2021 級 信工計算機組 學生干皓丞 2101212850 人工智慧作業報告

Email : zxdfgcv@gmail.com
About me : https://kancheng.github.io/

0. 作業說明

此報告為人工智慧課程五篇閱讀報告中的第四篇 Bayesian Network，全報告包含 agent、search、Markov decision process、Bayesian Network、Reinforcement Learning。因為考量自身到對該領域知識的掌握程度不足，全報告採心得與翻譯。

GitHub Project : https://github.com/kancheng/kan-readpaper-cv-and-ai-in-2021

1. 原文獻資訊與作者
2. 報告內容心得與講述
3. 原研究文獻

1. 原文獻資訊與作者

DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks
無所畏懼的 DAG：深入了解學習貝葉斯網絡的持續優化

DennisWei; IBM Research; dwei@us.ibm.com
Tian Gao; IBM Research; tgao@us.ibm.com
Yue Yu; Lehigh University; yuy214@lehigh.edu

https://arxiv.org/abs/2010.09133
Comments: 40 pages, 8 figures, to appear at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)
Subjects: Machine Learning (cs.LG); Machine Learning (stat.ML)

2. 報告內容心得與講述

(1) Motivation 动机

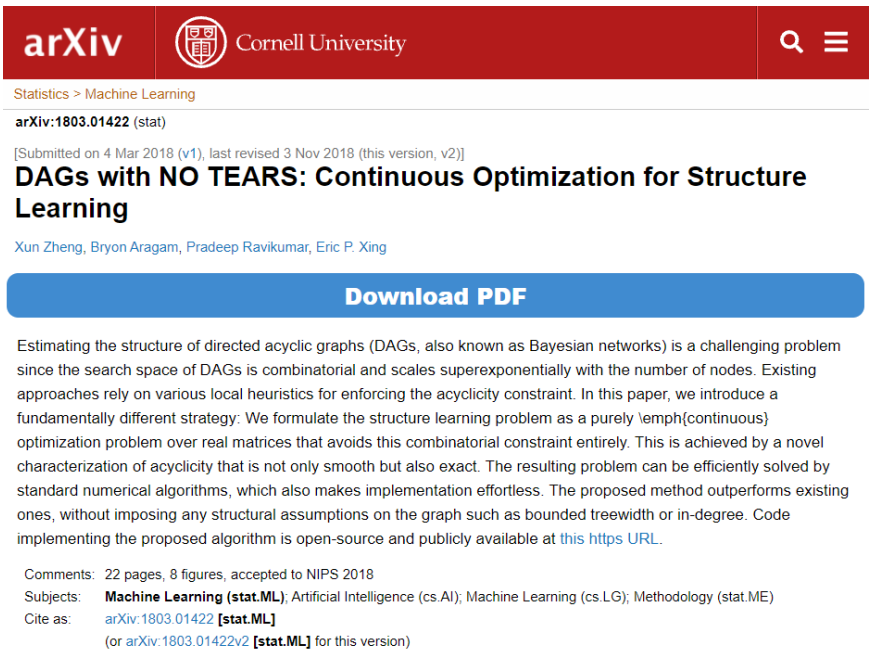
該研究重新審視了一個名為 NOTEARS 的連續優化框架，想將其使用在貝葉斯網絡學習上。研究者首先將非循環性的現有代數特徵(existing algebraic characterizations of acyclicity)推廣到一類矩陣多項式(a class of matrix polynomials)上面，接著重點關注每邊一個參數的設置，結果表明，排除在不影響整體地的細節情況下，不能滿足 NOTEARS 公式的 Karush-Kuhn-Tucker (KKT) 最優性條件，並解釋該演算法的行為。

並且研究者用數學去推導出等效重構 KKT 條件，來證明該條件在這之中是必要的，並將它們與圖中某些邊不存在的顯式約束相關聯。根據 KKT 條件，提出了一種局部搜索後處理算法(a local search post-processing algorithm is proposed)，並證明該演算法可以顯著和普遍地改善所有測試演算法的結構漢明距離(the structural Hamming distance)，通常提高 2 倍或更多。與本地搜索的某些組合比原始 NOTEARS 更準確、更有效。

在此有三個專有名詞需要注意 !!!

* Structural Hamming distances (SHD)
* Learning Bayesian Networks
* NOTEARS 公式的 Karush-Kuhn-Tucker (KKT)
* DAGs - directed acyclic graphical model

[33] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, December 2018.



https://arxiv.org/abs/1803.01422

[33] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In Advances in Neural Information Processing Systems, pages 9472–9483, December 2018.

1. 貝葉斯定理

條件機率（又稱後驗機率）就是事件 A 在另外一個事件 B 已經發生條件下的發生機率。其條件機率表示為 P(A|B)，讀作“在 B 條件下所發生 A 的機率”。

若在同一個樣本空間Ω中的事件或者子集 A 與 B，如果隨機從Ω當中選出的一個元素屬於 B，那麼這個隨機選擇的元素還屬於 A 的機率就定義為在 B 的前提下 A 的條件機率：

$P(A|B)=P(A\cap B)/P(B)$

聯合機率：$P(A\cap B)$ or P(A,B)
先驗機率：P(A) or P(B)

2. 貝葉斯網絡(Bayesian network)，又稱信念網絡(Belief Network)，或有向無環圖模型(directed acyclic graphical model)，是一種機率圖模型。1985 年由 Judea Pearl 首先提出。它是一種模擬人類推理過程中因果關係的不確定性處理模型，其網絡拓樸結構是一個有向無環圖(DAG)。

貝葉斯網絡的有向無環圖中的節點表示隨機變量{X1,X2,...,Xn}

它們可以是可觀察到的變量，或隱變量、未知參數等。認為有因果關係（或非條件獨立）的變量或命題則用箭頭來連接。若兩個節點間以一個單箭頭連接在一起，表示其中一個節點是“因(parents)”，另一個是“果(children)”，兩節點就會產生一個條件概率值。
舉例說明
假設有兩個伺服器 (S1,S2)，會傳送封包到使用者端（以 U 表示之），但是第二個伺服器的封包傳送成功率會與第一個伺服器傳送成功與否有關，因此此貝氏網路的結構圖可以表示成如圖二的型式。就每個封包傳送而言，只有兩種可能值：T（成功）或 F（失敗）。則此貝氏網路之聯合機率分配可以表示成：
$P(U, S1, S2) = P(U|S1, S2) * P(S2|S1)*P(S1)$

