Notes on Portfolios

JeaSung Park
Soongsil University, Seoul

15 January 2018

Abstract

안녕하십니까? 컴퓨터비전을 활용해 인간에게 도움이 되는 좋은 소프트웨어를 만들고 싶은 박재성입니다. 병렬 처리를 통해 빠른 연산을 해내는 그래픽 처리 장치에 매료되어 이 분야에 첫 발을 내딛게 되었고, 확률 그래프 모델을 활용한 이미지 분석에 관심을 가지고 2 년간 공부하였습니다. 이 부서에 지원하게 된 까닭은 딥러닝 모델 설계의 자동화를 통해 기존 모델 대비 정확도가 향상된 모델을 만들고 싶어서이며, 해당 분야를 연구함에 있어 저의 기계학습 전반에 대한 이해와 그래프 모델을 응용한 접근이 도움이 될 것이라고 생각했기 때문입니다. 기술 본위의 IT 회사인 카카오에서 우수한 연구원분들과 함께 협업하며 좋은 결과를 이끌어내고 싶습니다. 감사합니다.

Chapter 1

Covariance Selection

1.1 Sparse inverse covariance estimation with the graphical lasso

Before explaining some of the important properties in multivariate statistical analaysis, I will introduce some notations for ease of reading.

Definition An *exterior differential form* of degree r in $\mathbb{R}^m$ is an expression of the type

$$\sum_{t_1<t_2<\cdots<t_r} h_{t_1,\cdots,t_r}(x) dx_{t1} \wedge \cdots \wedge dx_{t_r}$$

where the $h_{t_1,\cdots,t_r}$ are analytic functions of $x_1, \cdots, x_m$ □

This definition comes with an important property that is fundamental to changing variables from one to another.

Theorem If dy is an $m \times 1$ vector of differentials and if $dx = Bdy$, where B is an $m \times m$ nonsingular matrix(so that dx is a vector of linear differential forms), then

$$\bigwedge_{i=1}^{m} dx_i = detB \bigwedge_{i=1}^{m} dy_i$$

□

There is another notation I need to introduce. This is important because there is no other ways to express the sampling distribution of some distribution clearer than this.

Definition Let $A = (a_{i,j})$ be a $p \times q$ matrix and $B = (b_{i,j})$ be an $r \times s$ matrix. The Kronecker product of A and B, denoted by $A \otimes B$, is the $pr \times qs$ matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{11}B & a_{12}B & \cdots & a_{1q}B \\ \vdots & & & \\ a_{p1}B & a_{p2}B & ... & a_{pq}B \end{bmatrix}$$

□

We now have notations sufficient for explaining a sampling distribution of mean and covariance. Define the mean and covariance as follows.

$$\begin{gathered} \bar{X} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} X_i \\ S = \frac{1}{n}A \end{gathered} \tag{1.1}$$

where $A = \sum_{i=1}^{N}(X_i - \bar{X})(X_i^T - \bar{X})^T$

Then, we can prove the following result.

Theorem If the $N \times m$ matrix X is $N(1\mu^T, I_N \otimes \Sigma)$ then $\bar{X}$ and A, defined by (1.1), are independently distributed; $\bar{X}$ is $N_m(\mu, \frac{1}{N}\Sigma)$ and A has the same distribution as Z^TZ, where the $n \times m(n = N - 1)$ matrix Z is $N(0, I_n \otimes \Sigma)$(i.e. the n rows of Z are independent $N_m(0^T, \Sigma)$ random vectors). □

Motivated by the above theorem, I now introduce a distribution called *Wishart distribution*

Definition If $A = Z^TZ$ where the $n \times m$ matrix Z is $N(0, I_n \otimes \Sigma)$, then A is said to have the *Wishart distribution* with n degrees of freedom and covariance Σ. We will write that A is $W_m(n, \Sigma)$, the subscript on W denoting the size of the matrix A. □

We now have the definition of Wishart distribution. This comes with a natural question on the probability density function of Wishart distribution. The following theorem gives the density function of A when $n \geq m$.

