordemos que $\mu_N < 0$ porque le habíamos asignado el signo a la variable. Además asumiremos que P y N son independientes.

Bajo estas condiciones la variable $\tilde{I}$ tiene distribución $N(\theta = \alpha\mu_P + (1-\alpha)\mu_N, \Sigma = \sigma^2(\alpha^2 + (1-\alpha)^2))$.

Luego la probabilidad de saturación se puede calcular de la siguiente manera:

$$\begin{aligned} P(\tilde{I} < u) &\leq \beta \\ P\left(\frac{\tilde{I}-\theta}{\Sigma} < \frac{u-\theta}{\Sigma}\right) &\leq \beta \\ \Phi\left(\frac{u-\theta}{\Sigma}\right) &\leq \beta \\ \frac{u-\theta}{\Sigma} &\leq \Phi^{-1}(\beta) \end{aligned}$$

Como $\Sigma \geq (1-\sqrt{2}\alpha)^2\sigma$, encadenando desigualdades obtenemos:

$$\frac{u-\theta}{\Sigma} \leq \frac{u-\theta}{(1-\sqrt{2}\alpha)^2\sigma} \tag{7.11}$$

Entonces toda condición que pidamos sobre el miembro de la derecha de (7.11), implicará que también la cumple el primer miembro.

$$\begin{aligned} \frac{u-\theta}{\Sigma} \leq \frac{u-\theta}{(1-\sqrt{2}\alpha)^2\sigma} &\leq \Phi^{-1}(\beta) \\ u-\theta &\leq \Phi^{-1}(\beta)\sigma(1-\sqrt{2}\alpha)^2 \\ -\theta &\leq \Phi^{-1}(\beta)\sigma(1-\sqrt{2}\alpha)^2 - u \\ -\alpha(\mu_P-\mu_N)-\mu_N &\leq \Phi^{-1}(\beta)\sigma(1-\sqrt{2}\alpha)^2 - u \\ -\alpha(\mu_P-\mu_N) &\leq \Phi^{-1}(\beta)\sigma(1-\sqrt{2}\alpha)^2 - u + \mu_N \\ 0 &\leq \Phi^{-1}(\beta)\sigma(1-\sqrt{2}\alpha)^2 - u + \mu_N + \alpha(\mu_P-\mu_N) \\ 0 &\leq 2\Phi^{-1}(\beta)\sigma\alpha^2 + (-2\sqrt{2}\Phi^{-1}(\beta)\sigma + (\mu_P-\mu_N))\alpha + \\ &\quad + \mu_N - u + \Phi^{-1}(\beta)\sigma \end{aligned} \tag{7.12}$$

La fórmula (7.12) es una expresión cuadrática en α. El estudio del signo de esta expresión nos permitirá encontrar la condición que debe cumplir el α para lograr la probabilidad de saturación menor que la β.

Veamos para que valores de (7.12) tiene raíces reales.

Utilicemos la siguiente notación, $\sigma\Phi^{-1}(\beta) = Q$ y $\mu_P - \mu_N = D$. Con esta notación la desigualdad (7.12) queda:

$$0 \leq 2Q\alpha^2 + (D - 2\sqrt{2}Q)\alpha + Q + \mu_N - u \tag{7.13}$$

Estudiemos el discriminante de (7.13) y veamos cuando es positivo.

$$\begin{aligned} \Delta = (D - 2\sqrt{2}Q)^2 - 8Q(Q + \mu_N - u) \geq 0 \\ (D - 2\sqrt{2}Q)^2 \geq 8Q(Q + \mu_N - u) \end{aligned} \tag{7.14}$$

Como $Q < 0$, esta desigualdad se cumple cuando $Q + \mu_N - u < 0$. Pero como Q, μ_N y $-u$ son negativos, entonces vale siempre.

