ÍNDICE

Objetivos

El objetivo principal del trabajo es la creación de un programa capaz de clasificar información multidimensional de manera no supervisada mediante la adaptación de un modelo de mecánica estadística, conocido como el modelo de Potts.

Modelo de Potts

Introducción

El modelo de Potts[1] es una generalización del modelo propuesto por Ising[2][3] para el estudio de la dinámica de las transiciones de fase. Dicho modelo permite explicar como las interacciones entre los componentes fundamentales de un sistema dan origen a un comportamiento colectivo y ordenado. En particular, puede ser utilizado para obtener los distintos dominios magnéticos de un material ferromagnético cuando es sometido a un campo magnético externo y se varía su temperatura.

El modelo consiste en una red de dimensión d conformada por N espines s_i tal que $s_i = 1,\ 2,\ \ldots\ q$. La configuración del sistema en un determinado instante n queda descripta por la combinación de todos los valores de espín $S_n = \{s_1, s_2,\ \ldots\ s_N\}$, siendo $\xi = \{S_1,\ S_2,\ \ldots\ S_{q^N}\}$ el espacio conformado por todas las configuraciones posibles. La energía del sistema en ausencia de campo magnético externo viene dada por

$$H(S) = \sum_{ij} J_{ij}\,(1 - \delta_{s_i s_j}) \ , \qquad (1)$$

donde $H(S)$ es el Hamiltoniano sobre una configuración de interés S y J_{ij} representa la magnitud de interacción entre dos espines s_i y s_j. Nótese que la contribución de un par de espines a H es nula si $s_i = s_j$; esto se debe a que espines paralelos "ayudan" al sistema a llegar a su estado fundamental de mínima energía. Para obtener el valor de una cantidad física de interés Φ sobre un espacio ξ a una determinada temperatura T, es necesario calcular su valor medio termodinámico

$$< \Phi > \ = \ \sum_{S} \Phi(S)\, P(S) \ , \qquad (2)$$

donde

$$P(S) = \frac{1}{Z}\, e^{-\frac{H(S)}{T}} \qquad (3)$$

es la función de distribución de Boltzmann y $Z = \sum_S e^{-\frac{H(S)}{T}}$ es la función de partición utilizada para normalizar $P(S)$. Una de las magnitudes físicas principales presentes en el modelo de Potts es la magnetización del sistema,

$$m = \frac{q\, N_{max}(S) - N}{(q-1)\, N} \ , \qquad (4)$$

con $N_{max}(S) = max\ \{ N_1(S), N_2(S), \ldots N_q(S) \}$, siendo $N_\beta(S) = \sum_i \delta_{s_i,\beta}$ la cantidad de espines en el estado β. La magnetización provee información acerca de la distribución de espines paralelos en relación a la cantidad de espines totales. Mientras más espines en un mismo estado haya, más fuerte será la magnetización, ya que hay un sentido "preferido" por sobre todos los demás. Otra magnitud física importante relacionada con la magnetización es la susceptibilidad magnética,

$$\chi = \frac{N}{T}\sigma_m^2 \ , \qquad (5)$$

donde σ_m^2 es la varianza de la magnetización entre las distintas configuraciones de espín, a una determinada temperatura T. Por último, se define la correlación espín-espín como

$$G_{ij} = < \delta_{s_i\, s_j} > \ , \qquad (6)$$

que representa la probabilidad de que dos espines s_i y s_j estén alineados a una determinada temperatura T.

Interpretación física

Cuando los espines están en una red ordenada y las interacciones entre vecinos cercanos es constante, $J_{ij} = J$, el sistema se denomina homogéneo. Dicho sistema tiene dos fases principales. A bajas temperaturas, el sistema se encuentra en su fase ferromagnética u ordenada (todos o la mayoría de los espines se encuentran paralelos); dado que $N_{max} \sim N$, $< m > \sim 1$ y $G_{ij} \sim 1$. A medida que aumenta la temperatura, el sistema sufre una transición abrupta hacia un estado paramagnético o desordenado (espines orientándose al azar por agitación térmica); donde $N_{max} \sim \frac{N}{q}$, $< m > \sim 0$ y G_{ij} decae a $\frac{1}{q}$ (el estado de espín de s_i es independiente del estado de espín s_j).