Theorem If A is $W_m(n, \Sigma)$ with $n \geq m$ then the density function of A is

$$\frac{1}{2^{mn/2}\Gamma_m(\frac{1}{2}n)(det(\Sigma)^{n/2}} exp(tr(-\frac{1}{2}\Sigma^{-1}A))(detA)^{(n-m-1)/2}$$

where $\Gamma_m(\bullet)$ denotes the multivariate gamma function. □

Suppose we want to estimate the covariance matrix given a data normally distributed. Then the sampling distribution will surely follow the Wishart distribution and as such, finding a maximum likelihood solution gives the desired result. So we propose an objective function below and try to find its global maximum(or minimum vice versa).

$$l_1(\Sigma, \phi) = tr(\Sigma^{-1}\phi) - logdet(\Sigma^{-1}\phi) - m \tag{1.2}$$

where $\phi = \hat{\Sigma}$ is a sampled estimate of the true covariance matrix.

If we restrict the solution space of the problem above, we can prove that the best solution is an unbiased estimate of the covariance matrix.

Theorem Using the loss function $l_1(\Sigma, \phi)$, the best(smallest risk) estimate of Σ having the form αA is the unibiased estimate $S = \frac{1}{n}A$. □

The thing is that we can do better than the sample covariance matrix S if we look outside the class of estimates of the form αA. James and Stein(1961) have shown using an *invariance argument* to prove this result. It is reasonable to require that if $\phi(A)$ estimates Σ and L is a nonsingular $m \times m$ matrix, then ϕ should satisfy

$$\phi(L^T AL) = L^T \phi(A) L$$

for $L^T AL$ is $W_m(n, L^T \Sigma L)$, so that $\phi(L^T AL)$ estimates $L^T \Sigma L$. Then

Theorem Using the loss function $l_1(\Sigma, \phi)$ given by (1.2), the best(smallest risk) estimate of Σ in the class of estimates satisfying

$$\phi(L^T AL) = L^T \phi(A) L$$

for all upper-triangular matrices L, is

$$\phi^*(A) = T^T \begin{bmatrix} \frac{1}{n+m-1} & & & 0 \\ & \frac{1}{n+m-3} & & \\ & & \ddots & \\ 0 & & & \frac{1}{n+m-2m+1} \end{bmatrix} T$$

where $A = T^T T$ with T upper-triangular □

This gives us an importance of regularizing the solution space so we add an l^1-penalty to the objective function.

$$l_1(\Sigma, \phi) = tr(\Sigma^{-1}\phi) - logdet(\Sigma^{-1}\phi) + \lambda \left\|\Sigma^{-1}\right\|_1$$

There are many ways to solve the objective. One way to solve this is to use a method called *coordinate descent*. Let W be the estimate of Σ. By partitioning the matrix W and solve each segment, we can achieve the goal. There is theoretical reason for this and I'll explain it later. First, partition W and $S := \phi$.

$$W = \begin{bmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{bmatrix}, S = \begin{bmatrix} S_{11} & s_{12} \\ s_{12}^T & s_22 \end{bmatrix}$$

The solution for w_{12} satisfies

$$w_{12} = argmin_y\{y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \lambda\} \tag{1.3}$$

Using convex duality, Banerjee and others(2007) go on to show that solving (1.3) is equivalent to solving the dual problem

$$min_\beta\{\tfrac{1}{2}\left\|W_{11}^{\frac{1}{2}}\beta - b\right\|^2 + \lambda \|\beta\|_1\} \tag{1.4}$$

where $b = W_{11}^{-\frac{1}{2}} s_{12}$.

The solution of this lasso regression problem is well known.

$$\hat{\beta}_j = \frac{S(u_j - \sum_{k \neq j} V_{kj} \hat{\beta}_k, \lambda)}{V_{jj}}$$

where $V = W_{11}$, $u = s_{12}$ and $S(x,t) = sign(x)(|x| - t)_+$

Algorithm 1 Graphical Lasso Algorithm

```
Start with $W = S + \lambda I$. The diagonal of $W$ remains unchanged in what follows.
while not converged do
    for $j = 1, 2, \cdots, p$ do
        Solve the lasso problem (1.4)
    end for
end while
```

We now show that why the coordinate descent algorithm works theoretically.