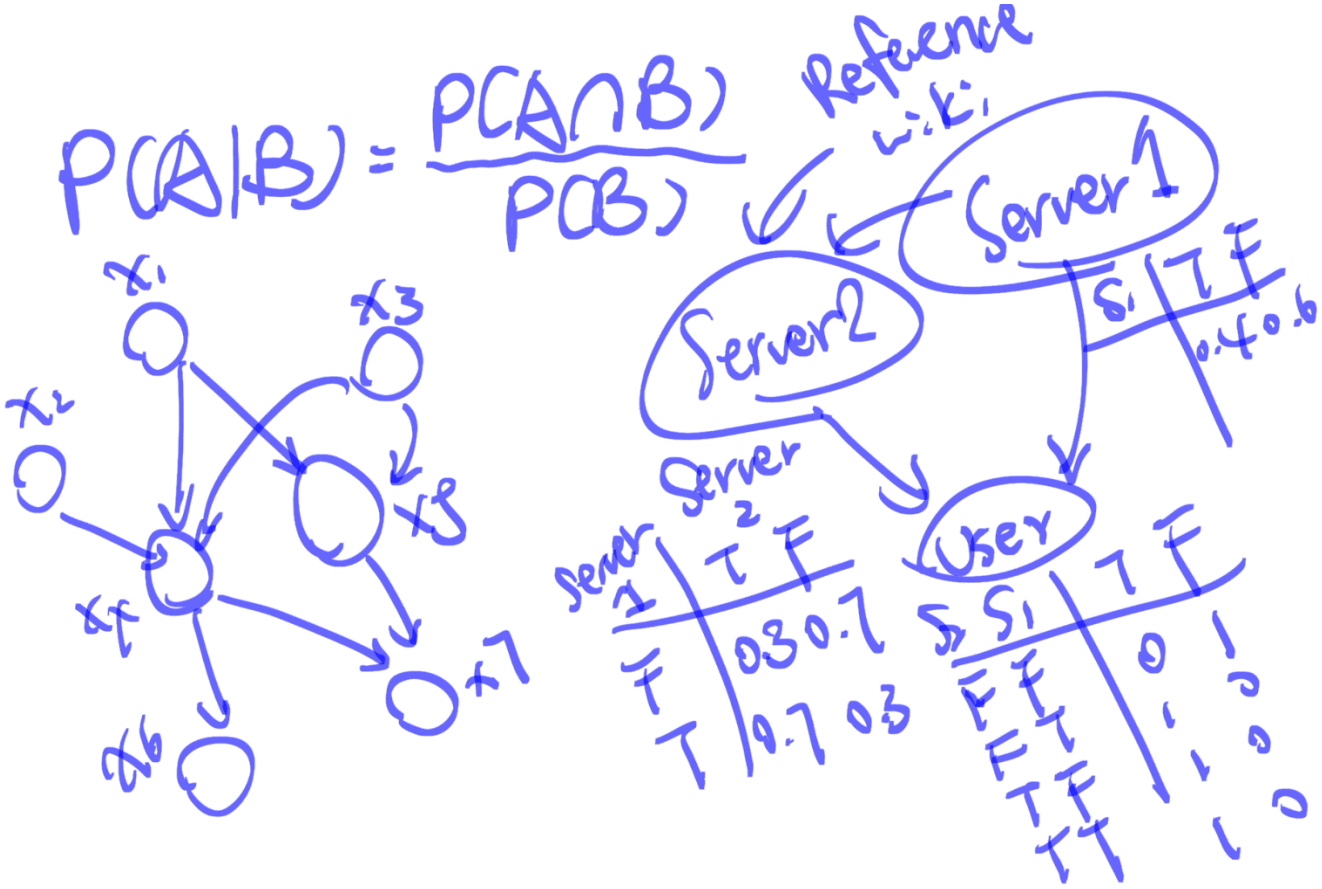
Reference & Note by Kan Horst – 為了理解該文件而額外查詢的知識

https://www.cnblogs.com/mantch/p/11179933.html
https://www.youtube.com/watch?v=bFZ-0FH5hfs
https://www.youtube.com/watch?v=S18XZQFPPvo
https://www.youtube.com/watch?v=artFW5By-Xw
https://www.youtube.com/watch?v=K_qNcLY3XUI

https://www.youtube.com/watch?v=EvXKAP5F_-4

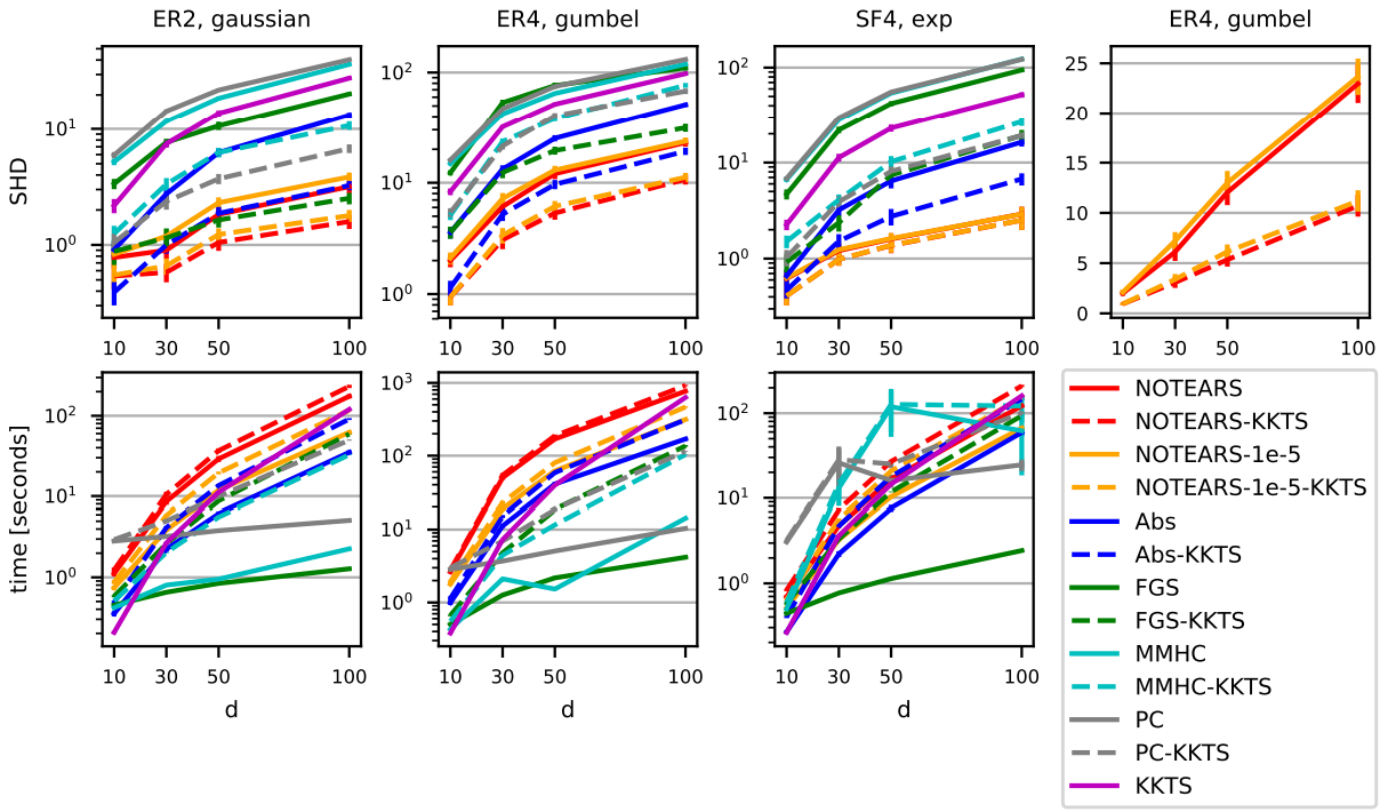
https://zh.wikipedia.org/wiki/%E8%B2%9D%E6%B0%8F%E7%B6%B2%E8%B7%AF

(2) Intuition 直觉



(3) Justification 理由

圖 1：結構漢明距離 (SHD; Structural Hamming distances) 相對於真實圖和 n = 1000 的求解時間。誤差條表示 100 次試驗的標準誤差。在 SF4 SHD 圖中，紅線與橙色重疊。右上角的面板側重於使用線性垂直刻度與 NOTEARS 的組合。可以看到該研究首先關注基本演算法（實線），其中 NOTEARS 在 SHD 方面顯然是最好的。



(4) Framework 框架

元算法(the meta-algorithm in Algorithm 1)，研究者將其稱為 KKT 知情本地搜索(KKT-informed local search)。

可以從該研究第 4.2 節所描述的實例看到名為 KKT-informed local search (KKTS)演算法

定理 9。如果 F(W) 是可分離的，則 KKT 通知的局部搜索會產生滿足 KKT 條件 (9) 的解，而研究的定理 7 和凸 F(W) 指出，定理 9 保證 KKT 通知的局部搜索將導致局部最小值。

The development in this subsection suggests the meta-algorithm in Algorithm 1, which we refer to as KKT-informed local search. An instantiation is described in Section 4.2.