Si se considera el caso de un sistema completamente inhomogéneo, donde los espines forman zonas de alta densidad o clusters magnéticos, a bajas temperaturas el sistema también se encontrará en estado ferromagnético, pero a medida que la temperatura ascienda podrá apreciarse una fase intermedia denominada superparamagnética. En esta fase los espines pertenecientes a dichos clusters tienden a alinearse entre sí mostrando propiedades ferromagnéticas locales, pero no necesariamente un orden relativo (o correlación) entre clusters distintos. En la temperatura de transición entre la fase ferromagnética y la superparamagnética, se observa un incremento abrupto de χ debido a la fluctuación magnética producida por los clusters, que se comportan como super-espines (todos sus integrantes se alinean simultáneamente, provocando una mayor varianza en la magnetización). A medida que la temperatura incrementa, se llega a la transición de la fase superparamagnética a la fase paramagnética; cada cluster se "desordena" abruptamente y χ disminuye por un factor proporcional al tamaño del cluster más grande. Debido a que cada cluster está compuesto a su vez por clusters más pequeños, es muy común encontrar sub-transiciones en la fase superparamagnética, también marcadas por picos en χ.

Relación con la clasificación de información

El modelo de Potts permite la clasificación de espines magnéticos según su estado y la distancia que los separa. ¿Que sucedería si ,en vez de partículas en un imán, se tratara con un vector de puntos multidimensionales? Si se graficara este vector (suponiendo que su dimensión no es mayor a 3), podrán observarse puntos distribuidos en un espacio cuyas coordenadas serán las características de dicho vector. La distancia relativa entre puntos será proporcional a la diferencia entre sus características, y aportará la información diferencial necesaria para realizar una clasificación. Por lo tanto, si tratamos al sistema como un imán, podremos abordarlo mediante el modelo de Potts y obtener los principales grupos que componen el set de datos; esto es, realizar un clustering superparamagnético[4]. Para poder efectuar esto, utilizamos el algoritmo de Swendsen-Wang[5], que itera sobre la dinámica del modelo de Potts a distintas temperaturas.

Algoritmo de Swendsen-Wang

Parámetros principales

Dado que el número de configuraciones posibles S crece exponencialmente con la cantidad de espines y estados (un sistema con $N = 1000$ y $q = 10$ posee 10^{1000} configuraciones posibles), resulta impráctico realizar el cálculo directo de sumas como (2). Para sobrepasar este obstáculo, es posible crear un subespacio de ξ mediante cadenas de Markov, definidas a través de simulaciones por Monte Carlo[6][7]. De esta manera, la expresión (2) queda reducida al promedio aritmético

$$< \Phi > \sim \frac{1}{M} \sum_{i}^{M} \Phi(S_i) \qquad (7)$$

donde M (la cantidad de configuraciones del subespacio ξ) es mucho menor que q^N. Uno de los algoritmos comunmente utilizados en estos casos es el algoritmo de Swendsen-Wang. La idea principal del algoritmo es mapear cada espín s_i a un punto x_i en un espacio d-dimensional y simular las interacciones entre los distintos puntos, tal que

$$p_{ij} = 1 - e^{-\frac{J_{ij}}{T} \delta_{s_i s_j}} \qquad (8)$$

donde p_{ij} es la probabilidad de interacción entre x_i y x_j. Nótese que $p_{ij} \neq 0 \Leftrightarrow s_i = s_j$. Se permite interactuar sólo a puntos que cumplan el requisito de K-vecindad mutua: se define que x_i es el K-ésimo vecino cercano de x_j si y sólo si x_j es el K-ésimo vecino cercano de x_i. Si se satisface la probabilidad de interacción, se crea un enlace entre x_i y x_j, y ambos puntos son asignados al mismo grupo o cluster de Swendsen-Wang (cluster SW). La fuerza de interacción local entre dos puntos es función de la distancia que los separa, y se define como

$$J_{ij} = \frac{1}{\widehat{K}} e^{-\frac{d_{ij}^2}{2a^2}} \qquad (9)$$

donde $\widehat{K}$ es la cantidad promedio de vecinos cercanos, a es la distancia promedio entre vecinos cercanos y $d_{ij} = \sqrt{|| x_i - x_j ||}$ es la distancia euclídea entre los vecinos cercanos x_i y x_j. Luego, para construir los clusters definitivos, es necesario definir la conectividad entre puntos como

$$C_{ij} = 1 \Leftrightarrow x_i \text{ pertenece al mismo cluster SW que } x_j \qquad (10)$$