Definition The **characteristic function** of X is the function $\phi : \mathbb{R} \to \mathbb{C}$ defined by

$$\phi(t) = E(e^{itX})$$

□

There is two important theorems which can be derived by using results in Fourier analysis.

Theorem If X is continuous with density function f and characteristic function ϕ, then

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt$$

at every pointx at which f is differentiable. □

Theorem Suppose that $F_1, F_2, \cdots$ is a sequence of distribution function with corresponding characteristic functions $\phi_1, \phi_2, \cdots$. If $\phi(t) = lim_{n \to \infty} \phi_n(t)$ exists and is continuous at t = 0, then ϕ is the characteristic function of some distribution function F, and $F_n \to F$. □

In other words, if two distributions have the same characteristic function almost everywhere, the two distributions are essentially the same. So, finding the characteristic function of Wishart distribution will give us a way to find useful properties.

Theorem If A is $W_m(n, \Sigma)$ then the characteristic function of A(that is, the joint characteristic function of the $\frac{1}{2}m(m+1)$ variables a_{ij}, $1 \leq i \leq j \leq m$) is

$$\phi(\Theta) = E(exp(i \sum_{j \leq k}^{m} \theta_{jk} a_{jk})) = det(I_m - i\Gamma\Sigma)^{-n/2}$$

where $\Gamma = (\gamma_{ij})$, $i, j = 1, \cdots, m$ with $\gamma_{ij} = (1 + \delta_{ij})\theta_{ij}$, $\theta_{ij} = \theta_{ji}$ and δ_{ij} is the Kronecker delta. □

By exploiting the above theorem, we get the following consequence.

Theorem If A is $W_m(n, \Sigma)$ and M is $k \times m$ of rank k, then MAM^T is $W_k(n, M\Sigma M^T)$. □

The reason why this proposition holds is because the characteristic function of MAM^T and $W_k(n, M\Sigma M^T)$ are the same. so by the theorem in the previous page, we can get the desired result. Likewise, the following theorem also holds.

Theorem Suppose that A is $W_m(n, \Sigma)$, where A and Σ are partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}$$

and put $A_{11 \cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ and $\Sigma_{11 \cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Then $A_{11 \cdot 2}$ is $W_k(n - m + k, \Sigma_{11 \cdot 2})$ and is independent of A_{12} and A_{22}.

The reason why this theorem is important is because the sub-problem should be invariant under partitioning. Since the invariance of Schur complement holds, we can use the coordinate descent method.

Chapter 2

Energy Minimization

2.1 Energy Minimization in Vision

Graphical models have been used in many areas including bayesian statistics, computer visions, bioinformatics and so on. In this chapter, I will introduce an application in image denoising.

Definition A *graph* is a set of vertices V and edges E that connect various pairs of vertices. A graph can be written $G = \{V, E\}$. Each edge can be represented by a pair of vertices, that is, $E \subset V \times V$. □

A very important probelm that can be solved efficiently seeks to maximize flow in a directed graph.

Problem In a directed graph identify one vertex as a source s and another as a target t. Associate with each directed edge e, a *capacity* $c(e)$, which is a non-negative number. A *flow* is a non-negative value $f(e)$ associated with each edge with the following properties. First $0 \leq f(e) \leq c(e)$. Second, at any vertex $v \in \{V - s - t\}$,

$$\sum_{e;Arriving\ at\ v} f(e) - \sum_{e;Leaving\ from\ v} f(e) = 0$$

The value of a flow is

$$\sum_{e;Arriving\ at\ t} f(e)$$

The objective is to maximize this quantity. □

There are many efficient algorithms to maximize the flow. A dual problem is also interesting. Decompose the vertices into two disjoint sets S and T, such that $s \in S$ and $t \in T$. This represents a *cut*. Then the following theorem holds.

Theorem (Max-flow Min-cut) If f is a flow in a flow network $G = \{V, E\}$ with source s and sink t, then the following conditions are equivalent:

1. f is a maximum flow in G
2. $|f| = c(s, T)$ for some cut (S, T) of G

□

The following implementation computes the maximum flow in a flow network, $G = (V, E)$.