Entonces la expresión (7.12) o (7.13) con la notación de Q y D siempre tiene raíces reales. Estas son:

$$\alpha_1 = \frac{\sqrt{2}}{2} - \frac{D}{4Q} + \frac{\sqrt{D^2 - 4\sqrt{2}QD - 8Q(\mu_N - u)}}{4Q}, \tag{7.15}$$

y

$$\alpha_2 = \frac{\sqrt{2}}{2} - \frac{D}{4Q} - \frac{\sqrt{D^2 - 4\sqrt{2}QD - 8Q(\mu_N - u)}}{4Q}. \tag{7.16}$$

Como $Q < 0$ observemos que $\alpha_2 > \alpha_1$.

Si pedimos que $-4\sqrt{2}Q - 8Q(\mu_N - u) < 0$, se puede probar que $\alpha_1 > \dfrac{\sqrt{2}}{2}$ y que $\alpha_2 < \dfrac{\sqrt{2}}{2} - \dfrac{D}{2Q}$.

Recordemos que el valor del parámetro α debe estar entre las raíces, es decir que debe ser:

$$\frac{\sqrt{2}}{2} < \alpha < \text{mín}(1, \frac{\sqrt{2}}{2} - \frac{D}{2Q}). \tag{7.17}$$

Luego, si α verifica (7.17), entonces la probabilidad de que el proceso de interés global sea menor que u es menor que β.

Tiene una interpretación razonable este resultado porque significa que el proceso de la mezcla que debe ser preponderante es el positivo.

Esto quiere decir que para tener una probabilidad de saturación requerida, el menor valor de las afinidades positivas debe ser mayor que $\frac{\sqrt{2}}{2}$ *y la menor afinidad negativa no debe ser menor que* $\frac{\sqrt{2}}{2} - 1$.

En el ejemplo anterior trabajamos con hipótesis usuales de normalidad de los procesos e independencia. Se puede realizar un cálculo similar para otra función de distribución y de manera análoga encontrar el parámetro de mezcla adecuado. El conocimiento de este valor mínimo nos permite por ejemplo seleccionar algún perfil con muchos seguidores que fuerce la atención positiva, por medio de promociones, concursos o juegos de preguntas y respuestas de manera de evitar la saturación y el alejamiento de usuarios de la red social.

Conclusiones y Trabajos Futuros

7.3. Conclusiones

En esta tesis presentamos un resumen del desarrollo del concepto red social y algunos datos de cómo las redes sociales impactan en nuestra realidad cercana, a pesar de ser redes que transcurren en un espacio virtual tienen consecuencias medibles y concretas.

Esta posibilidad de abordar su estudio nos llevó a proponer un modelo estocástico para Twitter con el que analizamos algunos aspectos de su dinámica y nos permitió sacar conclusiones sobre la misma.

Respecto de esto podemos asegurar que si el modelo de emisión de tweets tiene ciertas características, un usuario que por alguna circunstancia inicial comenzó a seguir a un perfil con el que no desarrolló una afinidad positiva, será saturado por el contenido de este y en la eventualidad que este proceso se produzca en varias oportunidades, abandonará o suspenderá el uso de Twitter.

Para analizar el modelo considerando múltiples fuentes hemos elegido el enfoque de cópulas, ya que nos permite evitar supuestos de independencia entre perfiles que en la mayoría de los casos no aportan a una descripción realista del fenómeno.

Para ajustar las cópulas al modelo de aproximaciones fluidas, es decir, modelar la emisión de tweets con un proceso continuo; desarrollamos uno de los aportes originales de esta tesis.

Con este aporte generalizamos el estimador no paramétrico a partir de procesos empíricos de la cópula del proceso y a partir de él podemos tener en cuenta la dependencia del mismo al momento de estudiar las probabilidades de saturación.

7.4. Trabajos futuros

En el desarrollo de esta tesis, nos fuimos encontrando con muchos problemas que quedaron planteados y que podrían ser objeto de interés de futuros trabajos.

Podemos distinguirlos en dos grandes temas:

7.4.1. Trabajos enfocados en Twitter:

- Aplicación y análisis de probabilidad de saturación a partir de datos reales. Realizando seguimiento de perfiles y toda las consideraciones correspondientes para determinar las segmentaciones y los coeficientes de afinidad. Con estos datos, estudiar probabilidades de saturación y determinación del comportamiento post-saturación, de umbrales, etc.