Algorithm 1 KKT-informed local search (KKTS)

Require: Initial set $\mathcal{Z}$ of edge absence constraints. Solve (10).

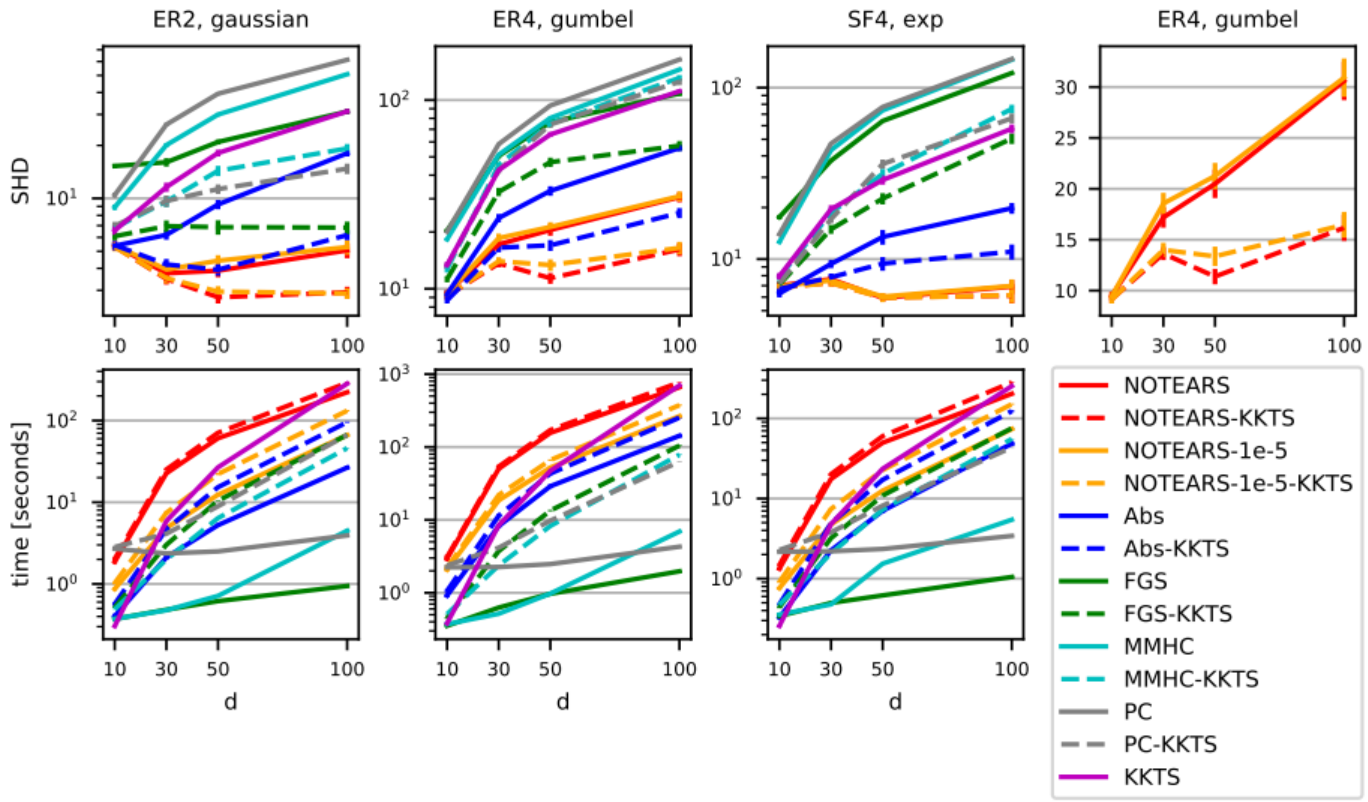
1: **while** $W^*(\mathcal{Z})$ infeasible **do**
2: Select edge(s) in cycle $((W^*(\mathcal{Z}))_{ij} \neq 0, \left(\nabla h(A^*(\mathcal{Z}))\right)_{ij} > 0)$. Add to $\mathcal{Z}$. Re-solve (10).
3: **end while**
4: **while** $\mathcal{Z}$ reducible **do**
5: Remove one or more unnecessary constraints $(i, j) \in \mathcal{Z}$ for which $\left(\nabla h(A^*(\mathcal{Z}))\right)_{ij} = 0$ (see Lemma 8). Re-solve (10).
6: **end while**

Theorem 9. *If $F(W)$ is separable, KKT-informed local search yields a solution satisfying the KKT conditions* (9).

When combined with Theorem 7 and a convex F(W), Theorem 9 guarantees that KKT-informed local search will result in local minima.

(5) Result 结果

Figure 2: (SHD) with respect to true graph and solution times for n = 2d.
從研究結果結論的圖 2 可以看到結構漢明距離 (SHD;Structural Hamming distances) 相對於真實圖和 n = 2d 的求解時間，在 SF4 SHD 圖中，紅線與橙色重疊，且可以看到在圖二右上角的面板側重於使用線性垂直刻度與 NOTEARS 的組合。



3. 原研究文獻

Abstract 摘要

This paper re-examines a continuous optimization framework dubbed NOTEARS for learning Bayesian networks.
本文重新審視了一個名為 NOTEARS 的連續優化框架，用於學習貝葉斯網絡。

We first generalize existing algebraic characterizations of acyclicity to a class of matrix polynomials.
我們首先將非循環性的現有代數特徵推廣到一類矩陣多項式。

Next, focusing on a one-parameter-per-edge setting, it is shown that the Karush-Kuhn-Tucker (KKT)

optimality conditions for the NOTEARS formulation cannot be satisfied except in a trivial case, which explains a behavior of the associated algorithm.
接下來，重點關注每邊一個參數的設置，結果表明，除了在微不足道的情況下，不能滿足 NOTEARS 公式的 Karush-Kuhn-Tucker (KKT) 最優性條件，並解釋該演算法的行為。

We then derive the KKT conditions for an equivalent reformulation, show that they are indeed necessary, and relate them to explicit constraints that certain edges be absent from the graph.
然後我們推導出等效重構的 KKT 條件，表明它們確實是必要的，並將它們與圖中某些邊不存在的顯式約束相關聯。

If the score function is convex, these KKT conditions are also sufficient for local minimality despite the non-convexity of the constraint.
如果得分函數是凸的，儘管約束不凸，這些 KKT 條件也足以滿足局部極小性。

Informed by the KKT conditions, a local search post-processing algorithm is proposed and shown to substantially and universally improve the structural Hamming distance of all tested algorithms, typically by a factor of 2 or more.
根據 KKT 條件，提出了一種局部搜索後處理算法，並證明它可以顯著和普遍地改善所有測試算法的結構漢明距離，通常提高 2 倍或更多。

Some combinations with local search are both more accurate and more efficient than the original NOTEARS.
與本地搜索的某些組合比原始 NOTEARS 更準確、更有效。

1 Introduction 前言

Bayesian networks are directed probabilistic graphical models used to model joint probability distributions of data in many applications [21, 27].
貝葉斯網絡是有向概率圖形模型，用於在許多應用中對數據的聯合概率分佈進行建模 [21, 27]。

Automatic discovery of their directed acyclic graph (DAG) structure is important to research areas from causal inference to biology.
自動發現它們的有向無環圖 (DAG) 結構對於從因果推斷到生物學的研究領域都很重要。