Algorithm 2 Ford-Fulkerson

```
for each edge $(u, v) \in G.E$ do
    $(u, v).f = 0$
end for
while there exists a a path $p$ from $s$ to $t$ in the residual network $G_f$ do
    $c_f(p) = min\{c_f(u, v) : (u, v) \text{ is in } p\}$
    for each edge $(u, v)$ in $p$ do
        if $(u, v) \in G.E$ then
            $(u, v).f = (u, v).f + c_f(p)$
        else
            $(v, u).f = (v.u).f - c_f(p)$
        end if
    end for
end while
```

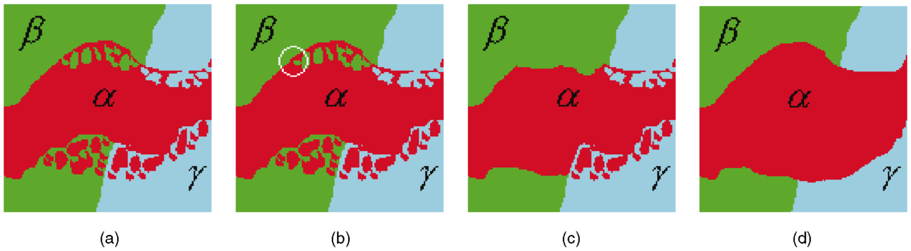
So how do we exploit this nice property in computer vision? Many tasks in computer vision involve assigning a label(such as disparity) to every pixel. A common constraint is that the labels should vary smoothly almost everywhere while preserving sharp discontinuities that may exist, e.g., at object boundaries. These tasks are naturally stated in terms of energy minimization. Again, I will give you another problem formulation.

Problem In image denoising or segmentation, every pixel $p \in P$ should be assigned a label in some finite set L. The goal is to find a labeling f that assigns each pixel $p \in P$ a label $f_p \in L$, where f is both piecewise smooth and consistent with the observed data. This can be achieved by minimizing the energy

$$E(f) = E_{smooth}(f) + E_{data}(f)$$

Here, E_{smooth} measures the extent to which f is not piecewise smooth, while E_{data} measures the disagreement between f and observed data. □

So here is how it goes. Any labeling f can be uniquely represented by a partition of image pixels $P = \{P_l : l \in L\}$, where $P_l = \{p \in P : f_p = l\}$ is a subset of pixels assigned label l. Since there is an obvious one to one correspondence between labelings f and partition P, we can use these notations interchangingly.



(a) (b) (c) (d)

Figure 2.1: Examples of standard and large moves from a given initial labeling (a). The number of labels is $|L| = 3$. A standard moves, (a)→(b), changes the label of a single pixel(in the circled area). Strong moves, $\alpha - \beta - swap$ (a)→(c) and $\alpha - expansion$ (a)→(d), allow large number of pixels to change their labels simultaneously.

Definition Given a pair of labels α, β, a move from a partition P to $P^{'}$ is called an $\alpha - \beta - swap$ if $P_l = P_l^{'}$ for any label $l \neq \alpha, \beta$. □

Definition Given a label α, a move from a partition P to a new partition $P^{'}$ is called an $\alpha - expansion$ if $P_\alpha \subset P_\alpha^{'}$ and $P_l^{'} \subset P_l$ for any label $l \neq \alpha$. □

Given an input labeling f(partition P) and a pair of labels α, β, we wish to find a labeling $\hat{f}$ that minimizes E over all labelings within one $\alpha - \beta$swap of f. The following algorithm shows how to minimize such energy function.

Algorithm 3 Algorithm $\alpha - \beta$ Swap

```
 1: Start with an arbitrary labeling $f$
 2: Set success := 0
 3: **for** each pair of labels $\{\alpha, \beta\} \subset L$ **do**
 4:    Find $\hat{f} = argminE(f^{'})$ among $f^{'}$ within one $\alpha - \beta$ swap of $f$
 5:    **if** $E(\hat{f}) < E(f)$, **then**
 6:       Set $f := \hat{f}$ and success := 1
 7:    **end if**
 8: **end for**
 9: **if** success == 1 **then**
10:    Go to 2:
11: **end if**
```

To let you understand how an image is modeled into graph, I will give you an example of 1-dimensional image.