- Analizar las dependencias de los procesos de emisión según las distintas formas de agregarlos, por segmentos, por temas, etc. Con esta información se podrían llegar a testear hipótesis sobre los medios de comunicación del tipo: "un twitter de futbol genera ciertos minutos de contenido en televisión y en radio", etc.

7.4.2. Trabajos a partir del estimador de cópulas:

- Para el estimador de cópulas de difusiones pedíamos una serie de hipótesis, una pregunta obligada es si son esenciales o si alguna de estas hipótesis se puede relajar y hasta cuanto.

- Qué pasaría con el estimador para otro tipo de procesos a tiempo continuo pero con saltos como el de Poisson.

- A partir del estimador se pueden proponer distintos test de hipótesis para la cópula que pueden ser de suma utilidad al momento de estudiar dependencias y riesgos.

Por último quisiera mencionar como una posible linea de trabajo a futuro la aplicación de cópulas a finanzas. En la que ya se están haciendo muchos avances y que se muestra como una herramienta muy interesante y que puede aportar mucho en esta área. No por nada los principales desarrollos en este tema surgieron en el interior de las finanzas y se propagaron hacia otras ciencias.

Bibliografía

[1] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, *Characterizing user behavior in online social networks*, in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, IMC '09, New York, NY, USA, 2009, ACM, pp. 49–62.

[2] P. Billingsley, *Probability and measure*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, third ed., 1995. A Wiley-Interscience Publication.

[3] ——, *Convergence of probability measures*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, second ed., 1999. A Wiley-Interscience Publication.

[4] G. E. Box and N. R. Draper, *Empirical Model-building and Response Surfaces*, Wiley, 1987.

[5] D. M. Boyd and N. B. Ellison, *Social network sites: Definition, history, and scholarship*, Journal of Computer-Mediated Communication, 13(1) (2007). http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html.

[6] S. Brain, *"twitter statistics - statistic brain"*. `http://www.statisticbrain.com/twitter-statistics/`, Mayo 2012.

[7] F. Carrero. `http://franciscocarrero.com/2008/06/page/2/`.

[8] A. Charpentier, J. D. Fermanian, and O. Scaillet, *The Estimation of Copulas: Theory and Practice*, in Copulas: From theory to application in finance, J. Rank, ed., Risk Books, 2007, pp. 35–62.

[9] R. Cintas del Río, *Teoría de Cópulas y control de riesgo Financiero*, PhD thesis, Universidad Complutense de Madrid. Facultad de Ciencias Matemáticas. Departamento de Estadística e Investigación Operativa., 2007.

[10] W. F. Darsow, B. Nguyen, and E. T. Olsen, *Copulas and Markov processes*, Illinois J. Math., 36 (1992), pp. 600–642.

[11] P. Deheuvels, *La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance*, Acad. Roy. Belg. Bull. Cl. Sci. (5), 65 (1979), pp. 274–292.

[12] A. L. Depaoli, *¿cuanto tardamos en escribir un tweet?* `http://mandomando.com/2012/05/30/cuanto-tardamos-en-escribir-un-tweet/`, Mayo 2012.

[13] Diario La Nación Artículo, *La polémica de darín.* `http://www.lanacion.com.ar/1543919-un-tuit-cada-120-segundos-por-la-polemica-de-darin`.

[14] P. Doukhan, *Mixing: Properties and Examples*, no. 85 in Lecture Notes in Statistics, Springer-Verlag, 1994.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, Wiley, January 1968.

[16] J. D. Fermanian and O. Scaillet, *Nonparametric estimation of copulas for time series*, fame research paper series, International Center for Financial Asset Management and Engineering, 2003.

[17] L. C. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science*, Empirical Press, 2004.

[18] B. Furht, *Handbook of Social Network Technologies and Applications*, Springer, 2010.

[19] M. Granovetter, *Economic action and social structure: The problem of embeddedness*, The American Journal of Sociology, 91 (1985), pp. 481–510.