However, DAG structure learning is in general an NP-hard problem [8].
然而，DAG 結構學習通常是一個 NP-hard 問題 [8]。

Many learning algorithms have been proposed to circumvent exhaustive search in the discrete space of DAGs, including those for discrete variables [7, 1, 26, 16, 9, 32, 12] and continuous variables [6, 29].
已經提出了許多學習算法來規避 DAG 離散空間中的窮舉搜索，包括離散變量 [7, 1, 26, 16, 9, 32, 12] 和連續變量 [6, 29] 的算法。

Recently, Zheng et al. [33] proposed a continuous optimization formulation, referred to as NOTEARS, in which acyclicity of the graph is enforced by a trace of matrix exponential constraint on a weighted adjacency matrix.
最近，Zheng et al. [33] 提出了一種連續優化公式，稱為 NOTEARS，其中圖的非循環性由加權鄰接矩陣上的矩陣指數約束的跡線強制執行。

Several works have since successfully extended the formulation to nonlinear and nonparametric models [31, 20, 18, 34].
此後，有幾項工作成功地將公式擴展到非線性和非參數模型 [31, 20, 18, 34]。

This paper takes further steps toward fulfilling the promise of [33] in opening the door to continuous optimization techniques for score-based structure learning.
本研究採取進一步措施來實現 [33] 的承諾，為基於分數的結構學習打開了持續優化技術的大門。

We contribute in particular to theoretical understanding of this framework, leading to significant algorithmic improvements.
我們特別有助於對該框架的理論理解，使之有導致顯著的演算法改進。

First, in Section 2, the matrix exponential constraint of [33] and the matrix polynomial constraint of [31] are generalized to a class of matrix polynomials with positive coefficients whose traces characterize acyclicity.
首先，在第 2 節中，[33] 的矩陣指數約束和 [31] 的矩陣多項式約束被推廣到一類具有正係數的矩陣多項式，其跡線表徵非循環性(traces characterize acyclicity)。

We also provide a characterization involving the gradient of functions in this class, which is not only essential to proving later results but also has an intuitive graphical interpretation.
我們還提供了涉及此類中函數梯度的表徵，這不僅對證明以後的結果至關重要，而且還具有直觀的圖形解釋。

In Section 3.1, we revisit the NOTEARS formulation of [33] in which a weighted adjacency matrix is obtained by element-wise squaring of the parameter matrix.
在第 3.1 節中，我們重新審視了 [33] 的 NOTEARS 公式，其中通過參數矩陣的元素平方獲得加權鄰接矩陣。

It is shown that the Karush-Kuhn-Tucker (KKT) optimality conditions for this constrained optimization cannot be satisfied except in a trivial case.
結果表明，除非在微不足道的情況下，否則無法滿足此約束優化的 Karush-Kuhn-Tucker (KKT; Karush-Kuhn-Tucker) 最優性條件。

This negative result is somewhat surprising given the empirical success of the augmented Lagrangian algorithm of [33], and we use the result to explain why the algorithm does not converge to an exactly acyclic solution even when the penalty parameters are very high.
考慮到 [33] 的增廣拉格朗日演算法(Lagrangian algorithm)的經驗成功，這個負面結果有些令人驚

訝，我們使用結果來解釋為什麼即使懲罰參數非常高，算法也不會收斂到完全非循環的解決方案。

In Section 3.2, we consider an equivalent reformulation in which the adjacency matrix is given by the absolute value of the parameter matrix, motivated in part by the failure to satisfy KKT conditions in Section 3.1, and in part by the connection between the `1 norm and sparsity.
在 3.2 節中，我們考慮了一個等效的重構，其中鄰接矩陣由參數矩陣的絕對值給出，部分原因是由於無法滿足 3.1 節中的 KKT 條件，部分原因是`1 範數之間的聯繫 和稀疏性(the `1 norm and sparsity)。

We show that the KKT conditions for this reformulation are indeed necessary conditions of optimality, i.e. they are satisfied by all local minima, although even here common constraint qualification methods turn out to fail.
我們表明，這種重新制定的 KKT 條件確實是最優性的必要條件，即所有局部最小值都滿足它們，儘管即使在這裡常見的約束限定方法也失敗了。

If the score function is convex, then the KKT conditions are also sufficient for local minimality, despite the non-convexity of the constraint.
如果得分函數是凸的，那麼 KKT 條件對於局部極小也是足夠的，儘管約束是非凸的。

We then relate the KKT conditions to the optimality conditions for score optimization subject to explicit edge absence constraints.
然後，我們將 KKT 條件與受顯式邊緣缺失約束的分數優化的最優性條件相關聯。

The KKT conditions can thus be understood through edge absences: together these must be sufficient to ensure acyclicity, but each absence must also be necessary in preventing the completion of a cycle.
因此，可以通過邊緣缺失來理解 KKT 條件：這些條件一起必須足以確保非循環性，但每次缺失也必須是防止循環完成的必要條件。

The theoretical development of Section 3.2 naturally suggests two algorithms: an augmented Lagrangian algorithm as in [33] with an absolute value adjacency matrix instead of quadratic, and a local search algorithm, KKTS, informed by the KKT conditions and proven to satisfy them.
3.2 節的理論發展自然地提出了兩種算法：[33] 中的增廣拉格朗日算法，使用絕對值鄰接矩陣而不是二次矩陣，以及局部搜索算法 KKTS，根據 KKT 條件提供信息並證明滿足它們。

KKTS (a) adds edge absence constraints to break cycles, (b) removes constraints that are unnecessary, and (c) swaps constraints (reverses edges) to combat non-convexity. We find in Section 5 that neither of these two algorithms yields state-of-the-art accuracy by itself.
KKTS (a) 添加邊缺失約束以打破循環，(b) 刪除不必要的約束，以及 (c) 交換約束（反轉邊）以對抗非凸性。 我們在第 5 節中發現，這兩種算法本身都不能產生最先進的準確性。

However, when combined with other algorithms, KKTS substantially reduces structural Hamming distance (SHD) with respect to the true graph, typically by a factor of at least 2.

然而，當與其他算法結合時，KKTS 相對於真實圖顯著減少了結構漢明距離 (SHD)，通常至少減少 2 倍。

Moreover, this improvement is consistent across dimensions and base algorithms.
此外，這種改進在維度和基礎算法上是一致的。

In the case of NOTEARS, new state-of-the-art accuracy is obtained, while other combinations can outperform NOTEARS and take less time.
在 NOTEARS 的情況下，獲得了新的最先進的準確性，而其他組合可以勝過 NOTEARS 並且花費的時間更少。

More on related work Bayesian network structure learning has long been an active research area.
更多相關工作 貝葉斯網絡結構學習長期以來一直是一個活躍的研究領域。

Constraint- and score-based methods utilize independence tests and graph scores respectively to learn the DAG structure.
基於約束和基於分數的方法分別利用獨立性測試和圖分數來學習 DAG 結構。

Optimization methods such as greedy search [7], dynamic programming [19], branch and bound [10], A* search [32, 30], local-to-global search [13] as well as approximation methods [23] have all been proposed.
優化方法，如貪心搜索 [7]、動態規劃 [19]、分支定界 [10]、A* 搜索 [32, 30]、局部到全局搜索 [13] 以及近似方法 [23] 都被提出了。

As mentioned, this paper is most closely related to the continuous framework of [33] and subsequent works [31, 34].
如前所述，本文與 [33] 的連續框架和後續工作 [31, 34] 的關係最為密切。

Regression-based methods for DAG learning, without the matrix exponential constraint, have also been carefully studied [24, 6, 2, 14].
沒有矩陣指數約束的基於回歸的 DAG 學習方法也得到了仔細研究 [24, 6, 2, 14]。

