CIMAT

Centro de Investigación en Matemáticas, A.C.

Distribución de direcciones en el Gibbs sampler generalizado

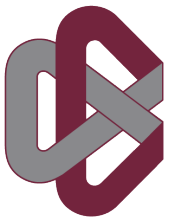
T E S I S

Que para obtener el grado de
Maestro en Ciencias
con especialidad en
Probabilidad y Estadística

P r e s e n t a:
José Cricelio Montesinos López

Director de tesis:
Dr. José Andrés Christen Gracia

Guanajuato, Gto. Noviembre de 2016



CIMAT

Centro de Investigación en Matemáticas, A.C.

DISTRIBUCIÓN DE DIRECCIONES EN EL GIBBS SAMPLER GENERALIZADO

José Cricelio Montesinos López

Dirigido por:
Dr. José Andrés Christen Gracia

Noviembre, 2016

A mi familia

Agradecimientos

Agradezco a mi asesor, el Dr. J. Andrés Christen Gracia, por creer en mí y por darme la oportunidad de enriquecerme con sus valiosas ideas, por su disposición y paciencia, y por enseñarme a ver la vida con optimismo y buen humor. Gracias Dr. Andrés por su motivación y por su apoyo durante toda la elaboración de este trabajo. También doy las gracias a mis sinodales, Dr. Rogelio Ramos Quiroga y Dr. Marcos Aurelio Capistrán Ocampo, por el tiempo que destinaron para la revisión de este trabajo.

Gracias al Centro de Investigación en Matemáticas (CIMAT), por darme una educación de calidad, y por permitirme usar sus instalaciones y recursos para concluir este trabajo. Agradezco también al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo económico brindado en los dos años de maestría.

Gracias a mis padres porque siempre han confiado en mi y por mantenerme con los pies en la tierra. Agradezco a mis hermanos por el apoyo incondicional que siempre me han brindado, y por compartir conmigo sus experiencias que han sido de gran utilidad en mi aprendizaje.

Gracias a mis compañeros por hacer que la estancia en la maestría fuera muy placentera. A mis amigos les agradezco por todo el cariño brindado, y por ser parte esencial en esta etapa de aprendizaje.

Resumen

En este trabajo se dan dos justificaciones de por qué funciona la propuesta de distribución de direcciones dada en el artículo "*Optimal Direction Gibbs Sampler for Truncated Multivariate Normal Distributions*", por Christen et al. (2015). La primera justificación consiste en minimizar la traza de la matriz de covarianzas de dos pasos consecutivos de la cadena, de este modo, la dirección óptima estaría reduciendo la dependencia, entrada a entrada, del estado actual $\boldsymbol{X}$ y el nuevo vector generado $\boldsymbol{Y}$. Para la segunda justificación, obtenemos la información mutua de la i-ésima entrada de $\boldsymbol{Y}$ con el estado actual $\boldsymbol{X}$, y después se minimiza la suma de funciones crecientes de estas informaciones. Con esto, la dirección óptima estaría reduciendo la dependencia de cada entrada de $\boldsymbol{Y}$ con la de todo el vector $\boldsymbol{X}$.

Además, aquí se propone una distribución de direcciones, distinta a la que aparece en el artículo, con el que se crea un algoritmo para generar muestras de la distribución Normal Truncada Multivariada. La propuesta está basada en minimizar las informaciones mutuas de cada entrada del vector $\boldsymbol{Y}$ con el estado actual $\boldsymbol{X}$. Ésta distribución de direcciones tiene como soporte a las columnas normalizadas de la matriz de varianzas y covarianzas de la distribución objetivo; la probabilidad de selección de cada dirección es más grande para aquellas que hacen que las entradas del vector generado $\boldsymbol{Y}$ sean lo menos dependientes del estado actual $\boldsymbol{X}$. El desempeño del algoritmo propuesto se evalúa con experimentos de simulación y se comparan con los resultados obtenidos en Christen et al. (2015).

Índice

Resumen III

1. Introducción 1

2. Gibbs Direccional, Información Mutua y Covarianza 4

2.1. Gibbs Direccional 4

2.2. Información Mutua para el Caso Normal 7

2.3. Matriz de covarianzas 9

2.4. Información Mutua marginal 14

3. Propuesta de distribución de direcciones 21

3.1. Caso Normal 21

3.1.1. Seleccionando un conjunto de direcciones 24

3.2. Normal Truncada 26

4. Experimentos numéricos 29

4.1. Caso Normal 29

4.2. Normal Truncada 32

5. Discusión 39

Referencias 41

Índice de figuras

4.1. Desviaciones estándar progresivamente más contrastantes. 30

4.2. Comportamiento del IAT/n: **(a)** contra la dimensión, para cada nivel α y **(b)** contra α, para diferentes dimensiones. Caso Normal Multivarada. . 31

4.3. Muestras de la Normal bivariada completa (puntos negros) y de la NT bivariada (puntos azules), con 5000 iteraciones, para **(a)** $\alpha = 0$, **(b)** $\alpha = 5$, **(c)** $\alpha = 10$ y **(d)** $\alpha = 20$. 33

4.4. Muestras de la Normal bivariada completa (puntos negros) y de la NT bivariada con el Algoritmo ODG1 (azul), ODG2 (rojo), KD (verde), DW (naranja), con 5000 iteraciones para **(a)** $\alpha = 0$, **(b)** $\alpha = 5$, **(c)** $\alpha = 10$ y **(d)** $\alpha = 20$. Fuente: Christen et al. (2015). 34

4.5. Comportamiento del IAT/n: **(a)** contra la dimensión, para cada nivel α y **(b)** contra α, para diferentes dimensiones. Caso NTM. 37

4.6. Muestras de la distribución NM completa (puntos negros) y de la NT de dimensión $n = 20$ (puntos azules), para **(a)** $\alpha = 0$, **(b)** $\alpha = 15$, **(c)** $\alpha = 30$ y **(d)** $\alpha = 50$; las curvas en verdes representan los contornos de la densidad truncada bivariada. 38

Índice de tablas

4.1. IAT/n para muestras de la Normal Multivariada. Cada cantidad es el promedio de 30 cadenas de longitud 10000. 31
4.2. Número de simulaciones (τ) requeridas para obtener una muestra pseudoindependiente. Cada cantidad es el promedio de 30 cadenas de longitud 5000. Normal Truncada bivariada. Los datos de los algoritmos KD y DW, fueron tomados de Christen et al. (2015). 35
4.3. IAT/n para muestras de la distribución NTM. Cada cantidad es el promedio de 30 cadenas de longitud 10000. 36

CAPÍTULO 1

Introducción

Los métodos de simulación de Cadenas de Markov Monte Carlo (MCMC, por su siglas en inglés) son algoritmos empleados para producir muestras de una distribución π, sin necesidad de simular directamente de dicha distribución. Estos métodos están basados en la construcción de una cadena de Markov Ergódica, cuya distribución estacionaria sea precisamente π. Los métodos MCMC han resultado ser de gran utilidad en diversas áreas, particularmente en Estadística Bayesiana, ya que permiten producir muestras, aún cuando la distribución de π es compleja.

El muestreador de Gibbs es un algoritmo MCMC que simula de manera sistemática o aleatoria de las distribuciones condicionales sobre un conjunto de direcciones. Un caso general del muestreador de Gibbs es el Gibbs Direccional Óptimo, el cual elige una dirección arbitraria sobre el espacio que se simula, y posteriormente se muestrea de la distribución condicional total de la dirección elegida. Esto puede ser escrito como,

$$\boldsymbol{X}^{(t+1)} = \boldsymbol{x}^{(t)} + r\boldsymbol{e},$$

donde $\boldsymbol{e} \in \mathbb{R}^n$ indica la dirección y $r \in \mathbb{R}$ representa la longitud de la transición. Notemos que, si tomamos al conjunto de direcciones $\boldsymbol{e}$ como las direcciones canónicas

y estas se eligen de manera sistemática, se obtiene como resultado el muestreador de Gibbs estándar. Mientras que si las direcciones canónicas se toman de forma aleatoria, se obtiene el Random Scan Gibbs Sampler.

La necesidad de muestrear de la distribución Normal Truncada Multivariada (NTM) es frecuente en la inferencia Bayesiana y en problemas inversos.

Una forma muy simple para muestrear de la distribución NTM sería generar valores de la Normal Multivariada y solamente aceptar aquellas muestras que estén dentro de la región de interés. Este método, el cual se conoce como Rejection Samplig, trabaja bien cuando la tasa de aceptación es razonablemente alta. Sin embargo, resulta ser muy ineficiente cuando la tasa de aceptación es baja, como en el caso de alta dimensión y/o cuando el soporte esta estrechamente acotado.

La mayoría de los métodos disponibles para muestrear de la distribución NTM están basados en el muestreador de Gibbs, el cual es simple de usar y tiene la ventaja de aceptar todas las propuestas generadas y, por lo tanto, no se ve afectada por tasas de aceptación pobres. El inconveniente que se tiene con las muestras producidas por el muestreador de Gibbs es que no son independientes, el grado de correlación depende tanto de la matriz de covarianzas como de la dimensionalidad.

En Christen et al. (2015) exploran un criterio de optimalidad para el algoritmo MCMC Gibbs direccional. Dicho criterio consiste en minimizar la información mutua entre dos pasos consecutivos de la cadena de Markov. También proponen, de forma heurística, una distribución de direcciones para el caso en que la distribución objetivo es la distribución NTM.

En éste trabajo se dan razones teóricas de por qué funciona, en los experimentos, las distribución de direcciones dada en Christen et al. (2015). Además, se propone una distribución de direcciones óptima, distinta a la que aparece en dicho artículo. Para evaluar el desempeño del algoritmo con la distribución de direcciones propuesta, se comparan con los resultados obtenidos en el mismo artículo.

Para exponer lo anterior, la tesis se divide en cinco capítulos. En el Capítulo 2 se darán dos justificaciones de por qué el criterio de minimizar $\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$, para la distribución de direcciones propuesto en Christen et al. (2015), en efecto tiene sentido. La primera consiste en minimizar la traza de la matriz de covarianzas de dos pasos consecutivos de la cadena, de este modo, la dirección óptima $\boldsymbol{e}$, estaría reduciendo la dependencia entrada a entrada del estado actual de la cadena $\boldsymbol{X}$ y el nuevo vector generado $\boldsymbol{Y}$. Para la segunda justificación, obtenemos la información mutua de la $i-$ésima entrada de $\boldsymbol{Y}$ con el estado actual $\boldsymbol{X}$ y después se minimiza la suma de funciones crecientes de estas informaciones mutuas. Con esto, la dirección óptima $\boldsymbol{e}$, estaría reduciendo la dependencia de cada entrada de $\boldsymbol{Y}$ con la de todo el vector $\boldsymbol{X}$.

A lo largo del Capítulo 3 se desarrolla el algoritmo MCMC Gibbs direccinal, que se utilizara para simular de una distribución NTM. En el Capítulo 4 se evalua el desempeño del algoritmo propuesto con experimentos de simulación y se comparan con los resultados obtenidos en Christen et al. (2015).

Por último, en el Capítulo 5 se presentan una discusión de los resultados obtenidos, mencionando las ventajas y dificultades del algoritmo propuesto.

CAPÍTULO 2

Gibbs Direccional, Información Mutua y Covarianza

En este Capítulo se brindan dos justificaciones de por qué el criterio de minimizar $\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$, para la distribución de direcciones propuesto en Christen et al. (2015), en efecto tiene sentido. La primera consiste en minimizar la traza de la matriz de covarianzas de dos pasos consecutivos, de este modo estaríamos reduciendo la dependencia, entrada a entrada, del estado actual de la cadena, $\boldsymbol{X}$, y el nuevo vector generado, $\boldsymbol{Y}$. En la segunda obtenemos la información mutua de la $i-$ésima entrada de $\boldsymbol{Y}$ con el estado actual $\boldsymbol{X}$ y después se minimiza la suma de funciones crecientes de estas informaciones mutuas. Con esto, estaríamos reduciendo la dependencia de cada entrada de $\boldsymbol{Y}$ con la de todo el vector $\boldsymbol{X}$.

2.1. Gibbs Direccional

Los métodos de simulación MCMC son algoritmos empleados para producir muestras de una distribución π, que suele ser compleja, sin necesidad de simular directamente

de dicha distribución. Estos métodos están basados en la construcción de una cadena de Markov Ergódica $X^{(t)}$ cuya distribución estacionaria sea precisamente π.