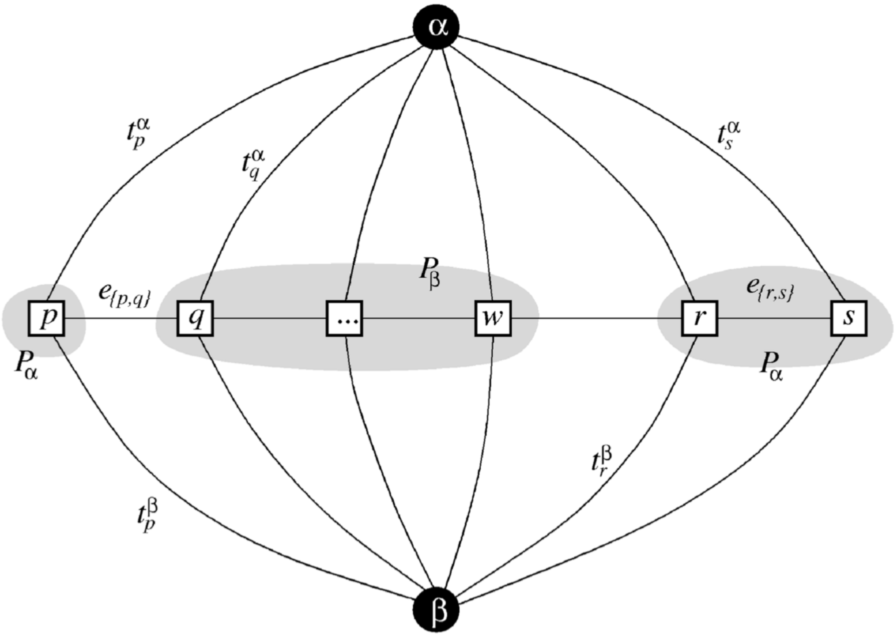


Figure 2.2: An example of the graph $G_{\alpha\beta}$ for a 1D image. The set of pixels in the image is $P_{\alpha\beta} = P_\alpha \cup P_\beta$ where $P_\alpha = \{p, r, s\}$ and $P_\beta = \{q, \cdots, w\}$

Table 2.1: Energy settings for each edge element

Edge	Weight	for
t_p^α	$D_p(\alpha) + \sum_{q \in N_p, q \notin P_{\alpha\beta}} V(\alpha, f_q)$	$p \in P_{\alpha\beta}$
t_p^β	$D_p(\beta) + \sum_{q \in N_p, q \notin P_{\alpha\beta}} V(\beta, f_q)$	$p \in P_{\alpha\beta}$
$e_{p,q}$	$V(\alpha, \beta)$	$\{p, q\} \in N, p, q \in P_{\alpha,\beta}$

By setting energy like the table above and processing Algorithm 3, we can achieve the goal of the problem. This is because the Ford-Fulkerson gives us the minimum cut of a graph while preserving one-to-one correspondence between pixels and labels. This can be justified by the following theorem.

Theorem There is one-to-one correspondence between cuts C on $G_{\alpha\beta}$ and labelings that are one $\alpha - \beta$ swap from f. Moreover, the cost of a cut C on $G_{\alpha\beta}$ is $|C| = E(f^C)$ plus a constant. □

It seems like we can implement the algorithm straightaway. However, we can improve the efficiency of the algorithm 3 by choosing different algorithm in line 4.

Definition A blocking flow of some network is such a flow that every path from s to t contains at least one edge which is saturated by this flow.

Definition A level graph of a network G is a graph built in the following way

1. For each vertex, we calculate $level[v]$, the shortest path(unweighted) from s to this vertex using only edges with positive capacity.
2. We keep only those edges (u, v) for which $level[v] + 1 = level[u]$.

Obviously, this network is acyclic.