2 Characterizations of acyclicity - 非週期性的特徵

In this first section, we provide algebraic characterizations of acyclicity for a directed graph in terms of its adjacency matrix.
在第一部分中，我們根據鄰接矩陣為有向圖提供了非循環性的代數特徵。

For a directed graph G = (V, E) with vertices V = {1, …, d} and directed edges (i, j) ∈ E, a non-negative matrix A is a (weighted) adjacency matrix for G if Aij > 0 for (i, j) ∈ E and Aij = 0 otherwise.
對於頂點 V = {1, …, d} 和有向邊 (i, j) ∈ E 的有向圖 G = (V, E)，

如果 Aij > 0 for (i, j) ∈ E 並且 Aij = 0 ，則非負矩陣 A 是 G 的（加權）鄰接矩陣。

We consider a class of functions h(A) corresponding to matrix polynomials of degree d with positive coefficients,
我們考慮一類函數 h(A) 對應於具有正係數的 d 次矩陣多項式，

$$P(A) = c_0 I + c_1 A + \cdots + c_d A^d = \sum_{p=0}^{d} c_p A^p, \qquad c_p > 0, \quad p = 1, \ldots, d,$$

from which we define
我們從中定義

$$h(A) = \operatorname{tr}(P(A)) - c_0 d = \sum_{p=1}^{d} c_p \operatorname{tr}(A^p). \qquad (1)$$

This class includes the function h(A) = tr ((I+A/d)^d)-d from [31], which corresponds to cp =(dp)/dp, and the trace of matrix exponential from [33], h(A)= … = … .
此類包括函數 h(A) = tr ((I+A/d)^d)-d 來自 [31]，對應於 cp =(dp)/dp，以及來自 [33] 的矩陣指數跡 , h(A)= ⋯ = ⋯ 。

$$h(A) = \operatorname{tr}\big((I + A/d)^d\big) - d$$

$$c_p = \binom{d}{p} / d^p$$

$$h(A) = \operatorname{tr}(e^A) - d = \sum_{p=1}^{\infty} \frac{\operatorname{tr}(A^p)}{p!}.$$

Although (2) appears to be an infinite power series, it can be rewritten as a finite series with no powers higher than d using the Cayley-Hamilton theorem [15], which equates A^d to a linear combination of I, A , … , A^(d-1), and similarly for all higher powers of A.
儘管 (2) 看起來是一個無限冪級數，但它可以使用凱萊-漢密爾頓定理(Cayley-Hamilton theorem) [15] 重寫為一個沒有高於 d 的冪的有限級數，該定理將 A^d 等同於 I、A 的線性組合， … , A^(d-1) ，對於 A 的所有更高次冪也類似。

Any function h(A) of the form in (1) can characterize acyclicity, as stated below.
(1) 中形式的任何函數 h(A) 都可以表徵非循環性，如下所述。

We defer all proofs to Appendix A.

我們將所有證明推遲到附錄 A。

Theorem 1. A directed graph G is acyclic if and only if its (weighted) adjacency matrix satisfies h(A) = 0 for any h defined by (1).
定理 1. 有向圖 G 是無環的，當且僅當其（加權）鄰接矩陣對於由 (1) 定義的任何 h 滿足 h(A) = 0。

The proof of Theorem 1 is facilitated by Lemma 1 below.
下面的引理 1 促進了定理 1 的證明。

We recall that a matrix B is said to be nilpotent if Bp = 0 for some power p ∈ N (and consequently all higher powers).
我們回想一下，如果對於某個冪 p ∈ N（以及所有更高的冪），如果 Bp = 0，則矩陣 B 被稱為冪零。

Equivalent characterizations are that all eigenvalues of B are zero, and most usefully here, that tr(Bp) = 0 for all p ∈ N.
等效的特徵是 B 的所有特徵值都為零，並且在這裡最有用的是，對於所有 p ∈ N，tr(Bp) = 0。

We call attention to the lemma as there may be independent interest in alternative ways of enforcing nilpotency.
我們提請注意引理，因為可能對強制執行冪零的替代方法有獨立的興趣。

Lemma 1. A directed graph G is acyclic if and only if its (weighted) adjacency matrix A is nilpotent.
引理 1. 有向圖 G 是無環的當且僅當其（加權）鄰接矩陣 A 是冪零的。

*1 ∇ Del 算子或稱 Nabla 算子，在中文中也叫向量微分算子、劈形算子、倒三角算子。
https://en.wikipedia.org/wiki/Del

*2 Δ Delta

The gradient of h(A) in (1) is a matrix-valued function given by ∇h(A) … .
(1) 中 h(A) 的梯度是由 ∇h(A) … 給出的矩陣值函數。

$$\nabla h(A) = \sum_{p=1}^{d} p c_p \left(A^{p-1}\right)^T . \qquad (3)$$

We make the following elementary observation for later reference.
我們作如下初步觀察，供以後參考。

Lemma 2. For non-negative matrices A, ∇h(A) is non-negative and h(A) is therefore a nondecreasing function in the sense that h(A) ≥ h(B) if A – B ≥ 0.

引理 2. 對於非負矩陣 A，∇h(A) 是非負的，因此 h(A) 是一個非遞減函數，如果 A – B ≥ 0，則 h(A) ≥ h(B)。

Off-diagonal elements ∇h(A))ij have an intuitive interpretation in terms of directed walks from j to i, i.e. a sequence of edges (j, i1), (i1, i2); … ; (il-1, i) ∈ E.
非對角線元素 ∇h(A))ij 在從 j 到 i 的有向行走方面具有直觀的解釋，即邊序列 (j, i1), (i1, i2)； ……; (il-1, i) ∈ E。

Off-diagonal elements $(\nabla h(A))_{ij}$ have an intuitive interpretation in terms of *directed walks* from j to i, i.e. a sequence of edges $(j, i_1), (i_1, i_2), \ldots, (i_{l-1}, i) \in \mathcal{E}$. If there is a directed walk from j to i, then there is also a *directed path*, i.e. a directed walk in which all vertices $j, i_1, \ldots, i_{l-1}, i$ are distinct [5].

If there is a directed walk from j to i, then there is also a directed path, i.e. a directed walk in which all vertices j, i1, … , il-1, i are distinct [5].
如果存在從 j 到 i 的有向行走，那麼也存在有向路徑，即所有頂點 j, i1, ⋯ , il-1, i 都是不同的有向行走 [5]。

Lemma 3. For any h(A) defined by (1) and i ≠ j, (∇h(A))ij > 0 if and only if there exists a directed walk from j to i in G.
引理 3. 對於由 (1) 定義的任何 h(A) 且 i ≠ j，(∇h(A))ij > 0 當且僅當在 G 中存在從 j 到 i 的有向遊走。

The gradient ∇h(A) can also be used to characterize acyclicity, which will prove useful in the sequel.
梯度 ∇h(A) 也可用於表徵非循環性，這將在後續中證明是有用的。

Lemma 4. A directed graph G is acyclic if and only if the Hadamard product A。∇h(A) = 0 for any h defined by (1).
引理 4. 有向圖 G 是無環的當且僅當 Hadamard 乘積 A。∇h(A) = 0 對於由 (1) 定義的任何 h。

With the help of Lemma 3, we can give a simple graphical interpretation of Lemma 4: If a directed graph is acyclic, then for every pair (i, j), we must either not have an edge from i to j, i.e. Aij = 0, or not have a return path from j to i, i.e. (∇h(A))ij = 0.
借助引理 3，我們可以給出引理 4 的簡單圖形解釋：如果有向圖是無環的，那麼對於每一對 (i, j)，我們必須要么沒有從 i 到 j 的邊，即 Aij = 0，或者沒有從 j 到 i 的返迴路徑，即 (∇h(A))ij = 0。

3 Analysis of continuous acyclicity-constrained optimization - 連續無週期約束優化分析

In the remainder of the paper, we address the problem of learning a Bayesian network (a probabilistic directed graphical model) for the joint distribution of a d-dimensional random vector X, given a data matrix of n samples X ∈ Rnxd.
在本研究的其他部分，我們解決了學習貝葉斯網絡在概率有向圖模型上的問題，該網絡用於 d 維隨機向量 X 的聯合分佈，給定 n 個樣本 X ∈ Rnxd 的數據矩陣。

$\mathbf{X} \in \mathbb{R}^{n \times d}$.