Dada la definición de los métodos MCMC, uno puede proponer una infinidad de implementaciones, pero querríamos que dichas implementaciones fueran óptimas en algún sentido, tales como: rápida convergencia, o que el número de muestras necesarias para obtener una muestra pseudo-independiente sea lo más pequeño posible.

El muestreador de Gibbs es un algoritmo MCMC que simula de manera sistemática o aleatoria de las distribuciones condicionales sobre un conjunto de direcciones. Un caso general del muestreador de Gibbs es el Gibbs Direccional Óptimo, el cual elige una dirección arbitraria $\boldsymbol{e} \in \mathbb{R}^n$ tal que $\|\boldsymbol{e}\| = 1$, donde $\|\cdot\|$ es la norma euclidiana, y posteriormente se muestrea de la distribución condicional total de la dirección elegida. Esto puede ser escrito como

$$\boldsymbol{X}^{(t+1)} = \boldsymbol{x}^{(t)} + r\boldsymbol{e},$$

donde $\boldsymbol{e}$ indica la dirección y $r \in \mathbb{R}$ representa la longitud de la transición, la cual, condicionado a $\boldsymbol{e}$ y a $\boldsymbol{x}^{(t)}$, tiene distribución proporcional a $\pi\left(\boldsymbol{x}^{(t)} + r\boldsymbol{e}\right)$ (Liu, 2008). El algoritmo funciona de la siguiente manera:

Algoritmo 1 Gibbs Direccional Óptimo

Dado $\boldsymbol{X}^{(t)} = \boldsymbol{x}$,

1. Se genera $\boldsymbol{e} \sim h\left(\boldsymbol{e}\right)$.
2. Se genera $r \sim g\left(r|\boldsymbol{e}, \boldsymbol{x}\right)$.
3. Hacemos $\boldsymbol{X}^{(t+1)} = \boldsymbol{x} + r\boldsymbol{e}$.

Se puede ver que el Kernel de transición que surge del Algoritmo anterior esta en balance detallado con π y, suponiendo $\pi - irreducible$, la cadena de Markov generada tiene a π como distribución ergódica. De forma simple, $\pi - irreducible$ implica que hay

una probabilidad positiva de moverse de $\boldsymbol{x}$ a $\boldsymbol{y}$ en un número finito de pasos, para cada $\boldsymbol{x}$ y $\boldsymbol{y}$ en el soporte de π, consultar Robert y Casella (2013) para más detalles. De hecho, una vez supuesto la irreducibilidad, cualquier cadena tendrá como distribución ergódica a π, por lo que el desempeño dependerá de que tan dependientes sean $\boldsymbol{X}^{(t)}$ y $\boldsymbol{X}^{(t+1)}$.

La pregunta natural que surge es si nosotros podemos tomar cualquier dirección en el Gibbs sampler, y si es así, cómo elegir cuál dirección tomar. De hecho se pueden seleccionar direcciones arbitrarias, tomando un nuevo punto en la cadena de Markov simulado de la distribución condicional sobre esa dirección (Liu, 2008). Sin embargo, no es claro cómo elegir tal dirección y qué criterio usar para optimizar la cadena.

Para responder a la pregunta: ¿cómo escoger la distribución de las direcciones $\boldsymbol{e}$ de forma que se optimice (en algún sentido) el desempeño del algoritmo Gibbs Direccional?, se debe definir primero un criterio de optimalidad. Una vez hecho esto, entonces se debe seleccionar una distribución de probabilidad h, para $\boldsymbol{e}$, que satisfaga dicho criterio. Aquí es importante recalcar que el algoritmo será *óptimo* sólo en el sentido de ese criterio en particular.

Por lo dicho anteriormente sobre el desempeño del algoritmo, una forma de atacar el problema de optimización sería encontrar una medida de dependencia entre dos variables aleatorias y después minimizarla. Con esto, estaremos reduciendo la dependencia en la cadena, y así tendríamos más observaciones pseudo-independientes con pocas iteraciones.

En Christen et al. (2015) proponen como medida de dependencia la información mutua entre dos variables aleatorias $\boldsymbol{X}$ e $\boldsymbol{Y}$, la cual mide la divergencia de Kullback-Leibler entre el modelo conjunto $f_{\boldsymbol{Y},\boldsymbol{X}}$ y el alternativo independiente $f_{\boldsymbol{Y}}(\boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{x})$, esto es,

$$I\left(f_{\boldsymbol{Y}\boldsymbol{X}}, f_{\boldsymbol{Y}} f_{\boldsymbol{X}}\right) = \int \int f_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x}) \log \frac{f_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x})}{f_{\boldsymbol{Y}}(\boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{x})} \, d\boldsymbol{x} \, d\boldsymbol{y}. \tag{2.1}$$

De las propiedades heredadas de la divergencia de Kullback-Leibler se tiene que $I \geq 0$ y, a partir de la desigualdad de Jensen, se puede probar que $I = 0$ si y sólo si $f_{\boldsymbol{Y},\boldsymbol{X}} = f_{\boldsymbol{Y}}(\boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{x})$, i.e., si y sólo si $\boldsymbol{X}$ e $\boldsymbol{Y}$ son independientes .

A partir de (2.1), Christen et al. (2015) exploran un criterio de optimalidad para el algoritmo Gibbs Direccional. Dicho criterio consiste en minimizar la información mutua entre dos pasos consecutivos ($\boldsymbol{X}^{(t)}$ y $\boldsymbol{X}^{(t+1)}$) de la cadena de Markov generada por el algoritmo.

2.2. Información Mutua para el Caso Normal

Supongamos que la distribución objetivo π es una Normal $n-variada$ con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{A}$, la cual es la inversa de la matriz de covarianzas. Si $\boldsymbol{Z} \sim \pi$, entonces su función de densidad esta dada por

$$\pi(\boldsymbol{z}) = \left(\frac{|\boldsymbol{A}|}{(2\pi)^n}\right)^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{A}(\boldsymbol{z}-\boldsymbol{\mu})\right\}.$$

Haciendo $\boldsymbol{X} = \boldsymbol{X}^{(t)}$ y $\boldsymbol{Y} = \boldsymbol{X}^{(t+1)}$ y suponiendo que $\boldsymbol{X}^{(t)} \sim \pi$, es decir, suponemos que ya estamos en la distribución estacionaria, vemos que $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \pi(\boldsymbol{y})$ y $f_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) = K(\boldsymbol{x},\boldsymbol{y})\pi(\boldsymbol{x})$. Christen et al. (2015) encontraron que,

$$\begin{aligned} I_{\boldsymbol{e}}(f_{\boldsymbol{XY}}, f_{\boldsymbol{X}} f_{\boldsymbol{Y}}) &= \int\int K(\boldsymbol{x},\boldsymbol{y})\pi(\boldsymbol{x}) \log\frac{K(\boldsymbol{x},\boldsymbol{y})}{\pi(\boldsymbol{y})}\, d\boldsymbol{x}\, d\boldsymbol{y} \\ &= C + \frac{1}{2}\log\left(\boldsymbol{e}^T \boldsymbol{A}\boldsymbol{e}\right), \end{aligned} \tag{2.2}$$

donde $C = n - \frac{1}{2} + \frac{n-1}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{A}|$, es una constante que no depende de $\boldsymbol{e}$.

De acuerdo a la ecuación (2.2), la mejor dirección es aquella que minimiza $C + \frac{1}{2}\log\left(\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\right)$, o de forma equivalente, es aquella dirección que minimice $\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$, que para este caso es el eigenvector $\boldsymbol{e}_n$ que corresponde al eigenvalor más pequeño λ_n de la matriz de precisión $\boldsymbol{A}$.

Sin embargo, no podemos tomar únicamente la mejor dirección, porque la cadena resultante no será irreducible y claramente no estaremos muestreando de π, ya que la cadena sólo recorrería una línea en el espacio de estados. La cadena debe ser $\pi-irreducible$

para que tenga a π como distribución ergódica. De hecho, si la distribución de direcciones h tiene soporte en la esfera S^n, entonces la cadena de Markov resultante es irreducible con kernel de transición dado por,

$$K(\boldsymbol{x}, \boldsymbol{y}) = \int K_{\boldsymbol{e}}(\boldsymbol{x}, \boldsymbol{y})\, h(\boldsymbol{e})\; d\,\boldsymbol{e}.$$

Por lo tanto, necesitamos una distribución completa para $\boldsymbol{e}$. Christen et al. (2015) tomaron el resto de los eigenvectores para garantizar que la cadena sea $\pi-$irreducible. Para ello hicieron que las direcciones $\boldsymbol{e}$ fueran los eigenvectores de la matriz de precisión $\boldsymbol{A}$, así $\boldsymbol{e} \in \{\boldsymbol{e}_1, \boldsymbol{e}, \ldots, \boldsymbol{e}_n\}$. La $i-$ésima dirección la tomaron con probabilidad proporcional a λ_i^{-b}, donde λ_i es el eigenvalor correspondiente al $i-$ésimo eigenvector, $i = 1, 2, \ldots, n$, y b es una variable aleatoria con distribución $Beta(\alpha, \beta)$. Luego

$$h_1(\boldsymbol{e}_i) = k\,(\lambda_i)^{-b},$$

donde $k = \left(\sum_{i=1}^n \lambda_i^{-b}\right)^{-1}$, y observaron que la dirección $\boldsymbol{e}_n$ correspondiente al eigenvalor más pequeño λ_n de $\boldsymbol{A}$ es óptima. Note que al minimizar la ecuación (2.2),

$$\begin{aligned}
\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1} I_{\boldsymbol{e}}(f_{\boldsymbol{XY}}, f_{\boldsymbol{X}} f_{\boldsymbol{Y}}) &= \min_{\|\boldsymbol{e}\|=1} \left(C + \frac{1}{2}\log\left(\boldsymbol{e}^T \boldsymbol{A}\boldsymbol{e}\right)\right) \\
&= C + \frac{1}{2}\log\left(\min_{\|\boldsymbol{e}\|=1}\left\{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\right\}\right) \\
&= C + \frac{1}{2}\log\lambda_n.
\end{aligned}$$

y el mínimo se alcanza cuando $\boldsymbol{e} = \boldsymbol{e}_n$, es el eigenvector asociado a λ_n, es decir, dan más peso a la dirección donde se encuentra la máxima variabilidad (ver Johnson y Wichern (2014)).

Dado que todos los eigenvectores tienen probabilidad positiva de ser elegidos, y además forman una base de $\mathbb{R}^n$, el Gibbs direccional resultante será ergódico. El algoritmo que ellos proponen, para simular de la distribución Normal Multivariada, trabaja de la siguiente manera:

Algoritmo 2 Gibbs Direccional Óptimo: Normal Multivariada

Dado $\boldsymbol{X}^{(t)} = \boldsymbol{x}$,

1. Proponer una dirección $\boldsymbol{e}$ de la distribución h_1.
2. Proponer una logitud $r \sim N\left(-\frac{\boldsymbol{e}^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}, \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\right)$.
3. Hacer $\boldsymbol{X}^{(t+1)} = \boldsymbol{x} + r\boldsymbol{e}$.

El problema que se encontró en el criterio propuesto por Christen et al. (2015) fue que, por definición, la información mutua debe ser ser positiva o cero para cualquier dirección $\boldsymbol{e}$. Sin embargo (2.2) puede resultar negativa. Este problema viene de la dependencia entre $\boldsymbol{X}^{(t+1)}$ y r. Notemos que, ya que fijamos $\boldsymbol{x}^{(t)}$ y $\boldsymbol{e}$, entonces, $\boldsymbol{X}^{(t+1)} = \boldsymbol{x}^{(t)} + r\boldsymbol{e}$ esta completamente determinada por r, de modo que la densidad de $\boldsymbol{Y}$ vive en la línea $\boldsymbol{x}^{(t)} + r\boldsymbol{e}$, es decir, la densidad condicional $f(\boldsymbol{y}|\boldsymbol{x})$ que se empleo para el cálculo de la información mutua no fue la correcta.

Lo que motivo esta tesis fue responder a la pregunta, ¿por qué funcionaron, en los experimentos, las distribuciones de direcciones dadas en Christen et al. (2015)? En las siguientes dos secciones se dan argumentos del por qué el criterio de minimizar $\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$, sí tiene sentido.

2.3. Matriz de covarianzas

El siguiente resultado será la base de esta sección.

Si $\boldsymbol{X} \sim NM(\boldsymbol{\mu_1}, \Sigma_1)$ y $\boldsymbol{Y} \sim NM(\boldsymbol{\mu_2}, \Sigma_2)$, entonces $\boldsymbol{X}$ y $\boldsymbol{Y}$ son independientes si y solo si $Cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{0}$.