These definitions come with an important lemma which is crucial in improving algorithm 2.

lemma The distances from s to each vertex don't decrease after each iteration, i.e. $level_{i+1}[v] \geq level_i[v]$. □

Inspired by the preceding lemma, I now introduce Dinic's Algorithm

Algorithm 4 Dinic's Algorithm

```
for each edge $(u, v) \in G$ do
    $(u, v).f = 0$
end for
Construct the level graph $G_L$ from a graph $G$
if $level[t] == \infty$ then
    Stop and return $f$
end if
while there exists a a path $p$ from $s$ to $t$ in the residual network $G_L$ do
    $c_f(p) = min\{c_f(u, v) : (u, v) \text{ is in } p\}$
    for each edge $(u, v)$ in $p$ do
        if $(u, v) \in G.E$ then
            $(u, v).f = (u, v).f + c_f(p)$
        else
            $(v, u).f = (v.u).f - c_f(p)$
        end if
    end for
end while
```

The algorithm runs in $O(V^2E)$ time and is significantly improved compared to naive Ford-Fulkerson which runs in $O(Ef)$.

Chapter 3

Bayesian Restoration of Images

3.1 The Ising Model

The restoration of degraded images is a branch of digital picture processing, closely related to image segmentation and boundary finding, and extensively studied for its evident practical importance as well as theoretical interest. In this project, I adopted a Bayesian approach and exploited a hierarchical stochastic model for the original image, based on the *Gibbs distribution*, and a new retoration algorithm, based on stochastic relaxation and *annealing*, for computing the maximum a posteriori(MAP) estimate of the original image given the degraded image.

Definition An Ising model is an array of spins (e.g. atoms that can take states ±1) that are magnetically coupled to each other. If one spin is, say, in the +1 state then it is energetically favorable for its immediate neighbours to be in the same state, in the case of *ferromagnetic* model, and in the opposite state, in the case of an *antiferromagnet.*

Let the state x of an Ising model with N spins be a vector in which each component x_n takes values -1 or $+1$. If two spins m and n are neighbours we write $(m, n) \in N$. The coupling between neighbouring spins is J. The energy of a state x is

$$E(x; J, H) = -\left[\frac{1}{2}\sum_{m,n} J_{mn}x_m x_n + \sum_n Hx_n\right]$$

where H is the applied field.

At equilibrium at temparature T, the probability that the state is x is

$$P(x|\beta, J, H) = \frac{1}{Z(\beta, J, H)} exp[-\beta E(x; J, H)]$$

where $\beta = \frac{1}{\kappa_B T}$, κ_B is Boltzmann's constant, and

$$Z(\beta, J, H) = \sum_x exp[-\beta E(x; J, H)]$$

In order to analyze some of the important properties in Ising model, I will introduce a few concepts in Statistical Physics. Consider for example the heat capacity of a system, which is defined to be

$$C = \frac{\partial}{\partial T}\bar{E}$$

where

$$\bar{E} = \frac{1}{Z}\sum_x exp(-\beta E(x))E(x)$$

To work out the heat capacity of a system, we might naively guess that we have to increase the temparature and measure the energy change. Heat capacity, however, is intimately related to energy *fluctuations* at constant temparature. Let's start from the partition function,

$$Z = \sum_x exp(-\beta E(x))$$

The mean energy is obtained by differentiation with respect to β:

$$\frac{\partial ln(Z)}{\partial \beta} = \frac{1}{Z}\sum_x -E(x)exp(-\beta E(x)) = -\bar{E}$$

A further differentiation spits out the variance of the energy:

$$\frac{\partial^2 ln(Z)}{\partial \beta^2} = \frac{1}{Z}\sum_x E(x)^2 exp(-\beta E(x)) - \bar{E}^2 =< E^2 > -\bar{E}^2 = var(E)$$

This result can be interpreted by letting $\lambda = 0$ in a cumulant generating function in a statistical sense.

$$ln\left[E(e^{\lambda E(x)})\right]$$

But the heat capacity is also the derivative of $\bar{E}$ with respect to temparature:

$$\frac{\partial \bar{E}}{\partial T} = -\frac{\partial}{\partial T}\frac{\partial lnZ}{\partial \beta} = -\frac{\partial^2 lnZ}{\partial \beta^2}\frac{\partial \beta}{\partial T} = -var(E)\left(-\frac{1}{\kappa_B T^2}\right)$$

So for any system at temparature T,

$$C = \frac{var(E)}{\kappa_B T^2} = \kappa_B \beta^2 var(E)$$

Thus if we can observe the variance of the energy of a system at equilibrium, we can estimate its heat capacity. Notice that when $var(E)$ is constant, that is, a system is at its equilibrium, the heat capacity of the system is inversely proportional to T^2.