We assume that the Bayesian network is parametrized by a matrix W ∈ Rdxd such that the sparsity pattern of W corresponds to the adjacency pattern of the graph: Wij ≠ 0 if and only if (i, j) ∈ E.
我們假設貝葉斯網絡由矩陣 W ∈ Rdxd 參數化，使得 W 的稀疏模式對應於圖的鄰接模式：Wij ≠ 0 當且僅當 (i, j) ∈ E。

In other words, each edge is associated with a single parameter Wij.
換句話說，每條邊都與單個參數 Wij 相關聯。

The most straightforward instance of this setting is a linear structural equation model (SEM) given by Xj = WT.jX + zj , where W.j is the jth column of W and zj is random noise.
此設置的最直接實例是由 Xj = WT.jX + zj 給出的線性結構方程模型 (SEM)，其中 W.j 是 W 的第 j 列，zj 是隨機噪聲。

More general models such as generalized linear models E [Xj |X] = g(WT.jX) are also included.
還包括更通用的模型，例如廣義線性模型 E [Xj |X] = g(WT.jX)。

While we experiment only with continuous variables in Section 5, it is straightforward to accommodate binary variables as well:
雖然我們在第 5 節中只對連續變量進行了實驗，但也可以直接容納二元變量：

in a generalized linear structural equation, a single parameter Wij can account for the effect of a binary input variable Xi, while a suitable link function g (e.g. logistic) can be used for a binary output Xj.
在廣義線性結構方程中，單個參數 Wij 可以解釋二進制輸入變量 Xi 的影響，而合適的鏈接函數 g（例如邏輯）可用於二進制輸出 Xj。

This section analyzes the continuous optimization problem of minimizing a score function F(W) subject to the acyclicity constraint h(A) = 0 for any h defined by (1) (thanks to Theorem 1).
本節分析了在非循環約束 h(A) = 0 下，對於由 (1) 定義的任何 h 最小化得分函數 F(W) 的連續優化問題（感謝定理 1）。

For simplicity, it is assumed in this section that F(W) is continuously differentiable, although it is not hard to extend the analysis to account for an l1 penalty as in (13).
為簡單起見，在本節中假設 F(W) 是連續可微的，儘管不難擴展分析以解釋 (13) 中的 l1 懲罰。

We consider two ways of defining a weighted adjacency matrix A from W.
我們考慮從 W 定義加權鄰接矩陣 A 的兩種方法。

Section 3.1 re-examines the quadratic case A = W。W proposed in [33] and sheds light on their augmented Lagrangian algorithm.

第 3.1 節重新檢查了 [33] 中提出的二次情況 A = W。W 並闡明了他們的增廣拉格朗日算法。

We then propose and study the absolute value case A = |W| in Section 3.2.
然後我們提出並研究絕對值情況 A = |W| 在第 3.2 節中。

3.1 Quadratic adjacency matrix 二次鄰接矩陣

With A = W。W as the element-wise square of W, the optimization problem is min … .
以 A = W。W 作為 W 的元素平方，優化問題是 min ... 。

$$\min_{W} \quad F(W) \quad \text{s.t.} \quad h(W \circ W) \leq 0. \qquad (4)$$

The constraint h(W。W) ≤ 0 is equivalent to h(W。W) = 0 because h(A) ≥ 0 for non-negative A, as seen from (1).
約束 h(W。W) ≤ 0 等價於 h(W。W) = 0 因為 h(A) ≥ 0 對於非負 A，如從 (1) 中看到的。

The matrix exponential case of (4) with h(A) as in (2) was proposed in [33].
[33] 中提出了 (4) 的矩陣指數情況和 (2) 中的 h(A)。

Applying Lemma 4 yields the following consequence.
應用引理 4 產生以下結果。

Lemma 5. Let W be a feasible solution to problem (4). Then ∇W(h(W。W)) = 0.
引理 5. 令 W 是問題 (4) 的可行解。 則∇W(h(W。W)) = 0。

The vanishing gradient in Lemma 5 has theoretical and practical implications.
引理 5 中的消失梯度具有理論和實踐意義。

First, the Karush-Kuhn-Tucker (KKT) conditions of optimality [4] for problem (4), namely

∇F(W) + λ∇w(h(W。W)) = 0 (5)

with Lagrange multiplier λ≥0 , are not satisfied for any feasible solution, let alone a local minimum, except in a trivial case.
首先，問題（4）的 Karush-Kuhn-Tucker（KKT）條件[4]，即

∇F(W) + λ∇w(h(W。W)) = 0 (5)

$$\nabla F(W) + \lambda \nabla_W (h(W \circ W)) = 0 \qquad (5)$$

與拉格朗日乘子 λ≥0 ，不滿足任何可行的解決方案，更不用說局部最小值，除非是微不足道的情況。

Proposition 2. Let W be a feasible solution to problem (4).
命題 2. 設 W 是問題 (4) 的可行解。

Then unless W is an unconstrained stationary point of F(W), i.e. ∇F(W) = 0, the KKT condition (5) cannot hold.
那麼除非 W 是 F(W) 的無約束駐點，即 ∇F(W) = 0，否則 KKT 條件 (5) 不能成立。

In particular if F(W) is convex, the condition ∇F(W) = 0 holds only for unconstrained minimizers of F(W), so if these solutions are already acyclic, there is nothing more to be done.
特別是如果 F(W) 是凸的，條件 ∇F(W) = 0 僅適用於 F(W) 的無約束極小值，所以如果這些解已經是無環的，就沒有什麼可做的了。

On the practical side, Lemma 5 sheds light on the augmented Lagrangian algorithm proposed in [33].
在實踐方面，引理 5 闡明了 [33] 中提出的增強拉格朗日算法。

The augmented Lagrangian corresponding to (4) with penalty parameters α and ρ is
F(W) + σ h(W。W)+(ρ /2)h((W。W)^2), (6)
with gradient
∇F(W) + (α + ρ h(W。W)) ∇W(h(W。W)).
對應於具有懲罰參數α和ρ的（4）的增廣拉格朗日是具有梯度∇F(W)的 F(W)。

Proposition 3. 提案 3
Let W be a feasible solution to problem (4).
設 W 是問題 (4) 的可行解。

Then unless W is an unconstrained stationary point of F(W), i.e. ∇F(W) = 0, W cannot be a stationary point of the augmented Lagrangian (6).
那麼除非 W 是 F(W) 的無約束駐點，即 ∇F(W) = 0，W 不能是增廣拉格朗日 (6) 的駐點。

Proposition 3 explains the following observed behavior of the augmented Lagrangian algorithm,
命題 3 解釋了以下觀察到的增廣拉格朗日算法的行為，

namely that it does not converge to an exactly (or within machine precision) feasible solution of (4) even when the penalty parameters, are very high (ρ ~10^16).
意味著即使懲罰參數非常高（ ρ ~10^16），它也不會收斂到（4）的精確（或機器精度內）可行解。