Definamos a $\boldsymbol{V} := Cov\left(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t+1)}\right)$, la matriz de covarianzas de dos pasos consecutivos de la cadena. Por el resultado descrito anteriormente, un criterio de optimización sería encontrar la dirección $\boldsymbol{e}$ que haga que las entradas de la matriz de covarianzas $\boldsymbol{V}$ sean, en valor absoluto, lo más cercanamente posibles a cero. Trabajar con este criterio resulta bastante complicado. Pero si nos fijamos en las entradas de la diagonal principal de $\boldsymbol{V}$ y probamos que todas son mayores o iguales a cero, entonces, al minimizar la traza de $\boldsymbol{V}$, estaríamos haciendo mínima la dependencia, entrada a entrada, entre $\boldsymbol{X}^{(t)}$ y $\boldsymbol{X}^{(t+1)}$, y abríamos encontrado un criterio de optimalidad.

Siguiendo esta idea, lo que prosigue es: primero encontrar la matriz de covarianzas $\boldsymbol{V}$, después probar que los elementos de la diagonal principal $(v_{ii}, \quad \forall i = 1, \ldots, n)$ de $\boldsymbol{V}$ son mayores que cero, y por último minimizaremos la traza de $\boldsymbol{V}$.

Suponiendo que $\boldsymbol{X}^{(t)} \sim \pi$, tenemos

$$\begin{aligned} \boldsymbol{V} &= Cov\left(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t+1)}\right) \\ &= Cov\left(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t)} + r\boldsymbol{e}\right) \\ &= Cov\left(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t)}\right) + Cov\left(\boldsymbol{X}^{(t)}, r\boldsymbol{e}\right) \\ &= \boldsymbol{A}^{-1} + Cov\left(\boldsymbol{X}^{(t)}, r\right)\boldsymbol{e}^{T}. \end{aligned} \tag{2.3}$$

Para simplificar el lado derecho de (2.3) haremos uso del siguiente resultado.

Covarianza condicional. Si $\boldsymbol{X}$, $\boldsymbol{Y}$ y $\boldsymbol{Z}$ son variables aleatorias, entonces se cumple lo siguiente,

$$Cov\left(\boldsymbol{X}, \boldsymbol{Y}\right) = \mathbb{E}\left(Cov\left(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}\right)\right) + Cov\left(\mathbb{E}\left(\boldsymbol{X}|\boldsymbol{Z}\right), \mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{Z}\right)\right). \tag{2.4}$$

Ahora, sea g la densidad de la longitud r, ya que $g(r)$ es proporcional a $\pi\left(\boldsymbol{x} + r\boldsymbol{e}\right)$ (Liu, 2008), es decir,

$$g\left(r|\boldsymbol{e}, \boldsymbol{x}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{v} + r\boldsymbol{e}\right)^{T}\boldsymbol{A}\left(\boldsymbol{v} + r\boldsymbol{e}\right)\right\},$$

con $\boldsymbol{v} = \boldsymbol{x} - \mu$, entonces

$$r|\mathbf{e}, \boldsymbol{x} \sim N\left(-\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}, \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\right), \tag{2.5}$$

donde $\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$ es la precisión. Para ver lo anterior, notemos que

$$\begin{aligned}
g\left(r|\boldsymbol{e}, \boldsymbol{x}\right) &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{v} + r\boldsymbol{e}\right)^T \boldsymbol{A}\left(\boldsymbol{v} + r\boldsymbol{e}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(r^2\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e} + 2r\boldsymbol{e}^T A\boldsymbol{v} + \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(r^2\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e} + 2r\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e} + \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\left(\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)^2\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\left(r^2 + 2r\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}} + \left(\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)^2\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\left(r + \frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)^2\right\}.
\end{aligned}$$

Usando el resultado dado en (2.4) y por el de (2.5) se sigue,

$$\begin{aligned}
Cov\left(\boldsymbol{X}^{(t)}, r\right) &= \mathbb{E}\left(Cov\left(\boldsymbol{X}^{(t)}, r|\boldsymbol{X}^{(t)}\right)\right) + Cov\left(\mathbb{E}\left(\boldsymbol{X}^{(t)}|\boldsymbol{X}^{(t)}\right), \mathbb{E}\left(r|\boldsymbol{X}^{(t)}\right)\right) \\
&= 0 + Cov\left(\boldsymbol{X}^{(t)}, -\frac{\boldsymbol{e}^T\boldsymbol{A}\left(\boldsymbol{X}^{(t)} - \mu\right)}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right) \\
&= Cov\left(\boldsymbol{X}^{(t)}, -\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{X}^{(t)}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right) + Cov\left(\boldsymbol{X}^{(t)}, \frac{\boldsymbol{e}^T\boldsymbol{A}\mu}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right) \\
&= -Cov\left(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t)}\right)\frac{\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}} + 0 \\
&= -\frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{e} \\
&= -\frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\boldsymbol{e}.
\end{aligned} \tag{2.6}$$

Sustituimos (2.6) en (2.3) para obtener,

$$\boldsymbol{V} = \boldsymbol{A}^{-1} - \frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\boldsymbol{e}\boldsymbol{e}^T. \tag{2.7}$$

Ahora, para ver como son los elementos de la diagonal principal de $\boldsymbol{V}$, definamos $\sigma_{ij} = \left(\boldsymbol{A}^{-1}\right)_{ij}$ y sea $\boldsymbol{d}_i^T = (0, \ldots, 0, 1, 0, \ldots 0)$ un vector con un 1 en la $i-$ésima posición y ceros fuera de esta, también notemos lo siguiente,

$$\boldsymbol{e}\boldsymbol{e}^T = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix} \begin{bmatrix} e_1 & e_2 & \cdots & e_i & \cdots & e_n \end{bmatrix} = \begin{bmatrix} e_1^2 & e_1e_2 & \cdots & e_1e_i & \cdots & e_1e_n \\ e_2e_1 & e_2^2 & \cdots & e_2e_i & \cdots & e_2e_n \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ e_ie_1 & e_ie_2 & \cdots & e_i^2 & \cdots & e_ie_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e_ne_1 & e_ne_2 & \cdots & e_ne_i & \cdots & e_n^2 \end{bmatrix}$$

De (2.7) y con la notación anterior podemos ver que los elementos de la diagonal principal de $\boldsymbol{V}$ están dados por,

$$\begin{aligned} v_{ij} &= \left(\boldsymbol{A}^{-1}\right)_{ii} - \frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\boldsymbol{e}_i^2 \\ &= \sigma_{ii} - \frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}. \end{aligned} \tag{2.8}$$

A continuación enunciamos dos resultados que serán de utilidad a lo largo de la tesis (ver Johnson y Wichern (2014)).

Lema de Maximización. Sea $\boldsymbol{A}$ una matriz definida positiva y $\boldsymbol{d}$ un vector dado. Entonces, para un vector no nulo arbitrario $\boldsymbol{e}$,

$$\sup \frac{\left(\boldsymbol{e}^T\boldsymbol{d}\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}} = \boldsymbol{d}^T\boldsymbol{A}^{-1}\boldsymbol{d}, \tag{2.9}$$

y el máximo se obtiene cuando $\boldsymbol{e} = c\boldsymbol{A}^{-1}\boldsymbol{d}$ para cualquier constante $c \neq 0$.

Maximización de formas cuadráticas para un punto en la esfera unitaria. Sea $\boldsymbol{A}_{(nxn)}$ una matriz definida positiva con eigenvalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ y sean $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ los eigenvectores normalizados asociados. Entonces,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left(\frac{\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \right) = \lambda_n \qquad (\text{se alcanza cuando } \boldsymbol{x} = \boldsymbol{e}_n) \tag{2.10}$$

$$\max_{\boldsymbol{x}\in\mathbb{R}^n} \left(\frac{\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \right) = \lambda_1 \qquad (\text{se alcanza cuando } \boldsymbol{x} = \boldsymbol{e}_1)\,. \tag{2.11}$$

Además,

$$\max_{\boldsymbol{x}\perp\boldsymbol{e}_1,\ldots,\boldsymbol{e}_k} \left(\frac{\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \right) = \lambda_{k+1},$$

donde el símbolo $\perp$ se lee «es perpendicular a». Y el máximo se alcanza cuando $\boldsymbol{x} = \boldsymbol{e}_{k+1}, k = 1, 2, \ldots, p-1$.

Del lema (2.9) tenemos que,

$$\begin{aligned} \sup \frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}} &= \boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{d}_i \\ &= \left(\boldsymbol{A}^{-1}\right)_{ii} \\ &= \sigma_{ii}, \qquad \forall i = 1, \ldots, n, \end{aligned}$$

entonces,

$$\frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}} \leq \sigma_{ii}; \qquad \forall i = 1, \ldots, n, \tag{2.12}$$

de donde se sigue que,

$$v_{ii} = \sigma_i^2 - \frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}} \geq 0; \qquad \forall i = 1, \ldots, n.$$

Esto es, las entradas de la diagonal principal de la matriz de covarianzas de dos pasos consecutivos, son mayores o iguales a cero. De este modo, minimizar su traza tiene buena interpretación.

Por otro lado, de (2.7) y por ser la traza un operador lineal, podemos obtener una expresión para la traza de $\boldsymbol{V}$ de la siguiente manera,

$$\begin{aligned}\text{tr}\left(\boldsymbol{V}\right) &= \text{tr}\left(\boldsymbol{A}^{-1} - \frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\boldsymbol{e}\boldsymbol{e}^T\right)\\ &= \text{tr}\left(\boldsymbol{A}^{-1}\right) - \frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\text{tr}\left(\boldsymbol{e}\boldsymbol{e}^T\right)\\ &= \text{tr}\left(\boldsymbol{A}^{-1}\right) - \frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\text{tr}\left(\boldsymbol{e}^T\boldsymbol{e}\right)\\ &= \text{tr}\left(\boldsymbol{A}^{-1}\right) - \frac{\boldsymbol{e}^T\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}. \end{aligned} \tag{2.13}$$

A partir de (2.13) podemos notar que,

$$\begin{aligned} arg\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\text{tr}\left(\boldsymbol{V}\right) &= arg\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\left(\text{tr}\left(\boldsymbol{A}^{-1}\right) - \frac{\boldsymbol{e}^T\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\\ &= arg\max_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\left(\frac{\boldsymbol{e}^T\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\\ &= arg\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\left(\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{e}}\right). \end{aligned}$$

Pero, por (2.10)

$$\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\left(\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{e}}\right) = \lambda_n,$$

y se alcanza cuando $\boldsymbol{e} = \boldsymbol{e}_n$ es el eigenvector correspondiente al eigenvalor mas pequeño λ_n de $\boldsymbol{A}$. Es decir, el mínimo de la traza de $Cov\left(\boldsymbol{X^{(t)}}, \boldsymbol{X^{(t+1)}}\right)$ se alcanza con el eigenvector que corresponde al aigenvalor más pequeño de la matriz de precisión $\boldsymbol{A}$.

Lo anterior nos dice que al minimizar $\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{e}}$ estaríamos minimizando la dependencia, entrada a entrada, de dos pasos consecutivos de la cadena. De este modo, estaríamos dándole una interpretación al criterio de minimización propuesto por Christen et al. (2015).

2.4. Información Mutua marginal

Sean $\boldsymbol{X} = \boldsymbol{X^{(t)}}$ y $\boldsymbol{Y} = \boldsymbol{X^{(t+1)}}$ dos pasos consecutivos de la cadena, y denotemos por $X_i, Y_i, \quad i = 1, \ldots, n$, a los elementos de $\boldsymbol{X}$ y $\boldsymbol{Y}$, respectivamente. En esta sección

propondremos como medida de dependencia a la Información Mutua, pero ahora ya no de los vectores completos $\boldsymbol{X}$ e $\boldsymbol{Y}$, en lugar de eso, la obtendremos con la $i-$ésima entrada de $\boldsymbol{Y}$ y el vector completo $\boldsymbol{X}$, a la cual le llamaremos Información Mutua Marginal y escribiremos como $I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right)$. Después, se hará la minimización de la suma de una función creciente de estas cantidades. De este modo, estaríamos reduciendo la dependencia de cada entrada del nuevo punto generado $\boldsymbol{Y}$ con la del estado actual $\boldsymbol{X}$.