Utilizando el valor medio termodinámico de C_{ij} (probabilidad de encontrar a x_i y x_j en el mismo cluster SW) es posible obtener el estimador mejorado de la correlación espín-espín

$$G_{ij} = \frac{(q-1) < C_{ij} > + 1}{q} , \qquad (11)$$

Si $G_{ij} > \Theta$, se asignan x_i y x_j al mismo cluster de información. Dado que en la etapa superparamagnética $G_{SPM} \sim 1 - \frac{2}{q}$ y en la fase paramagnética $G_{PM} \sim \frac{1}{q}$, puede estimarse el umbral de correlación de la siguiente manera:

$$\Theta = \frac{G_{SPM} + G_{PM}}{2} \qquad (12)$$

Etapas del algoritmo

El algoritmo de Swendsen-Wang puede ser resumido en 3 etapas principales:

i. Generación del modelo

1. Asociar a cada espín s_i un punto x_i.
2. Identificar los K vecinos cercanos de cada punto x_i.
3. Calcular la interacción J_{ij} entre puntos vecinos x_i y x_j.

ii. Localización de la fase superparamagnética

1. Seleccionar la cantidad de iteraciones M y la temperatura máxima T_{max}.
2. Generar una configuración S al azar.
3. Asignar un enlace entre puntos cercanos x_i y x_j con probabilidad p_{ij}.
4. Localizar los clusters de Swendsen-Wang y resetear enlaces.
5. Generar una nueva configuración S. Los puntos pertenecientes al mismo cluster SW obtienen el mismo valor de espín con la misma probabilidad.
6. Calcular m y C_{ij}.
7. Ir al paso (3), a menos de haber alcanzado M.
8. Calcular $< m >$, χ y G_{ij}.
9. Ir al paso (2), a menos de haber alcanzado T_{max}.

iii. Obtención de los clusters

1. Ubicar el régimen superparamagnético mediante χ y estimar $T_{clustering}$.
2. Evaluar G_{ij} en $T = T_{clustering}$ y construir los clusters.

Resultados

Se presentan a continuación los resultados de dos simulaciones de dos sets de datos distintos. En ambas simulaciones se utilizó $q = 20$, $M = 5$, $T_{max} = 0.2$, $\Delta T = 0.01$ y $K = 10$. Se presenta en la figura 1 el resultado final del algoritmo de clustering sobre el primer set, compuesto por 2160 puntos bidimensionales de distribución angular uniforme en $[0, 2\pi]$ y distribución normal $N[r, \sigma]$, donde r representa el valor medio del radio y σ el desvío estándar alrededor del valor de r.