3.2 Monte Carlo simulation

In this section, I will take an example of two-dimensional planar Ising model and compute its equilibrium using a simple Gibbs-sampling method. Starting from some initial state, a spin n is selected at random, and the probability that it should be $+1$ given the state of the other spins and the temparature is computed,

$$P(+1|b_n) = \frac{1}{1 + exp(-2\beta b_n)}$$

where $\beta = \frac{1}{\kappa_B T}$ and b_n is the local field

$$b_n = \sum_{m:(m,n)\in N} Jx_m + H$$

After sufficiently many iterations, this procedure converges to the equilibrium distribution. An alternative and computationally friendly form of Gibbs-sampling formula is as follows.

$$P(accept; \Delta E, \beta) = \begin{cases} 1 & \Delta E \leq 0 \\ exp(-\beta \Delta E) & \Delta E > 0 \end{cases}$$

where $\Delta E = 2x_n b_n$.

3.3 Contrast with Schottky anomaly

So far, we have dealt with some theories that explains several properties of the Ising model. From now on, I will present an empirical result observed after simulating it in two different settings and show what's the difference between reality and the simulated environment.

A peak in the heat capacity, as a function of temparature, occurs in any system that has a finite number of energy levels; a peak is not in itself evidence of a phase transition. Such peaks were viewed as anomalies in classical thermodynamics, since 'normal' systems with infinite numbers of energy levels have heat capacities that are either constant of increasing function of temparature. In contrast, systems with a finite number of levels produced small blips in the heat capacity graph(Figure 3.1).

Let us refresh our memory of the simplest such system, a two-level system with states $x = 0$(energy 0) and $x = 1$(energy ϵ). The mean energy is

$$E(\beta) = \epsilon \frac{exp(-\beta\epsilon)}{1 + exp(-\beta\epsilon)} = \epsilon \frac{1}{1 + exp(-\beta\epsilon)}$$

and the derivatives with respect to β is

$$\frac{dE}{d\beta} = -\epsilon^2 \frac{exp(\beta\epsilon}{[1 + exp(\beta\epsilon)]^2}$$