Notemos que,

$$\begin{aligned}
I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right) &= \int\int \pi\left(\boldsymbol{x}, y_i|\boldsymbol{e}\right) \log \frac{\pi\left(\boldsymbol{x}, y_i|\boldsymbol{e}\right)}{\pi\left(\boldsymbol{x}|\boldsymbol{e}\right)\pi\left(y_i|\boldsymbol{e}\right)}\, dy_i\, d\boldsymbol{x} \\
&= \int\int \pi\left(\boldsymbol{x}|\boldsymbol{e}\right)\pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right) \log \frac{\pi\left(\boldsymbol{x}|\boldsymbol{e}\right)\pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right)}{\pi\left(\boldsymbol{x}|\boldsymbol{e}\right)\pi\left(y_i|\boldsymbol{e}\right)}\, dy_i\, d\boldsymbol{x} \\
&= \int\int \pi\left(\boldsymbol{x}\right)\pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right) \log \frac{\pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right)}{\pi\left(y_i|\boldsymbol{e}\right)}\, dy_i\, d\boldsymbol{x} \\
&= \int\int \pi\left(\boldsymbol{x}\right)\pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right) \log \pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right)\, dy_i\, d\boldsymbol{x} \\
&\quad - \int\int \pi\left(\boldsymbol{x}\right)\pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right) \log \pi\left(y_i|\boldsymbol{e}\right)\, dy_i\, d\boldsymbol{x} \\
&= \int \pi\left(\boldsymbol{x}\right)\left(\int \pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right) \log \pi\left(y_i|\boldsymbol{e}, \boldsymbol{x}\right) dy_i\right) d\boldsymbol{x} \\
&\quad - \int \log \pi\left(y_i|\boldsymbol{e}\right)\left(\int \pi\left(\boldsymbol{x}, y_i|\boldsymbol{e}\right) d\boldsymbol{x}\right) dy_i. \qquad (2.14)
\end{aligned}$$

Para resolver (2.14) necesitamos conocer tanto la distribución de $Y_i|\boldsymbol{e}, \boldsymbol{x}$ como la de $Y_i|\boldsymbol{e}$. Para esto, tenemos que $\boldsymbol{Y} = \boldsymbol{x} + r\boldsymbol{e}$, entonces, $Y_i = x_i + re_i$, y como $r|\boldsymbol{e}, \boldsymbol{x} \sim N\left(-\frac{\boldsymbol{e}^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}, \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\right)$, se sigue que,

$$Y_i|\boldsymbol{e}, \boldsymbol{x} \sim N\left(x_i - \frac{\boldsymbol{e}^T\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i, \frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{e_i^2}\right), \qquad (2.15)$$

donde $\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{e_i^2}$ es la precisión. Y su función generadora de momentos esta dada por,

$$M_{Y_i|\boldsymbol{e},\boldsymbol{x}}\left(t\right) = \exp\left\{\left(x_i - \frac{\boldsymbol{e}^T\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i\right)t + \frac{\boldsymbol{e}_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\frac{t^2}{2}\right\}. \qquad (2.16)$$

A partir de (2.16), podemos obtener la función generadora de momentos de la variable $Y_i|\boldsymbol{e}$, y de este modo, conocer su distribución.

Haciendo uso de la ley de esperanzas iteradas y por (2.16) tenemos que,

$$\begin{aligned}
M_{Y_i|\boldsymbol{e}}(t) &= \mathbb{E}_{Y_i|\boldsymbol{e}}\left[e^{tY_i}\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{Y_i|\boldsymbol{e},\boldsymbol{X}}\left(e^{tY_i}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[M_{Y_i|\boldsymbol{e},\boldsymbol{X}}(t)\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[\exp\left\{\left(X_i - \frac{\boldsymbol{e}^T\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i\right)t + \frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\frac{t^2}{2}\right\}\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[\exp\left\{\left(X_i - \frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{X}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i + \frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{\mu}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i\right)t + \frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\frac{t^2}{2}\right\}\right] \\
&= \exp\left\{\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{\mu}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i t + \frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\frac{t^2}{2}\right\}\mathbb{E}_{\boldsymbol{X}}\left[e^{t\left(X_i - \frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{X}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i\right)}\right].
\end{aligned} \tag{2.17}$$

Con la función generadora de momentos, $M_{\boldsymbol{X}}(\cdot)$, de la distribución normal con media $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{A}$, y haciendo $\boldsymbol{d}_i^T = (0,\ldots,0,1,0,\ldots,0)$, podemos encontrar la esperanza de lado derecho de (2.17) de la siguiente forma,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}\left[e^{t\left(X_i - \frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{X}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i\right)}\right] &= \mathbb{E}_{\boldsymbol{X}}\left[\exp\left\{t\left(\boldsymbol{d}_i^T\boldsymbol{X} - e_i\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{X}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\right\}\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[\exp\left\{t\left(\boldsymbol{d}_i^T - e_i\frac{\boldsymbol{e}^T\boldsymbol{A}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\boldsymbol{X}\right\}\right] \\
&= M_{\boldsymbol{X}}\left[\left(\boldsymbol{d}_i - e_i\frac{\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t\right] \\
&= \exp\left\{\boldsymbol{\mu}^T\left(\boldsymbol{d}_i - e_i\frac{\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t\right. \\
&\quad \left. + \frac{1}{2}t\left(\boldsymbol{d}_i^T - e_i\frac{\boldsymbol{e}^T\boldsymbol{A}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\boldsymbol{A}^{-1}\left(\boldsymbol{d}_i - e_i\frac{\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t\right\} \\
&= \exp\left\{\left(\mu_i - e_i\frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t\right. \\
&\quad \left. + \frac{t^2}{2}\left(\boldsymbol{d}_i^T\boldsymbol{A}^{-1} - e_i\frac{\boldsymbol{e}^T}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\left(\boldsymbol{d}_i - e_i\frac{\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\right\}.
\end{aligned} \tag{2.18}$$

Pero, haciendo $\sigma_{ii} = \left(\boldsymbol{A}^{-1}\right)_{ii}$, vemos que,

$$\begin{aligned}\left(\boldsymbol{d}_i^T\boldsymbol{A}^{-1}-e_i\frac{\boldsymbol{e}^T}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\left(\boldsymbol{d}_i-e_i\frac{\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right) &= \boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{d}_i-e_i\frac{\boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\\ &\quad -e_i\frac{\boldsymbol{e}^T\boldsymbol{d}_i}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}+e_i^2\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{(\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e})^2}\\ &= \sigma_{ii}-e_i\frac{\boldsymbol{d}_i^T\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}-\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}+\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\\ &= \sigma_{ii}-\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}. \end{aligned} \tag{2.19}$$

Sustituimos (2.19) en (2.18),

$$\mathbb{E}_{\boldsymbol{X}}\left[e^{t\left(X_i-\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{X}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i\right)}\right]=\exp\left\{\left(\mu_i-e_i\frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t+\frac{t^2}{2}\left(\sigma_{ii}-\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\right\}. \tag{2.20}$$

Por último, sustituyendo (2.20) en (2.17) se sigue,

$$\begin{aligned}M_{Y_i|\boldsymbol{e}}(t) &= \exp\left\{\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{\mu}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_it+\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\frac{t^2}{2}\right\}\\ &\quad \exp\left\{\left(\mu_i-e_i\frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t+\frac{1}{2}\left(\sigma_{ii}-\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t^2\right\}\\ &= \exp\left\{\left(\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{\mu}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}e_i+\mu_i-e_i\frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)t+\left(\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}+\sigma_{ii}-\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\frac{t^2}{2}\right\}\\ &= \exp\left(\mu_it+\sigma_{ii}\frac{t^2}{2}\right).\end{aligned}$$

La cual corresponde a la función generadora de momentos de una variable aleatoria normal univariada, con media μ_i y varianza σ_{ii}. Por lo tanto,

$$Y_i|\boldsymbol{e}\sim N\left(\mu_i,\sigma_{ii}\right). \tag{2.21}$$

A continuación enunciamos una definición que servirá para resolver (2.14).

Entropía. Si X es una variable aleatoria con función de densidad f, la entropía de una variable aleatoria X esta definida como,

$$H(X) := -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx.$$

La entropía es un tipo especial de la información mutua. Si dos variables aleatorias son iguales entonces la información mutua es igual a la entropía.

Para el caso normal se tiene una forma explicita para la entropía.

Entropía caso Normal. Si $X \sim N(\mu, \sigma^2)$, entonces la entropía de X esta dada por,

$$H(X) = \frac{1}{2}\left(\log\left(2\pi e \sigma^2\right)\right), \tag{2.22}$$

aquí e representa el **número de Euler**, no hay confusión con las direcciones, la cuales se denotan en negritas, ni con los elementos de tales direcciones, las cuales tienen un subíndice.

Regresando a la ecuación (2.14), ya que sabemos que tanto la distribución de $Y_i|\boldsymbol{e}$ como la de $Y_i|\boldsymbol{e}, \boldsymbol{x}$ son normales, podemos usar la expresión (2.22) para obtener: por un lado,

$$\begin{aligned}
&\int \pi(\boldsymbol{x}) \left[\int \pi(y_i|\boldsymbol{e}, \boldsymbol{x}) \log \pi(y_i|\boldsymbol{e}, \boldsymbol{x})\, dy_i\right] d\boldsymbol{x} \\
&= \int \pi(\boldsymbol{x}) \left\{-\frac{1}{2}\left[\log\left(2\pi e \frac{e_i^2}{\boldsymbol{e}^T \boldsymbol{A} \boldsymbol{e}}\right)\right]\right\} d\boldsymbol{x} \\
&= -\frac{1}{2}\left[\log\left(2\pi e \frac{e_i^2}{\boldsymbol{e}^T \boldsymbol{A} \boldsymbol{e}}\right)\right] \int \pi(\boldsymbol{x})\, d\boldsymbol{x} \\
&= -\frac{1}{2}\log(2\pi e) - \frac{1}{2}\log\left(\frac{e_i^2}{\boldsymbol{e}^T \boldsymbol{A} \boldsymbol{e}}\right),
\end{aligned} \tag{2.23}$$

y por otro lado,

$$
\begin{aligned}
-\int \log \pi\left(y_i|\boldsymbol{e}\right)\left[\int \pi\left(\boldsymbol{x}, y_i|\boldsymbol{e}\right) d\boldsymbol{x}\right] dy_i &= -\int \pi\left(y_i|\boldsymbol{e}\right) \log \pi\left(y_i|\boldsymbol{e}\right) dy_i \\
&= \frac{1}{2}\left[\log\left(2\pi e \sigma_{ii}\right)\right] \\
&= \frac{1}{2}\log\left(2\pi e\right) + \frac{1}{2}\log\left(\sigma_{ii}\right). \qquad (2.24)
\end{aligned}
$$

Finalmente, sustituimos (2.23) y (2.24) en (2.14),

$$
\begin{aligned}
I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right) &= -\frac{1}{2}\log\left(2\pi e\right) - \frac{1}{2}\log\left(\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right) + \frac{1}{2}\log\left(2\pi e\right) + \frac{1}{2}\log\left(\sigma_{ii}\right) \\
&= \frac{1}{2}\left[\log\left(\sigma_{ii}\right) - \log\left(\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\right] \\
&= \frac{1}{2}\log\left(\sigma_{ii}\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{e_i^2}\right). \qquad (2.25)
\end{aligned}
$$

De la expresión (2.12) se puede ver que,

$$
\log\left(\frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right) = \log\left(\frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right) \leq \log\left(\sigma_i^2\right); \qquad \forall i = 1, \ldots, n,
$$

lo que garantiza que $I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right) \geq 0, \quad \forall i = 1, \ldots, n$, es decir, la información mutua esta bien definida.