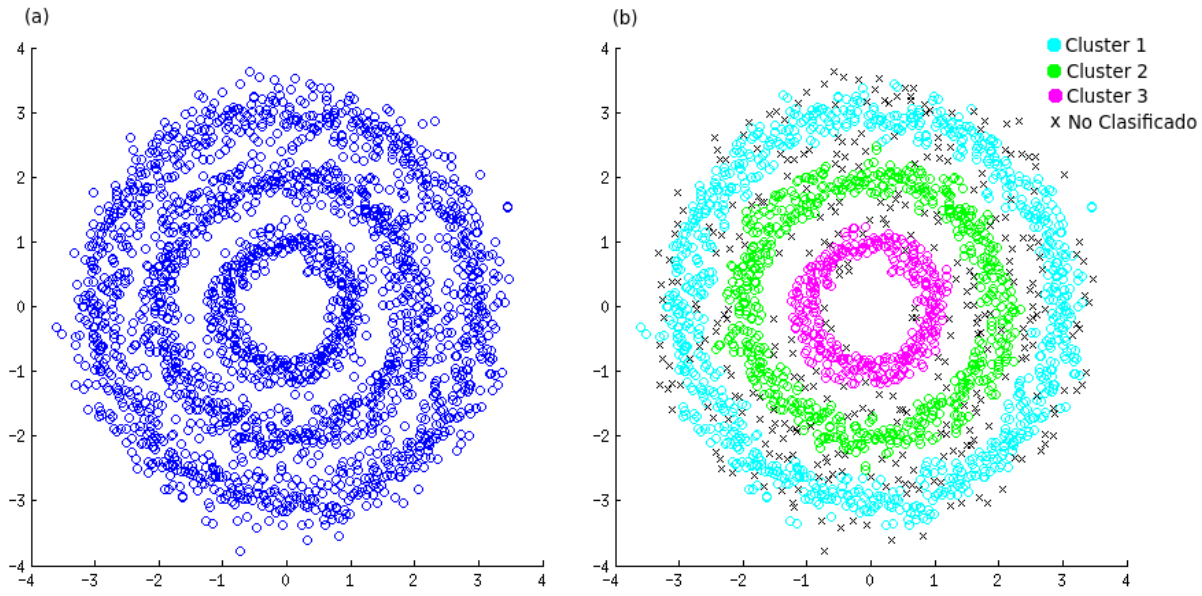

Fig. 1 - Resultado del algoritmo para el primer set. (a) Datos de entrada. (b) Clusters generados.

Los 3 grupos poseen 360, 720 y 1080 puntos cada uno, con $R_1 = N[1, 0.15]$, $R_2 = N[2, 0.2]$ y $R_3 = N[3, 0.25]$. La clasificación resultó de la siguiente manera:

	Puntos Originales	Puntos Clasificados	% Clasificación
Cluster 1	360	329	91,4%
Cluster 2	720	625	86,8%
Cluster 3	1080	874	80,9%
Total	**2160**	**1828**	**84,6%**

Fig. 2 - Clasificación de los puntos para el primer set.

Los 332 puntos restantes fueron asignados a un cluster de puntos sin clasificar, ya que no cumplieron con el criterio de asignación, derivado de las fuerzas de interacción y las correlaciones. En la figura 3 pueden observarse los parámetros del modelo de Potts y la variación de los clusters principales en función de la temperatura:

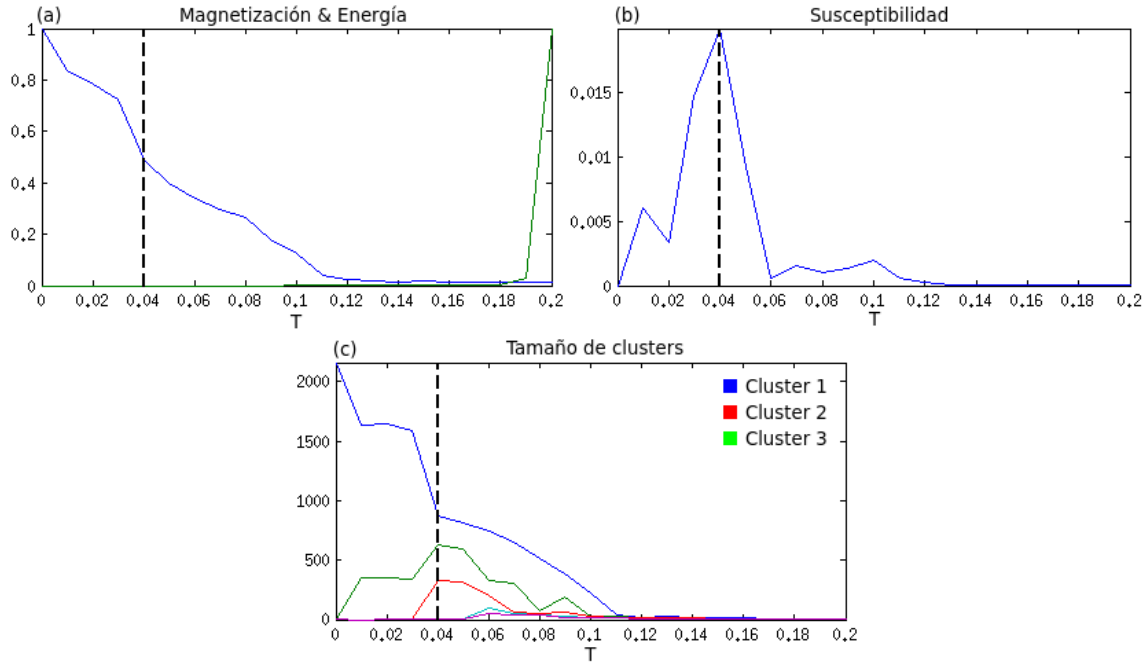

Fig. 3 - Parámetros del modelo para el primer set. Se indica con línea de puntos la temperatura de clustering utilizada. (a) Magnetización (curva azul) y Energía (curva verde) del sistema. (b) Susceptibilidad Magnética del sistema. (c) Tamaño de los 3 clusters principales.