The reason is that a minimizer of the augmented Lagrangian (6) cannot be a feasible solution to (4) except in the trivial case discussed above.
原因是增廣拉格朗日 (6) 的極小值不能是 (4) 的可行解，除非在上面討論的微不足道的情況下。

However, when α and ρ are very large, minimizers of (6) do tend to have gradients ∇W(h(W。W)) ≈ 0,

and accordingly h(W。W) ≈ 0 by continuity.
然而，當 α 和 ρ 非常大時，(6) 的極小值確實傾向於具有梯度 ∇W(h(W。W)) ≈ 0，因此通過連續性，h(W。W) ≈ 0。

Thus as α and ρ increase, the augmented Lagrangian algorithm yields solutions that are closer and closer to being feasible.
因此，隨著 α 和 ρ 的增加，增強拉格朗日算法會產生越來越接近可行的解決方案。

3.2 Absolute value adjacency matrix 絕對值鄰接矩陣

As an alternative, we consider defining adjacency matrix A as the absolute value of W, A = |W|, leading to the following constrained optimization:
作為替代方案，我們考慮將鄰接矩陣 A 定義為 W 的絕對值，A = |W|，導致以下約束優化：

$$\min_{W} \quad F(W) \quad \text{s.t.} \quad h(|W|) \leq 0. \tag{7}$$

Formulation (7) is motivated in part by the failure to satisfy KKT conditions in Section 3.1 and in part by the connection between the absolute value function/`1 norm and sparsity, which is needed for acyclicity.
公式（7）部分是由於無法滿足第 3.1 節中的 KKT 條件，部分是由於絕對值函數/`1 範數與非循環性所需的稀疏性之間的聯繫。

While it will be seen that (7) has different theoretical and numerical properties from (4), the two formulations are equivalent in a sense because acyclicity depends only on the sparsity pattern of W, which is clearly the same regardless of whether |W| or W。W is used.
雖然可以看出（7）與（4）具有不同的理論和數值性質，但兩種公式在某種意義上是等價的，因為非循環性僅取決於 W 的稀疏模式，無論 |W| 是否相同，W 顯然是相同的。 或 W。W 被使用。

3.2.1 An equivalent smooth optimization 等效的平滑優化

Problem (7) is not a smooth optimization because of the absolute value function.
由於絕對值函數，問題 (7) 不是平滑優化。

To avoid any issues with continuous differentiability, we make use of the following alternative formulation, which we show in Appendix A.3 to be equivalent to (7):
為避免連續可微性出現任何問題，我們使用以下替代公式，我們在附錄 A.3 中顯示其等效於 (7)：

$$\min_{W^+,W^-} \quad F\left(W^+ - W^-\right) \quad \text{s.t.} \quad h\left(W^+ + W^-\right) \leq 0, \quad W^+, W^- \geq 0. \tag{8}$$

Given any solution (W+, W-) to (8), a solution to (7) is obtained simply as W = W+ - W-.
給定 (W+, W-) 到 (8) 的任何解， (7) 的解可以簡單地通過 W = W+ - W- 獲得。

3.2.2 KKT conditions and constraint qualification - KKT 條件和約束條件

We proceed to analyze the KKT conditions for the smooth reformulation (8), which are as follows:
我們繼續分析平滑重構 (8) 的 KKT 條件，如下所示：

$$\pm\nabla F\left(W^{+}-W^{-}\right)+\lambda\nabla h\left(W^{+}+W^{-}\right)=M^{\pm}\geq 0 \qquad (9a)$$

$$W^{\pm}\circ M^{\pm}=0, \qquad (9b)$$

in addition to the feasibility conditions in (8).
除了（8）中的可行性條件。

The +- versions of (9a) result from taking gradients with respect to W+ and W- respectively, where $\lambda \geq 0$ is a Lagrange multiplier.
(9a) 的 +- 版本分別取自關於 W+ 和 W- 的梯度，其中 λ≥ 0 是拉格朗日乘數。

M+, M- are non-negative matrices of Lagrange multipliers corresponding to the non-negativity constraints in (8), with complementary slackness conditions (9b).
M+、M- 是與（8）中的非負約束相對應的拉格朗日乘子的非負矩陣，具有互補鬆弛條件（9b）。

As in Section 3.1, we must consider whether the KKT conditions are necessary conditions of optimality, i.e. whether a local minimum must satisfy them.
與 3.1 節一樣，我們必須考慮 KKT 條件是否是最優性的必要條件，即局部最小值是否必須滿足它們。

Theorem 6 gives an affirmative answer;
定理 6 給出了肯定的答案；

however, it turns out that common constraint qualifications used to establish necessity do not hold.
然而，事實證明，用於確定必要性的常見約束條件並不成立。

To begin, we recall that a feasible solution to an inequality-constrained problem such as (8) is said to be regular if the gradients of the active (i.e. tight) constraints are linearly independent.
首先，我們回想一下，如果活動（即緊）約束的梯度是線性無關的，則對諸如（8）這樣的不等式約束問題的可行解決方案被認為是規則的。

If a local minimum is regular, then the KKT conditions necessarily hold.
如果局部最小值是規則的，則 KKT 條件必然成立。

Proposition 4. A feasible solution (W+, W-) to problem (8) cannot be regular.
命題 4. 問題 (8) 的可行解 (W+, W-) 不可能是正則的。

Beyond regularity, we refer to the hierarchy of constraint qualifications presented in [4] and show that feasible solutions to (8) do not satisfy a weaker constraint qualification called quasinormality.
除了正則性之外，我們參考了 [4] 中提出的約束條件的層次結構，並表明 (8) 的可行解不滿足稱為準正態性的較弱約束條件。

Proposition 5. A feasible solution (W+, W-) to problem (8) cannot be quasinormal.
命題 5. 問題 (8) 的可行解 (W+, W-) 不可能是擬正規的。

In spite of these negative results, Appendix A.4 provides a direct proof that KKT conditions (9) are satisfied at a local minimum of (8).
儘管有這些負面結果，附錄 A.4 提供了直接證明 KKT 條件 (9) 在局部最小值 (8) 處得到滿足。

The proof uses the following lemma, which we highlight because of its graphical interpretation in terms of directed paths not being created/destroyed by the addition/removal of certain edges.
證明使用以下引理，我們強調它是因為它在有向路徑方面的圖形解釋不會因添加/刪除某些邊而被創建/破壞。

Lemma 6. For a non-negative matrix A, if $(\nabla h(A))_{ij} > 0$, changing the values of A_{kj} for any k cannot make $(\nabla h(A))_{ij} = 0$.
引理 6. 對於非負矩陣 A，如果 $(\nabla h(A))_{ij} > 0$，則改變任意 k 的 A_{kj} 值不能使 $(\nabla h(A))_{ij} = 0$。

Similarly if $(\nabla h(A))_{ij} = 0$, changing the values of A_{kj} for any k cannot make $(\nabla h(A))_{ij} > 0$.
類似地，如果 $(\nabla h(A))_{ij} = 0$，則改變任何 k 的 A_{kj} 值都不能使 $(\nabla h(A))_{ij} > 0$。

Theorem 6. Let (W+,W-) be a local minimum of problem (8).
定理 6. 令 (W+,W-) 是問題 (8) 的局部最小值。

Then there exist a Lagrange multiplier $\lambda > 0$ and matrices M+ > 0, M->0 satisfying the KKT conditions in (9).
則存在一個拉格朗日乘子 $\lambda > 0$ 且矩陣 M+ > 0, M->0 滿足 (9) 中的 KKT 條件。