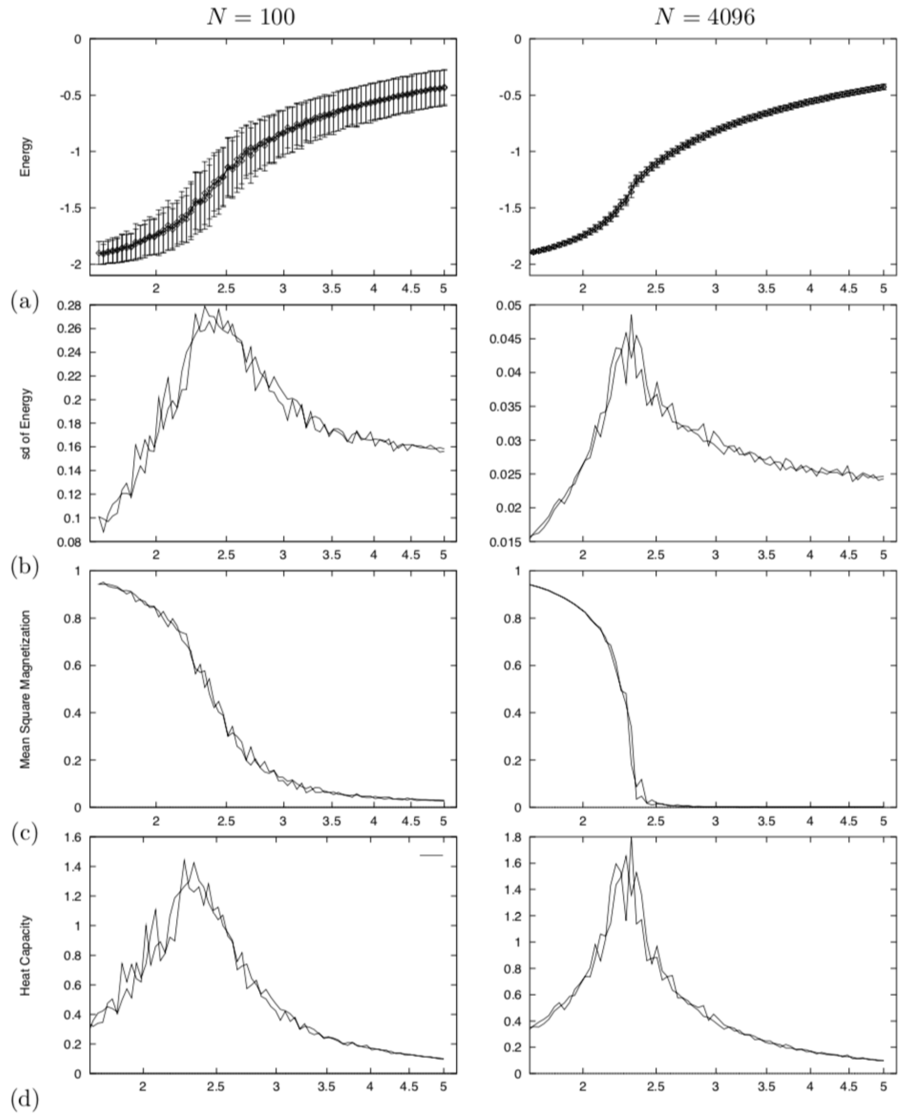


Figure 3.1: Detail of Monte Carlo simulations of rectangular Ising Models with $J = 1$. (a) Mean energy and fluctuations in energy as a function of temparature. (b) Fluctuations in energy (standard deviation). (c) Mean square magnetization. (d) Heat capacity.

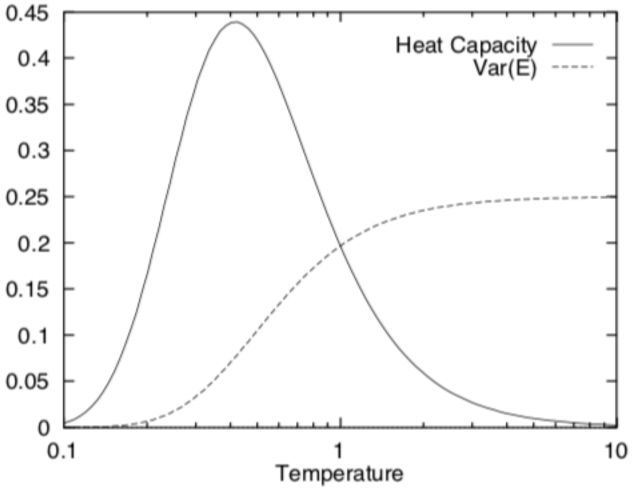


Figure 3.2: Schottky anomaly. Heat capacity and fluctuations in energy as a function of temparature for a two-level system with separation $\epsilon = 1$ and $\kappa_B = 1$

So the heat capacity is

$$C = \frac{dE}{dT} = -\frac{dE}{d\beta}\frac{1}{\kappa_B T^2} = \frac{\epsilon^2}{\kappa_B T^2}\frac{exp(\beta\epsilon)}{[1+exp(\beta\epsilon)]^2}$$

and the fluctuations in energy are given by $var(E) = C\kappa_B T^2 = -\frac{dE}{d\beta}$.

The heat capacity and fluctuations are plotted in figure 3.2. Notice that there is no peak in their fluctuations; the variance of the energy simply increases monotonically with temparature to a value proportional to the number of independent spins. As a consequence, the heat capacity has a peak around 0.7 and converges to 0 as temparature goes to infinity.