El segundo set de datos está conformado por tres centroides de coordenadas $(x,y,z) = \{(0,0,0),(15,15,15),(-15,15,-15)\}$, de 300 puntos cada uno y distribución no uniforme. En la siguiente figura se presenta el resultado del algoritmo:

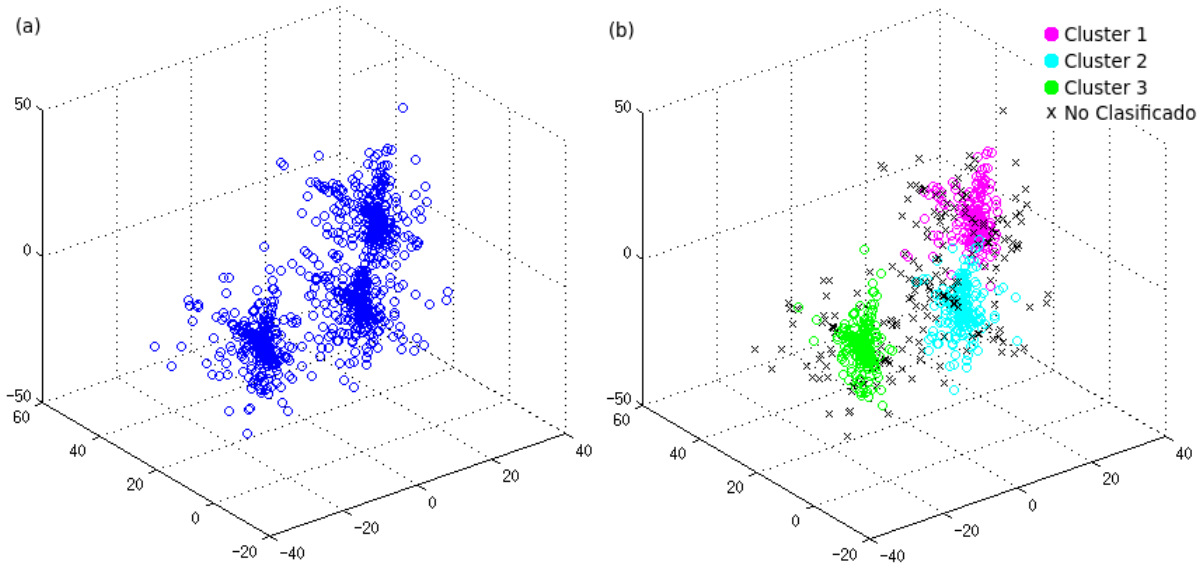

Fig. 4 - Resultado del algoritmo para el segundo set. (a) Datos de entrada. (b) Clusters generados.

Los puntos se clasificaron de la siguiente manera:

	Puntos Originales	Puntos Clasificados	% Clasificación
Cluster 1	300	245	81,6%
Cluster 2	300	250	83,3%
Cluster 3	300	252	84%
Total	**900**	**747**	**83%**

Fig. 5 - Clasificación de los puntos para el segundo set.

Los 153 puntos restantes no fueron clasificados.

Puede apreciarse en la figura 6 los parámetros obtenidos:

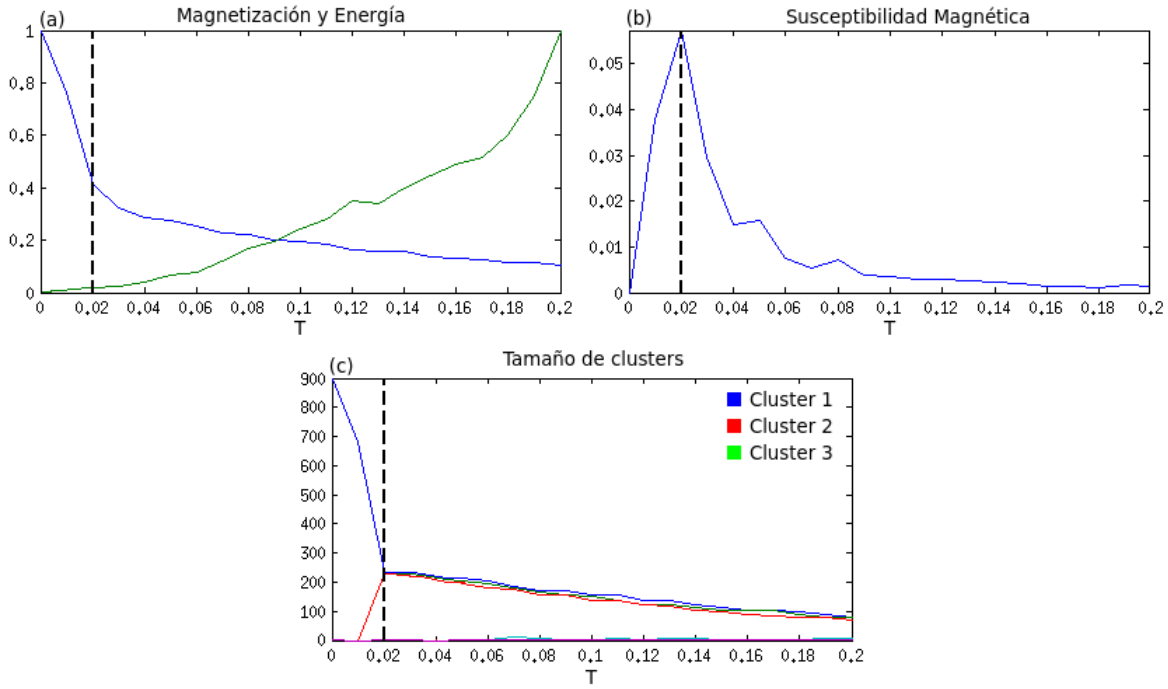

FIg. 6 - Parámetros del modelo para el segundo set. Se indica con línea de puntos la temperatura de clustering utilizada. (a) Magnetización (curva azul) y Energía (curva verde) del sistema. (b) Susceptibilidad Magnética del sistema. (c) Tamaño de los 3 clusters principales.

Conclusiones

Como puede observarse en los resultados, el algoritmo permite clasificar información de manera no supervisada. La efectividad del modelo es independiente de la dimensionalidad y la naturaleza de la información; es más, es posible realizar el clustering sólo conociendo la matriz de distancias d_{ij}, ya que a partir de ella se calcula J_{ij}. Los resultados expuestos en este trabajo corresponden a corridas de Monte Carlo con un número relativamente bajo de iteraciones ($M = 5$) en comparación con la bibliografía, donde suelen utilizar $M > 100$. Gracias a ello, el algoritmo puede correr relativamente rápido (no más de 5 segundos), mientras que en la bibliografía se reportan tiempos de simulación de entre 40 y 60 minutos. Una gran desventaja de esto es, claramente, la baja eficiencia relativa de clasificación (de aproximadamente un 80%) contra el porcentaje de más del 90% observado en la bibliografía. Se aprecia en la figura 2, además, como la clasificación pierde efectividad a medida que aumenta el desvío estándar (y por lo tanto la distancia entre puntos). Para reforzar las interacciones entre puntos relativamente lejanos pero pertenecientes a un mismo cluster de información, se recomienda aumentar la cantidad de iteraciones máximas. Esto permitirá obtener un valor más fiel del umbral de correlación Θ y por lo tanto aumentará la efectividad del algoritmo.

Bibliografía

1. Wu, Fa-Yueh. "The potts model." *Reviews of modern physics* 54.1 (1982): 235.
2. Cai, Wei. "Handout 12. Introduction to Statistical Mechanics." (2011).
3. Cipra, Barry A. "An introduction to the Ising model." *American Mathematical Monthly* 94.10 (1987): 937-959.
4. Blatt, Marcelo, Shai Wiseman, and Eytan Domany. "Data clustering using a model granular magnet." *Neural Computation* 9.8 (1997): 1805-1842.
5. Swendsen, Robert H, and Jian-Sheng Wang. "Nonuniversal critical dynamics in Monte Carlo simulations." *Physical review letters* 58.2 (1987): 86.
6. K. P. N. Murthy. “An introduction to Monte Carlo simulations in statistical physics.” (2003).
7. Helmut G. Katzgraber. “Introduction to Monte Carlo Methods.” (2011).