Ahora, tomaremos funciones crecientes g_i's de las Informaciones Mutuas marginales, con $g_i\left(x\right) := -\frac{\sigma_{ii}}{\exp(2x)}, \quad \forall i = 1, \ldots, n$, y minimizaremos la suma de estas funciones, ya que el argumento que las minimice sera el mismo. Así,

$$\begin{aligned}\sum_{i=1}^{n} g_i\left(I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right)\right) &= \sum_{i=1}^{n} g_i\left(\frac{1}{2}\log\left(\sigma_{ii}\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{e_i^2}\right)\right)\\ &= \sum_{i=1}^{n}\left(-\frac{\sigma_{ii}}{\exp\left\{2\left[\frac{1}{2}\log\left(\sigma_{ii}\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{e_i^2}\right)\right]\right\}}\right)\\ &= -\sum_{i=1}^{n}\frac{\sigma_{ii}}{\sigma_{ii}\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{e_i^2}}\\ &= -\frac{1}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\sum_{i=1}^{n} e_i^2\\ &= -\frac{\boldsymbol{e}^T\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}.\end{aligned} \tag{2.26}$$

De (2.26) se sigue que,

$$\begin{aligned}arg\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\sum_{i=1}^{n} g_i\left(I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right)\right) &= arg\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\left(-\frac{\boldsymbol{e}^T\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\\ &= arg\max_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\left(\frac{\boldsymbol{e}^T\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\\ &= arg\min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1}\left(\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{e}}\right).\end{aligned}$$

Lo anterior nos dice que, al minimizar $\frac{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{e}}$, estaríamos minimizando la dependencia de cada entrada del nuevo punto generado $\boldsymbol{Y}$ con el estado actual de la cadena $\boldsymbol{X}$. Es decir, hemos encontrado otra interpretación al criterio de minimización propuesto por Christen et al. (2015).

CAPÍTULO 3

Propuesta de distribución de direcciones

En este capítulo se propone dos algoritmos MCMC, que son una generalización del muestreador de Gibbs estándar, el primero se usará para simular de la distribución Normal Multivariada y el segundo, bajo pequeñas modificaciones del primero, se utilizara para simular de una distribución NTM. La distribución de direcciones que proponemos, se basa en minimizar las informaciones mutuas marginales descritas en la Sección 2.4, de este modo, estaríamos reduciendo la dependencia de cada entrada del nuevo punto generado con la del estado actual.

3.1. Caso Normal

Supongamos que la distribución objetivo π es una Normal $n - variada$ con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{A}$, la cual es la inversa de la matriz de covarianzas, es decir, que el vector aleatorio del que queremos simular es $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, \boldsymbol{A})$.

Sean $\boldsymbol{X} = \boldsymbol{X}^{(t)}$ y $\boldsymbol{Y} = \boldsymbol{X}^{(t+1)}$ dos pasos consecutivos de la cadena, y denotemos por $X_i, Y_i, \quad i = 1, \ldots, n$, a los elementos de $\boldsymbol{X}$ y $\boldsymbol{Y}$, respectivamente. El soporte de la distribución de direcciones que proponemos son aquellas direcciones que minimicen las

Informaciones Mutuas Marginales que definimos en Sección 2.4.

Podemos reescribir a la Información Mutua marginal dada en (2.25) como,

$$I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right) = \frac{1}{2}\log\left(\sigma_{ii}\right) - \frac{1}{2}\log\left(\frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right),$$

donde $\boldsymbol{d}_i^T = (0, \ldots, 0, 1, 0, \ldots 0)$. Anteriormente probamos que $I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right) \geq 0$, $\forall i = 1, \ldots, n$, entonces,

$$\begin{aligned} arg \min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1} I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right) &= arg \min_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1} \left[\frac{1}{2}\log\left(\sigma_{ii}\right) - \frac{1}{2}\log\left(\frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}\right)\right] \\ &= arg \max_{\boldsymbol{e}\in\mathbb{R}^n:\|\boldsymbol{e}\|=1} \frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}. \end{aligned} \tag{3.1}$$

Si hacemos $\sigma_{ij} = \left(\boldsymbol{A}^{-1}\right)_{ij}$, $i, j = 1, \ldots, n$, del Lema 2.9 tenemos que,

$$\begin{aligned} \max_{\boldsymbol{e}\in\mathbb{R}^n} \frac{\left(\boldsymbol{e}^T\boldsymbol{d}_i\right)^2}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}} &= \boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{d}_i \\ &= \sigma_{ii}, \end{aligned}$$

y el máximo se obtiene cuando $\boldsymbol{e} = c\boldsymbol{A}^{-1}\boldsymbol{d}_i$, para cualquier constante $c \neq 0$. Si tomamos $c = \left\|\boldsymbol{A}^{-1}\boldsymbol{d}_i\right\|^{-1}$, entonces, podemos ver que la dirección $\boldsymbol{e}$ que optimiza (3.1), sería la $i - esima$ columna normalizada de la matriz de varianzas y covarianzas $\boldsymbol{A}^{-1}$.

Ahora, si denotamos a $\boldsymbol{e}_i = c\boldsymbol{A}^{-1}\boldsymbol{d}_i$, como la dirección optima que se obtiene al minimizar $I_{\boldsymbol{e}}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right)$ con respecto a $\boldsymbol{e}$, entonces $\forall j = 1, \ldots, n$, los valores de la

Información Mutua marginal con ésta dirección son,

$$\begin{aligned} I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right) &= \frac{1}{2}\log\left(\sigma_{jj}\right) - \frac{1}{2}\log\left(\frac{\left(\boldsymbol{e}_i^T\boldsymbol{d}_j\right)^2}{\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{e}_i}\right) \\ &= \frac{1}{2}\log\left(\sigma_{jj}\right) - \frac{1}{2}\log\left(\frac{\left(c\boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{d}_j\right)^2}{c\boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{A}c\boldsymbol{A}^{-1}\boldsymbol{d}_i}\right) \\ &= \frac{1}{2}\log\left(\sigma_{jj}\right) - \frac{1}{2}\log\left(\frac{\left(\boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{d}_j\right)^2}{\boldsymbol{d}_i^T\boldsymbol{A}^{-1}\boldsymbol{d}_i}\right) \\ &= \frac{1}{2}\log\left(\sigma_{jj}\right) - \frac{1}{2}\log\left(\frac{\sigma_{ij}^2}{\sigma_{ii}}\right) \\ &= -\frac{1}{2}\log\left(\frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}}\right) \\ &= -\frac{1}{2}\log\left(\rho_{ij}^2\right), \end{aligned} \tag{3.2}$$

donde ρ_{ij} representa la correlación entre las variables Z_i y Z_j, con $\boldsymbol{Z} \sim \pi$.

De (3.2) podemos ver que si $i = j$ entonces $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right) = 0$, es decir, que si nos movemos en la dirección $\boldsymbol{e}_i = c\boldsymbol{A}^{-1}\boldsymbol{d}_i$, estaremos haciendo que la $i-$ésima entrada del nuevo vector generado $\boldsymbol{Y}$ sea independiente del estado actual $\boldsymbol{X}$. Mas aún, si $\rho_{ij}^2 = 1$, es decir, que si la variable Z_i tiene correlación perfecta con Z_j, entonces al movernos en la dirección $\boldsymbol{e}_i = c\boldsymbol{A}^{-1}\boldsymbol{d}_i$, haremos que tanto la $i-$ésima como la $j-$ésima entrada de $\boldsymbol{Y}$ sean independientes de $\boldsymbol{X}$. Por otro lado, conforme la correlación ρ_{ij} se valla acercando a cero, la información mutua $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right)$ se ira a infinito. Es importante mencionar que esto último no nos indica que la $j - esima$ entrada de $\boldsymbol{Y}$ se haga "más dependiente" de $\boldsymbol{X}$, pero si nos dice que, al menos, no serán independientes. De modo que podemos tomar las columnas normalizadas del la matriz de varianzas y covarianzas como el soporte de la distribución de direcciones, las cuales son una base de $\mathbb{R}^n$. La pregunta que surge es, ¿cómo asignamos las probabilidades (pesos) de selección de cada dirección?

3.1.1. Seleccionando un conjunto de direcciones

De manera natural, abría que darle mayor peso a las direcciones $\boldsymbol{e}_i$ que hagan que las $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right)$, para $i \neq j$, sean pequeñas. Así, una forma de dar prioridad a las direcciones óptimas individuales $\boldsymbol{e}_i$, $i = 1, \ldots, n$, es en base a,

$$I_i := \sum_{j=1}^{n} I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right) = -\frac{1}{2}\sum_{j=1}^{n} \log\left(\rho_{ij}^2\right). \tag{3.3}$$

La distribución de dirección alternativa que proponemos, hace que la cadena se mueva en un número finito de direcciones. Para ello, haremos que las direcciones $\boldsymbol{e}$ sean las columnas normalizadas de la matriz de varianzas y covarianzas $\boldsymbol{A^{-1}}$, así $\boldsymbol{e} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$. La $i-$ésima dirección se seleccionará con probabilidad proporcional a I_i^{-1}. Luego,

$$h\left(\boldsymbol{e}_i\right) = kI_i^{-1} \tag{3.4}$$

donde $k = \left(\sum_{i=1}^{n} I_i^{-1}\right)^{-1}$. De este modo, estaríamos dándole más pesos a las direcciones $\boldsymbol{e}_i$ ´s que hagan que las $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right)$´s sean pequeñas, es decir, darle más peso a las direcciones que hacen "más independientes" al resto de las entradas.

Pero habrá un problema al utilizar las expresión (3.3) cuando al menos dos variables, Z_i y Z_j, sean independientes, es decir, cuando $\rho_{ij}^2 = 0$ para alguna i y para alguna j, ya que, $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right)$ sería infinito, y en consecuencia se le daría pesos igual a cero, tanto a la dirección $\boldsymbol{e}_i$ como a la $\boldsymbol{e}_j$, y la cadena resultante no seria irreducible.

Para corregir este problema se pensaron en tres posibilidades. La primera, que surge de manera natural, consiste en simular de forma independiente a los bloques de variables con correlación cero. El segundo enfoque, trata de hacer solamente una modificación a la expresión dada en (3.3), de la siguente manera,

$$I_i := \sum_{j \in \boldsymbol{A}_i} I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right) = -\frac{1}{2}\sum_{j \in \boldsymbol{A}_i} \log\left(\rho_{ij}^2\right). \tag{3.5}$$

donde $\boldsymbol{A}_i = \{k : \rho_{ik}^2 \neq 0, \quad k = 1, \ldots, n\}$. Esto se pensó, ya que, si una variable (Z_i) es independiente de otra (Z_j), entonces la variable Z_j no tendría por que atribuir al peso

que se le de a la dirección $\boldsymbol{e}_i$, la cual hace independiente a la entrada $i-$ésima del nuevo vector.

La última opción que consideramos, y que es por la que optamos por simplicidad, consiste en sumar una cantidad considerablemente pequeña (ε) dentro del logaritmo de la definición dada en (3.3), para que, en caso de haber al menos dos pares de variables independientes, no resulten informaciones mutuas marginales infinito. De esta manera, eliminaríamos el problema de dar pesos cero a ciertas direcciones. Así, en lugar de usar (3.3) en las probabilidades de selección de cada dirección dadas en (3.4), usaremos

$$I_i = -\frac{1}{2}\sum_{j=1}^{n} \log\left(\rho_{ij}^2 + \varepsilon\right). \tag{3.6}$$

El algoritmo propuesto, considerando (3.6) en la distribución h dada en (3.4), es el siguiente:

Algoritmo 3 Gibbs Direccional Óptimo: Normal Multivariada

Dado $\boldsymbol{X}^{(t)} = \boldsymbol{x}$,

1. Se genera $\boldsymbol{e} \sim h(\boldsymbol{e})$.
2. Se genera una longitud $r|\boldsymbol{e}, \boldsymbol{x} \sim N\left(-\frac{\boldsymbol{e}^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}, \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\right)$.
3. Hacemos $\boldsymbol{X}^{(t+1)} = \boldsymbol{x} + r\boldsymbol{e}$.

Ya hemos visto que si nos movemos en la dirección $\boldsymbol{e}_i = c\boldsymbol{A}^{-1}\boldsymbol{d}_i$, entonces la información $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_i}, \pi_{\boldsymbol{X}}\pi_{Y_i}\right) = 0$, más aun, si $\rho_{ij}^2 \approx 1$ (si Z_i esta fuertemente correlacionada con Z_j), entonces $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right) \approx 0$. Por lo anterior, esperaríamos que el Algoritmo 3 funcionará mejor cuando las variables del vector $\boldsymbol{Z}$ tengan alta correlación. Ya que con esto, con cualquier dirección elegida $\boldsymbol{e}_i$, se tendrá que $I_{\boldsymbol{e}_i}\left(\pi_{\boldsymbol{X}Y_j}, \pi_{\boldsymbol{X}}\pi_{Y_j}\right) \approx 0$, esto es, que en cualquier dirección que nos movamos, las entradas del vector generado $\boldsymbol{Y}$ serán casi independientes del vector actual $\boldsymbol{X}$.