[20] P. Hawe, C. Webster, and A. Shiell, *A glossary of terms for navigating the field of social network analysis.*, J Epidemiol Community Health, 58 (2004), pp. 971–975.

[21] D. J. Higham, *An algorithmic introduction to numerical simulation of stochastic differential equations*, SIAM Review, 43 (2001), pp. 525–546.

[22] J. B. Hill, *Dependence properties and asymptotic theory.* University Lecture, 2012.

[23] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, vol. 113 of Graduate Texts in Mathematics, Springer-Verlag, New York, second ed., 1991.

[24] S. Karlin and H. M. Taylor, *A first course in stochastic processes*, Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, second ed., 1975.

[25] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, vol. 23 of Applications of Mathematics (New York), Springer-Verlag, Berlin, 1992.

[26] Y. A. Kutoyants and I. Negri, *On L_2 efficiency of empiric distribution for ergodic diffusion processes*, Teor. Veroyatnost. i Primenen., 46 (2001), pp. 164–169.

[27] C. H. Ling, *Representation of associative functions*, Publ Math Debrecen, 12 (1965), pp. 189–212.

[28] P. Mandl, *Analytical treatment of one-dimensional Markov processes*, Die Grundlehren der mathematischen Wissenschaften, Band 151, Academia Publishing House of the Czechoslovak Academy of Sciences, Prague; Springer-Verlag New York Inc., New York, 1968.

[29] R. B. Nelsen, *An introduction to copulas*, Springer Series in Statistics, Springer, New York, second ed., 2006.

[30] F. Olmos Maldonado, *Procesos estocásticos. integral estocástica ecuaciones diferenciales estocasticas: Lema de ito.* Lecturas de Curso, 2010.

[31] J. Quesada Molina and J. Rodryguez Lallena, *What are copulas?*, Monografías del Semin. Matem. Garcia de Galdeano, 27 (2003), pp. 499–506.

[32] P. Rakonczai, L. Márkus, and A. Zempléni, *Autocopulas: Investigating the interdependence structure of stationary time series*, Methodology and Computing in Applied Probability, 14 (2012), pp. 149–167.

[33] L. Rincón, *Introducción a los procesos estocásticos*, 2012.

[34] B. Schweizer, *Thirty years of copulas*, in Advances in Probability Distributions with Given Marginals, vol. 67 of Mathematics and Its Applications, Springer Netherlands, 1991, pp. 13–50.

[35] J. P. Scott, *Social Network Analysis: A Handbook*, SAGE Publications, Jan. 1987.

[36] G. Simmel, *On Individuality and Social Forms*, University of Chicago Press, 1971.

[37] M. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statist. Univ. Paris, 8 (1959), pp. 229–231.

[38] T. A. B. Snijders, C. E. G. Steglich, and M. Schweinberger, *Modeling the co-evolution of networks and behavior*, in In, 2006.

[39] D. Stefanova, *Dependence modeling of joint extremes via copulas: A dynamic portfolio allocation perspective*, tech. report, HEC Montréal, 2007.

[40] G. Thompson. `http://socialnetworking.lovetoknow.com/Forms_and_Types_of_Social_Media`.

[41] H. F. Trotter, *A property of Brownian motion paths*, Illinois J. Math., 2 (1958), pp. 425–433.

[42] I. Twitter. `http://support.twitter.com/entries/68916`, 2012.

[43] A. Y. Veretennikov, *On polynomial mixing bounds for stochastic differential equations*, Stochastic Process. Appl., 70 (1997), pp. 115–127.

[44] J. C. Westland and A. Odlyzko, *Social networks and mathematical models: A research commentary on "critical mass and willingness to pay for social networks"*, 2010.

[45] M. white. `http://socialnetworking.lovetoknow.com/What_Types_of_Social_Networks_Exist`.

[46] Wikipedia, *Twitter — wikipedia, the free encyclopedia*. `http://en.wikipedia.org/w/index.php?title=Twitter&oldid=540849864`, 2013. [Online; accessed 3-March-2013].

[47] D. Zarrella, *Is twitter a social network*, Junio 2009.