3.2. Normal Truncada

Supongamos que tenemos una distribución Normal $n - variada$ con matriz de precisión $\boldsymbol{A}$ y vector de medias $\boldsymbol{\mu}$, pero con soporte truncado en $x_i \in (a_i, b_i)$, $-\infty \leq a_i \leq b_i \leq \infty$, $i = 1, \ldots, n$. La función de densidad de la distribución NTM se puede expresar como:

$$\pi\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{a}, \boldsymbol{b}\right) = \frac{\exp\left\{\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T \boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\}}{\int_{\boldsymbol{a}}^{\boldsymbol{b}} \exp\left\{\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T \boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\} d\boldsymbol{x}},$$

para $\boldsymbol{a} \leq \boldsymbol{x} \leq \boldsymbol{b}$ y 0 de otro modo.

Una forma muy simple para muestrear de la NTM sería generar valores de la Normal Multivariada, por ejemplo con el Algoritmo 3, y solamente aceptar aquellas muestras que estén dentro del soporte, es decir, aquellas que satisfacen $x_i \in (a_i, b_i)$, $\forall i = 1, \ldots, n$. Este método, el cual se conoce como Rejection Samplig, trabaja bien cuando la tasa de aceptación es razonablemente alta, sin embargo, resulta ser muy ineficiente cuando la tasa de aceptación es baja, como en el caso de alta dimensión y/o cuando el soporte esta estrechamente acotado.

La mayoría de los métodos disponibles para muestrear de la distribución NTM están basados en el muestreador de Gibbs, el cual es simple de usar y tiene la ventaja de aceptar todas las propuestas generadas y, por lo tanto, no se ve afectada por tasas de aceptación pobres. El inconveniente que se tiene con las muestras producidas por el muestreador de Gibbs es que no son independientes, el grado de correlación depende tanto de la matriz de varianzas y covarianzas como de la dimensionalidad.

El muestreador de Gibbs estándar, muestrea de las distribuciones condicionales univariadas $\pi\left(x_i|\boldsymbol{x}_{-i}\right) = \pi\left(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right)$, las cuales son Normales Truncadas univariadas (Horrace, 2005; Kotecha y Djuric, 1999). Por lo tanto, el muestreador de Gibbs requiere simular de una distribución Normal Truncada univariada (NT) lo cual se puede hacer de una forma simple y eficiente (Kotecha y Djuric, 1999). En Damien y Walker (2001) presentan un esquema interesante al utilizar un slice sampler. Este algoritmo, introduce una variable latente, en el esquema de un muestreador de Gibbs en un

espacio aumentado por una variable, que convierte la distribución condicional completa en una distribución uniforme.

En esta sección mostraremos que el Algoritmo 3 se puede utilizar, bajo pequeñas modificaciones, para muestrear de la distribución Normal Truncada Multivariada.

Como se mencionó anteriormente, podemos muestrear de una Normal Multivarada seleccionando una dirección $\boldsymbol{e}$ y una longitud r, los cuales producen $\boldsymbol{X}^{(t+1)} = \boldsymbol{x}^{(t)} + r\boldsymbol{e}$. Notemos que, para la NTM, requerimos que $\boldsymbol{a} \leq \boldsymbol{x}^{(t)} + r\boldsymbol{e} \leq \boldsymbol{b}$; entonces $a_i \leq x_i + re_i \leq b_i, \quad \forall\, i \in \{1, \ldots, n\}$. Esto lleva a tener restricciones sobre r de la forma,

$$\begin{cases} \frac{a_i - x_i}{e_i} \leq r \leq \frac{b_i - x_i}{e_i} & \forall e_i > 0, \\ \frac{b_i - x_i}{e_i} \leq r \leq \frac{a_i - x_i}{e_i} & \forall e_i < 0. \end{cases}$$

No debemos preocuparnos para el caso $e_i = 0$ ya que no pone restricciones sobre r, esto debido a que la $i-$ésima coordenada no cambiaría. Tomando $r \in (c, d)$, con

$$c = \max_{i \in \{1,\ldots,n\}} \left(\left\{ \frac{a_i - x_i}{e_i} : e_i > 0 \right\} \cup \left\{ \frac{b_i - x_i}{e_i} : e_i < 0 \right\} \right), \tag{3.7}$$

$$d = \min_{i \in \{1,\ldots,n\}} \left(\left\{ \frac{a_i - x_i}{e_i} : e_i < 0 \right\} \cup \left\{ \frac{b_i - x_i}{e_i} : e_i > 0 \right\} \right), \tag{3.8}$$

garantizamos que $\boldsymbol{a} \leq \boldsymbol{X}^{(t+1)} \leq \boldsymbol{b}$. Ya que sabemos que $r|\boldsymbol{e}, \boldsymbol{x}^{(t)}$ sigue una distribución Normal, entonces la restricción $r \in (c, d)$ implica que $r|\boldsymbol{e}, \boldsymbol{x}^{(t)}, c, d$ sigue una distribución NT.

Para simular de $r \sim NT\,(\mu, \sigma^2, a, b)$, usaremos el método de la transformada inversa, como se describe en Chib (2001). Esto es,

$$r = \mu + \sigma\Phi^{-1}\left(\Phi\left(\frac{a-\mu}{\sigma}\right) + U\left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right)\right),$$

donde Φ es la función de distribución de la normal estándar y Φ^{-1} es su inversa, U es una variable aleatoria uniforme en $(0, 1)$.

Considerando la distribución de direcciones h dada en (3.4), el algoritmo que proponemos para simular de la distribución NTM procede de la siguiente manera:

Algoritmo 4 Gibbs Direccional Óptimo: Normal Truncada Multivariada

Dado $\boldsymbol{X}^{(t)} = \boldsymbol{x}$,

1. Proponemos una dirección $\boldsymbol{e}$ de la distribución h.
2. Proponemos una logitud r de una $NT\left(\mu_r, \tau_r, c, d\right)$ con media $\mu_r = -\frac{\boldsymbol{e}^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}$, precisión $\tau_r = \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$, c y d como en (3.7) y (3.8).
3. Hacemos $\boldsymbol{X}^{(t+1)} = \boldsymbol{x} + r\boldsymbol{e}$.

En el Capítulo 4 se realizan experimentos con el Algoritmo 4, al que le llamaremos Algoritmo ODG, para ver su comportamiento.

CAPÍTULO 4

Experimentos numéricos

4.1. Caso Normal

Los experimentos a realizar, se harán considerando una Normal $n-$variada con vector de medias $\boldsymbol{\mu} = \left(\sqrt{1/n}, \ldots, \sqrt{1/n}\right)$ y matriz de precisión $\boldsymbol{A}$, para diferentes dimensiones $n = 2, 3, 10, 20, 50, 100$.

La matriz de precisión la obtenemos como $\boldsymbol{A} = \boldsymbol{P}^T \boldsymbol{\Lambda} \boldsymbol{P}$. Aquí, $\boldsymbol{P}$ es una matriz aleatoria ortonormal, generada a partir de la descomposición QR de una matriz con entradas aleatorias uniformes; $\boldsymbol{P}$ representa una base ortonormal de eigenvectores de $\boldsymbol{A}$. Mientras que $\boldsymbol{\Lambda}$ es una matriz diagonal con los eigenvalores $\lambda_i = \sigma_i^{-2}$. La desviación estándar en cada dirección principal (eigen) es $\lambda_i^{-1/2} = \sigma_i = i^{-\alpha/n}$. Éstas representan desviaciones estándar decrecientes e incrementan inversamente conforme α incrementa; $\alpha = 0$ resulta en una distribución no correlacionada. Desviaciones estándar más contrastantes resulta en distribuciones más correlacionadas, ver Figura 4.1.

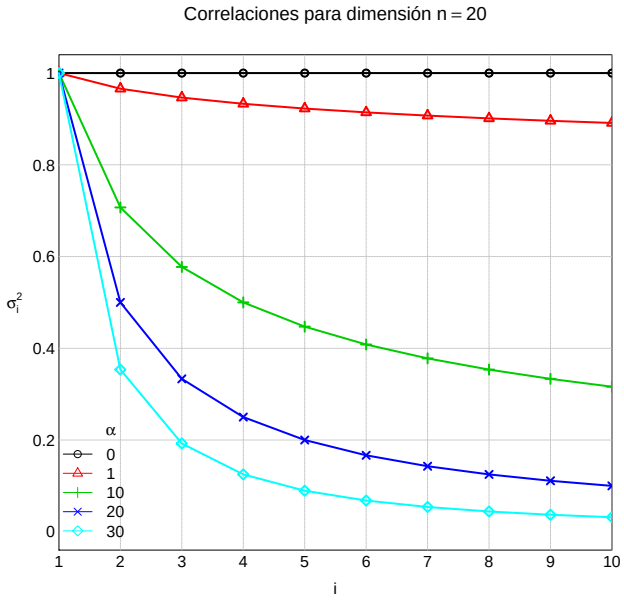


Figura 4.1: Desviaciones estándar progresivamente más contrastantes.

Para cada combinación de n y α calculamos el Integrated Autocorrelation Time (IAT) de la cadena resultante. El IAT da un índice de la eficiencia de una cadena o de un método de simulación MCMC (Roberts y Rosenthal, 2001); se usa para estimar el numero de simulaciones (τ) que debemos descartar de la cadena para obtener una muestra pseudo-independiente, ver Geyer (1992) para su estimación. Por lo que quisiéramos obtener valores bajos del IAT.

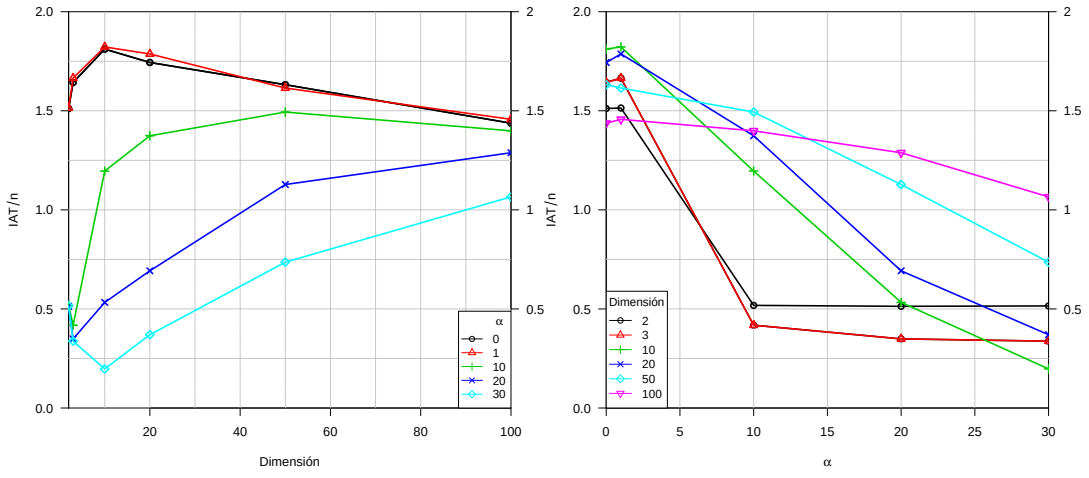
En la Tabla 4.1 se muestra el IAT dividido entre la dimensión (IAT/n) de las muestras obtenidas de la distribución Normal Multivariada para diferentes niveles de correlación α y diferentes dimensiones n.

En la Figura 4.2 se muestra el IAT/n contra la dimensión (lado izquierdo) y el IAT/n contra el nivel α (lado derecho), sacados de la Tabla 4.1. De la Figura 4.2a vemos que cuando hay poca correlación (α chico), el IAT/n se mantiene casi constante con respecto a la dimensión. Mientras que conforme aumenta la dimensión también crece el IAT/n, pero parece estabilizarse. De la Figura 4.2b podemos ver que conforme

	Dimensión (n)					
α	2	3	10	20	50	100
0	1.511	1.643	1.811	1.744	1.632	1.437
1	1.514	1.665	1.822	1.787	1.615	1.457
10	0.518	0.418	1.196	1.374	1.494	1.399
20	0.513	0.349	0.533	0.692	1.128	1.288
30	0.515	0.337	0.197	0.370	0.736	1.066

Tabla 4.1: IAT/n para muestras de la Normal Multivariada. Cada cantidad es el promedio de 30 cadenas de longitud 10000.

aumenta la correlación (aumenta α), el IAT/n va disminuyendo, y la velocidad con que disminuye depende de la dimensión. Como se había previsto, el algoritmo funciona mejor entre más correlacionada este la distribución objetivo.



(a) **(b)**

Figura 4.2: Comportamiento del IAT/n: **(a)** contra la dimensión, para cada nivel α y **(b)** contra α, para diferentes dimensiones. Caso Normal Multivarada.

Podemos ver al Gibbs Direccional Óptimo como una caminata aleatoria. Los IAT resultantes para todas las combinaciones de n y α son menores al IAT óptimo teórico $(IAT/n \approx 3)$ para una caminata aleatoria escalda óptima, como lo explican en Roberts y Rosenthal (2001). Más aún, se tiene que $IAT/n < 2$ para todas las combinaciones realizadas, lo cuál nos da un índice de la eficiencia del algoritmo propuesto.

4.2. Normal Truncada

Los experimentos que haremos en esta sección se harán considerando una Normal Truncada Multivariada $n-$dimensional, con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{A}$, obtenidos como en la Sección 4.1. El soporte estará restringido a $z_i \geq 0$, es decir, todas las entradas serán positivas. Es importante notar que el soporte truncado permanece no acotado. Nosotros no discutiremos como muestrear cuando el soporte esta estrechamente acotado, por ejemplo, un soporte de la forma $a_i < z_i < b_i$ con $-\infty < a_i < b_i < \infty$, para toda $i = 1, \ldots, n$. En el caso bidimensional, este soporte será un rectángulo, y podría representar básicamente un muestreo uniforme sobre el rectángulo, el cual es, en esencia, un problema de simulación diferente.

Comenzaremos analizando el caso $n = 2$ para poder comparar con los resultados obtenidos en Christen et al. (2015). Ellos comparan su algoritmo, al que le llamaremos ODG2, con los muestreadores de Gibbs presentados en Kotecha y Djuric (1999) y en Damien y Walker (2001), usando muchos ejemplos de la distribución NTM. Nosotros nos referiremos a estos dos últimos algoritmos como KD y DW, respectivamente. Haremos 5000 iteraciones con el Algoritmo ODG propuesto, comenzando en μ, así no será necesario el proceso de calentamiento (burn-in).

En la Figura 4.3 tenemos muestras de la distribución objetivo para $\alpha = 0, 5, 10, 20$. Puntos en negro corresponden a muestras de la distribución Normal bivaridad completa, mientras que los puntos en azul corresponden a las muestras de la distribución Normal Truncada bivarada, resultado de usar el Algoritmo ODG. Como se menciona anteriormente, con forme α incrementa, la correlación aumenta, y hace que las regiones sean

más difíciles de simular. Podemos ver que el Algoritmo ODG funciona bien en todos los casos, explora todo el soporte. Mientras que para $\alpha = 10$, tanto DW como KD tienen dificultades para explorar toda la región de interés. Para $\alpha = 20$, las muestras generadas por KD y DW se concentran en una región pequeña, ver Figura 4.4.

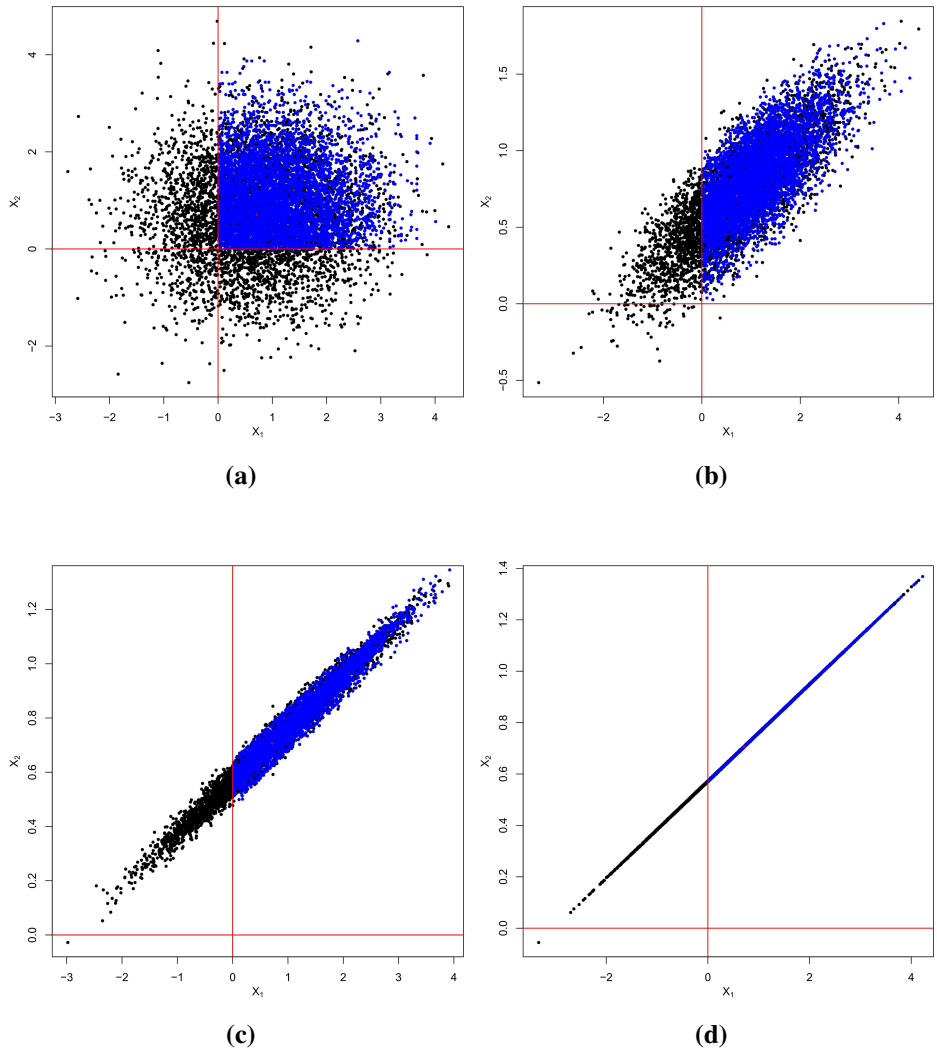


Figura 4.3: Muestras de la Normal bivariada completa (puntos negros) y de la NT bivariada (puntos azules), con 5000 iteraciones, para **(a)** $\alpha = 0$, **(b)** $\alpha = 5$, **(c)** $\alpha = 10$ y **(d)** $\alpha = 20$.

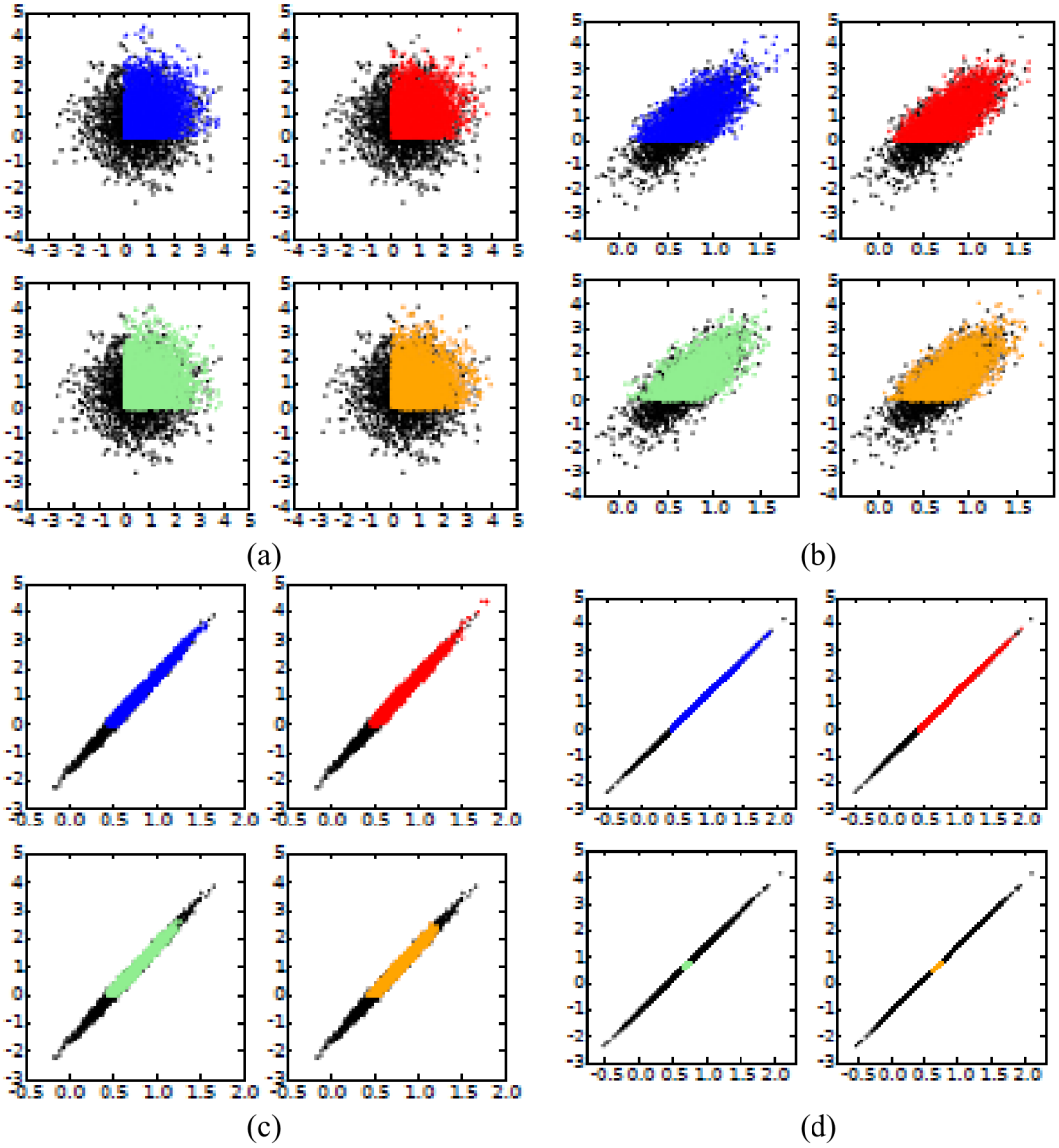


Figura 4.4: Muestras de la Normal bivariada completa (puntos negros) y de la NT bivariada con el Algoritmo ODG1 (azul), ODG2 (rojo), KD (verde), DW (naranja), con 5000 iteraciones para **(a)** $\alpha = 0$, **(b)** $\alpha = 5$, **(c)** $\alpha = 10$ y **(d)** $\alpha = 20$. Fuente: Christen et al. (2015).

Como se menciono anteriormente, el IAT es un indicador de la eficiencia del Algoritmo ODG. Sin embargo, el IAT no está completamente estudiado para cadenas no reversibles, como es el caso de los algoritmos KD y DW, ya que estos son muestreadores de Gibbs sistemáticos. En lugar del IAT, en Christen et al. (2015), calculan el Tamaño de Muestra Efectiva (TME) (Liu, 2008) y estiman τ como m/TME, donde m es la longitud de la cadena. En la Tabla 4.2 se reportan los resultados obtenidos en Christen et al. (2015) y los que se obtuvieron con el Algoritmo ODG, propuesto en este trabajo, para el caso bidimensional.

	Correlación (α)			
τ	0	5	10	20
ODG	3.0	2.5	1.1	1.0
KD	1.0	3.5	90.1	538.8
DW	1.4	3.8	100.5	467.7

Tabla 4.2: Número de simulaciones (τ) requeridas para obtener una muestra pseudoindependiente. Cada cantidad es el promedio de 30 cadenas de longitud 5000. Normal Truncada bivariada. Los datos de los algoritmos KD y DW, fueron tomados de Christen et al. (2015).

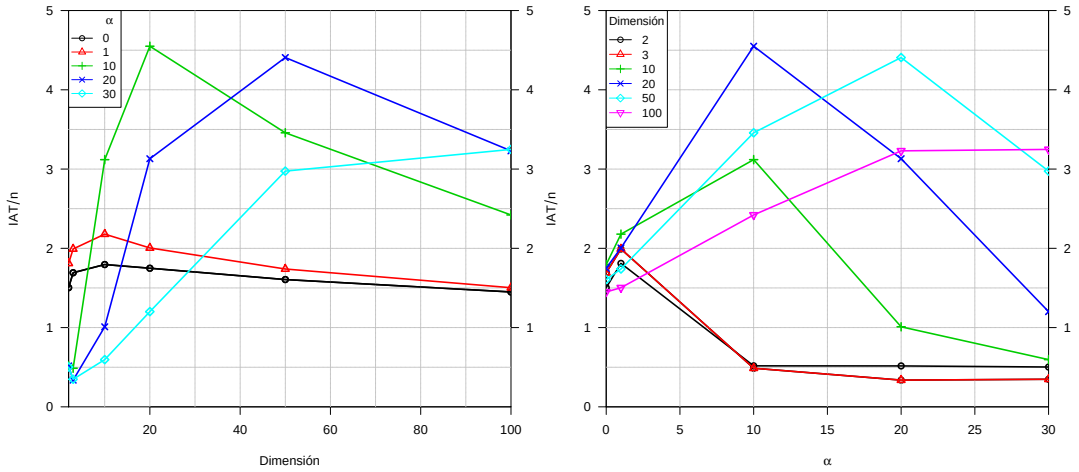
En la Tabla 4.2 podemos ver que, para el caso independiente $(\alpha = 0)$, los algoritmo KD y DW son más eficientes que ODG, ya que τ es menor. Si embargo, conforme incrementa la correlación, el algoritmo ODG tiene un mejor desempeño que los muestreadores de Gibbs sistemáticos, se observa una diferencia muy grande cuando la correlación es alta, siendo mucho mejor ODG. Cabe destacar, que aún para bajas correlaciones, el rendimiento del algoritmo ODG es comparable con los de KD y DW.

En la Tabla 4.3 se muestra el IAT dividido entre la dimensión (IAT/n) de la distribución NTM para diferentes niveles de correlación α y diferentes dimensiones n, obtenidos con el algoritmo ODG.

	Dimensión (n)					
α	2	3	10	20	50	100
0	1.504	1.692	1.795	1.749	1.606	1.450
1	1.810	1.992	2.179	2.005	1.739	1.503
10	0.517	0.487	3.119	4.550	3.458	2.421
20	0.516	0.337	1.010	3.130	4.407	3.230
30	0.502	0.348	0.596	1.201	2.974	3.248

Tabla 4.3: IAT/n para muestras de la distribución NTM. Cada cantidad es el promedio de 30 cadenas de longitud 10000.

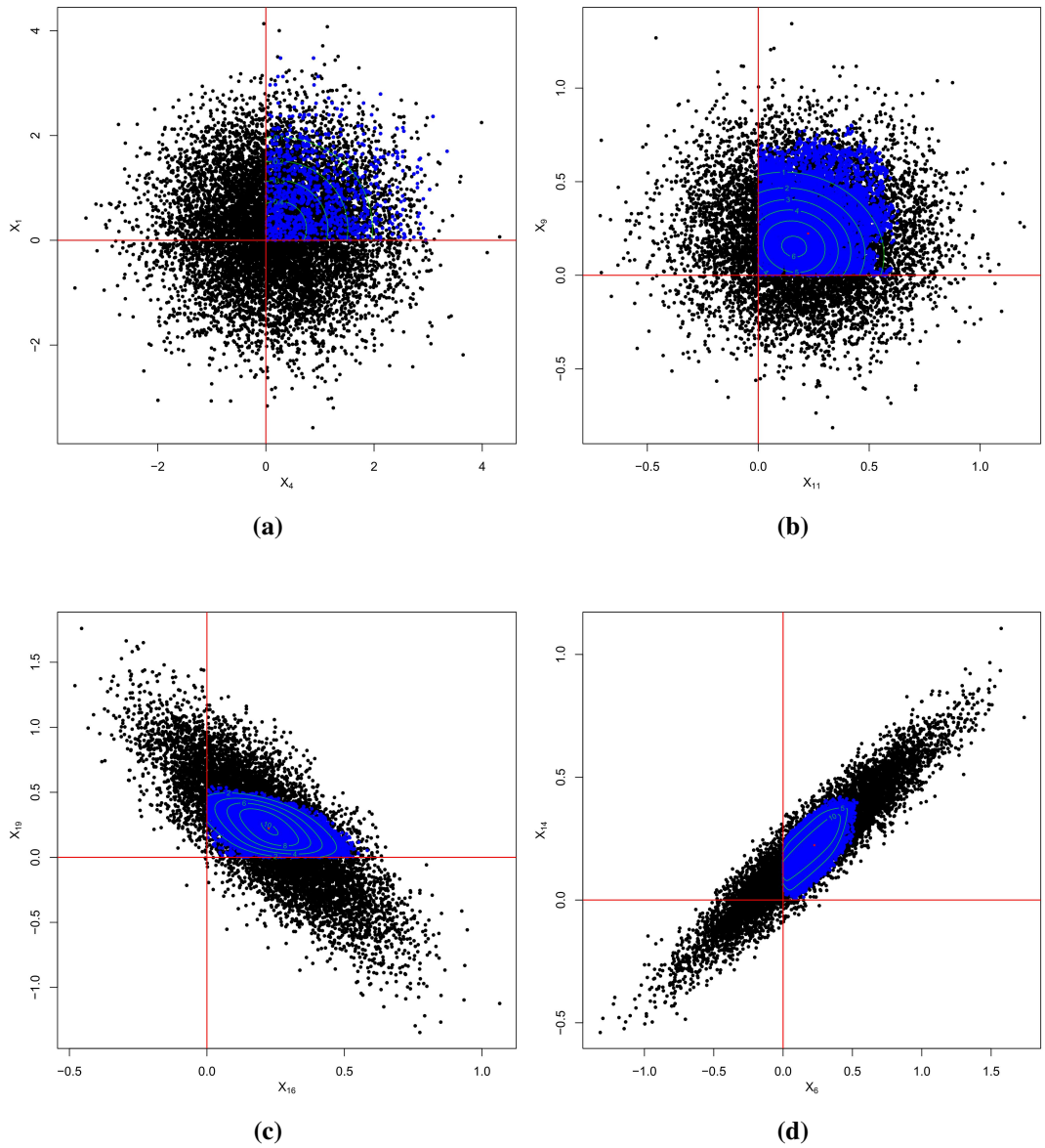
Para dar una mejor interpretación de la Tabla 4.3, en la Figura 4.5 se muestra el IAT/n contra la dimensión (lado izquierdo) y el IAT/n contra el nivel α (lado derecho). De la Figura 4.5a podemos ver que cuando la correlación es muy baja, el IAT/n no se ve afectado por la dimensión. También se observa que el comportamiento del algoritmo mejora si la distribución esta fuertemente correlacionada. Cabe mencionar, que para un mismo nivel α no se obtendrá el mismo orden de correlación para diferentes tamaños de muestras. Para un α fijo, la correlación disminuye conforme crece la dimensión. Por ejemplo, para $n = 2$ con un $\alpha = 10$, la distribución estará fuertemente correlacionada, mientras que para $n = 100$, con ese mismo nivel α, tendremos una distribución casi independiente, y para $n = 10$ tendremos una correlación "intermedia". Es por eso que se observa ese comportamiento en la Figura 4.5b: vemos que cuando la correlación es muy baja, el algoritmo funciona muy bien; conforme aumenta α, vemos al principio que el IAT/n es menor para dimensión pequeña y dimensión alta, esto se debe a que cuando la dimension es pequeña, la distribución se vuelve altamente correlacionada para un α pequeño, mientras que para dimensión alta $(n = 100)$ la distribución es casi independiente (necesitamos un valor de α muy grande para que la distribución se vuelva altamente correlacionada). Por esta razón el IAT/n es pequeño al inicio (caso independiente) y comienza a subir (correlación media), hasta cierto punto, donde empieza a hacerse pequeño nuevamente (correlación alta).



(a) **(b)**

Figura 4.5: Comportamiento del IAT/n: **(a)** contra la dimensión, para cada nivel α y **(b)** contra α, para diferentes dimensiones. Caso NTM.

En la Figura 4.6 se presentan simulaciones de la distribución NTM, obtenidas con el Algoritmo ODG, para el caso de dimensión $n = 20$, con diferentes niveles de correlación $\alpha = 0, 15, 30, 50$. Se hicieron 10000 iteraciones, comenzando en μ, de este modo no fue necesario el burn-in. Puntos en negro, corresponden a muestras de la Normal multivariada completa, mientras que los puntos en azul, corresponden a las muestras de la distribución NTM, resultado de usar el Algoritmo ODG. Para cada nivel α se seleccionaron, de forma aleatoria, parejas de índices, y con la función `marginal2()` del paquete **tmvtnorm**, implementado en R, se obtiene la función de densidad bivariada de la distribución Normal Truncada Multivariada, y con esta se sacan los contornos (curvas en verdes). De la Figura 4.6 podemos ver que el Algoritmo ODG funciona bien en todos los casos, explora toda la región de interés. Se realizaron muchos experimentos para ver que en efecto el algoritmo propuesto estuviera simulando de la distribución objetivo, y todos los resultados fueron satisfactorios.



(a) **(b)**

(c) **(d)**

Figura 4.6: Muestras de la distribución NM completa (puntos negros) y de la NT de dimensión $n = 20$ (puntos azules), para **(a)** $\alpha = 0$, **(b)** $\alpha = 15$, **(c)** $\alpha = 30$ y **(d)** $\alpha = 50$; las curvas en verdes representan los contornos de la densidad truncada bivariada.

CAPÍTULO 5

Discusión

En este trabajo de tesis, se dieron dos justificaciones de porque funciona la propuesta de distribución de direcciones dada en Christen et al. (2015). Con la primera, vemos que las direcciones propuestas son aquellas que reducen la dependencia, entrada a entrada, de dos pasos consecutivos de la cadena. Mientras que con la segunda justificación, se observa que se reduce la dependencia de cada entrada del nuevo vector generado, $\boldsymbol{Y}$, con todo el vector actual de la cadena, $\boldsymbol{X}$.

En la tesis, también se propuso una nueva distribución de direcciones, para generar muestras de una Normal Truncada Multivariada; esta basada en minimizar las informaciones mutuas de cada entrada de $\boldsymbol{Y}$ con el del estado actual $\boldsymbol{X}$, de este modo se esta reduciendo la dependencia de cada entrada de $\boldsymbol{Y}$ con la de $\boldsymbol{X}$. La distribución de dirección propuesta, tiene como soporte a las columnas normalizadas de la matriz de varianzas y covarianzas $\boldsymbol{A^{-1}}$; damos más pesos a aquellas direcciones que hacen que las entradas de $\boldsymbol{Y}$ sean lo menos dependientes del estado actual $\boldsymbol{X}$.

El algoritmo propuesto nos permite trabajar con distribuciones objetivos en altas dimensiones. La principal ventaja es que la distribución de direcciones utilizada es discreta y simular de ella es sumamente sencillo.

Vimos que el algoritmo se vuelve más eficiente cuando la densidad objetivo esta fuertemente correlacionada, y estos casos suelen ser los más complicados y es donde los muestreadores de Gibbs estándar resultan muy ineficientes, además, con frecuencia son de interés en muchas casos de estudio.

La dificultad que le vemos al algoritmo propuesto es que requiere tanto de la matriz de precisión $\boldsymbol{A}$ como de la matriz de varianzas y covarianzas $\boldsymbol{A}^{-1}$. Por un lado tenemos que las direcciones $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ son las columnas normalizadas de $\boldsymbol{A}^{-1}$ y, por otro lado, la media y precisión de la distribución de la longitud r están dadas por $\mu_r = -\frac{\boldsymbol{e}^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})}{\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}}$ y $\tau_r = \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$, respectivamente.

Referencias

Chib, S. (2001). Markov chain monte carlo methods: computation and inference. *Handbook of econometrics*, *5*, 3569–3649. (Citado en página 27.)

Christen, J. A., Fox, C., y Santana-Cibrian, M. (2015). Optimal direction gibbs sampler for truncated multivariate normal distributions. *Communications in Statistics-Simulation and Computation*, (just-accepted). (Citado en páginas III, III, V, VI, 2, 3, 4, 6, 7, 8, 9, 14, 20, 32, 34, 35 y 39.)

Damien, P. y Walker, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, *10*(2), 206–215. (Citado en páginas 26 y 32.)

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, *7*(4), 473–483. (Citado en página 30.)

Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis*, *94*(1), 209–221. (Citado en página 26.)

Johnson, R. A. y Wichern, D. W. (2014). *Applied multivariate statistical analysis*. Pearson Education Limited Essex. (Citado en páginas 8 y 12.)

Kotecha, J. H. y Djuric, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 3, (pp. 1757–1760). IEEE. (Citado en páginas 26 y 32.)

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media. (Citado en páginas 5, 6, 10 y 35.)

Robert, C. y Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media. (Citado en página 6.)

Roberts, G. O. y Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, *16*(4), 351–367. (Citado en páginas 30 y 32.)