3.2.3 Relationships with explicit edge absence constraints 與顯式邊缺失約束的關係

We now discuss relationships between the KKT conditions (9) and the optimality conditions for score optimization problems with explicit edge absence constraints, which correspond to zero-value constraints on the matrix W.
我們現在討論 KKT 條件 (9) 與具有顯式邊緣缺失約束的得分優化問題的最優性條件之間的關係，這對應於矩陣 W 上的零值約束。

Given a set Z of such constraints, we consider the problem min … and denote by W*(Z) an optimal solution.
給定一組這樣的約束 Z，我們考慮問題 min … 並用 W*(Z) 表示一個最優解。

$$\min_W F(W) \quad \text{s.t.} \quad W_{ij} = 0, \quad (i,j) \in \mathcal{Z} \tag{10}$$

The necessary conditions of optimality for (10) are (∇F(W))ij = 0 , Wij =0.
(10) 的最優性的必要條件是 (∇F(W))ij = 0 , Wij =0。

$$(\nabla F(W))_{ij} = 0, \quad (i,j) \notin \mathcal{Z}, \tag{11a}$$
$$W_{ij} = 0, \quad (i,j) \in \mathcal{Z}. \tag{11b}$$

In one direction, given a KKT point (W+, W-), we define the set P…(12) i.e. the set of (i, j) with directed walks from j to i, according to Lemma 3.
在一個方向上，給定一個 KKT 點 (W+, W-)，我們定義集合 P:={}(12)，即根據引理 3，從 j 到 i 的定向步行的 (i, j) 集合。

$$\mathcal{P} := \{(i,j) : (\nabla h(W^+ + W^-))_{ij} > 0\}, \tag{12}$$

Lemma 7. If (W+,W-) satisfies the KKT conditions in (9), then W = W+ - W- satisfies the optimality conditions in (11) for Z = P.
引理 7. 如果 (W+,W-) 滿足 (9) 中的 KKT 條件，則 W = W+ - W- 滿足 (11) 中 Z = P 的最優性條件。

If in addition F(W) is convex, then W is a minimizer of (10) for Z = P.
此外，如果 F(W) 是凸的，則 W 是 (10) 的最小化器，因為 Z = P。

Under the assumption that F is convex, we can use Lemma 7 to show that the KKT conditions (9) are sufficient for local minimality in (7), despite the constraint h(|W|)≤0 not being convex.
在 F 是凸的假設下，我們可以使用引理 7 來證明 KKT 條件（9）足以滿足（7）中的局部極小性，儘管約束 h(|W|)≤0 不是凸的。

Theorem 7. Assume that F(W) is convex. Then if (W+;W□) satisfies the KKT conditions in (9), W = W+ □W□ is a local minimum for problem (7).
定理 7. 假設 F(W) 是凸的。 然後，如果 (W+;W) 滿足 (9) 中的 KKT 條件，則 W = W+ W 是問題 (7) 的局部最小值。

In the opposite direction of Lemma 7, we focus on the case in which a minimizer W of (10) is feasible, i.e. h(A*(Z)) = 0 for A= |W*(Z)|.
在引理 7 的相反方向，我們專注於 (10) 的最小化 W 可行的情況，即 h(A*(Z)) = 0 for A= |W*(Z)|。

Then by Lemma 4, we must have (W*(Z))ij = 0 wherever (∇h(A*(Z)))ij > 0.
然後根據引理 4，無論 (∇h(A*(Z)))ij > 0，我們都必須有 (W*(Z))ij = 0。

If Z does not include such a pair (i, j), we may add (i, j) to Z while preserving the optimality of the existing

solution W*(Z) with respect to (10) (since it already satisfies the new constraint Wij = 0).
如果 Z 不包括這樣的一對 (i, j)，我們可以將 (i, j) 添加到 Z，同時保持現有解 W*(Z) 關於 (10) 的最優性（因為它已經滿足新的 約束 Wij = 0)。

Hence for feasible W*(Z), we adopt the convention that all (i, j) with (W*(Z))ij = 0 and $(\nabla h(A^*(Z)))_{ij}>0$ are included in Z.
因此，對於可行的 W*(Z)，我們採用所有 (i, j) 且 (W*(Z))ij = 0 和 (∇h(A*(Z)))ij>0 都包含在 Z 中的約定。

We call Z irreducible if it contains only pairs (i, j) for which (∇h(A*(Z)))ij> 0.
如果 Z 只包含 (i, j) 對 (∇h(A*(Z)))ij> 0，我們稱 Z 不可約。

Theorem 8. If a minimizer W*(Z) of (10) is feasible and Z is irreducible, then W+ = (W*(Z))+, W- = (W*(Z))_ satisfy the KKT conditions in (9) .
定理 8. 如果 (10) 的最小化器 W*(Z) 可行且 Z 不可約，則 W+ = (W*(Z))+, W- = (W*(Z))_ 滿足 KKT 條件 (9) 。

If W*(Z) is feasible but Z is not irreducible, then the following result guarantees that Z may be reduced to an irreducible set without losing feasibility.
如果 W*(Z) 是可行的，但 Z 不是不可約的，那麼下面的結果保證了 Z 可以減少到一個不可約的集合而不會失去可行性。

We assume that F(W) is separable (decomposable) as the following sum: F(W) = … .
我們假設 F(W) 是可分離的（可分解的），其總和如下： F(W) = ⋯ 。

$$F(W) = \sum_{j=1}^{d} F_j(W_{\cdot j}).$$

Lemma 8. Assume that the score function F(W) is separable.
引理 8. 假設評分函數 F(W) 是可分離的。

Suppose that W*(Z) in (10) is feasible and Z0(j) = {(i1, j), … , (iJ , j)} ⊆ Z is a subset for which$(\nabla h(A^*(Z)))_{ij}$ = 0, (i, j) ∈ Z0(j).
假設 (10) 中的 W*(Z) 是可行的並且 Z0(j) = {(i1, j), … , (iJ , j)} ⊆ Z 是其中 (∇h(A*(Z)))ij = 0, (i, j) ∈ Z0(j)。

Then W*(Z\Z0(j)) is also feasible.
那麼 W*(Z\Z0(j))也是可行的。

Since the removal of a constraint (i, j) ∈ Z for which $(\nabla h(A^*(Z)))_{ij}$ =0 does not affect feasibility, we call such a constraint unnecessary as a somewhat colloquial shorthand.
由於去除約束 (i, j) ∈ Z (∇h(A*(Z)))ij =0 並不影響可行性，我們稱這種約束是不必要的，因為它是一種

有點口語化的速記。

Lemma 8 removes a set of pairs (i, j) ∈ Z0(j) from Z for which (∇h(A*(Z)))ij =0 while maintaining feasibility.
引理 8 從 Z 中移除一組對 (i, j) ∈ Z0(j) ，其中 (∇h(A*(Z)))ij =0 同時保持可行性。

The resulting set is then checked again for irreducibility.
然後再次檢查結果集的不可約性。

Since each application of Lemma 8 removes at least one pair from Z, the re-optimization (20) has to be performed at most |Z| times to ensure a irreducible set.
由於引理 8 的每次應用都會從 Z 中刪除至少一對，因此最多必須執行重新優化（20）|Z|。 次以確保不可約集。

The development in this subsection suggests the meta-algorithm in Algorithm 1, which we refer to as KKT-informed local search.
本小節的發展暗示了算法 1 中的元算法，我們將其稱為 KKT 知情本地搜索。

An instantiation is described in Section 4.2.
第 4.2 節描述了實例化。

Algorithm 1 KKT-informed local search (KKTS)
算法 1 KKT-informed local search (KKTS)

Theorem 9. If F(W) is separable, KKT-informed local search yields a solution satisfying the KKT conditions (9).
定理 9。如果 F(W) 是可分離的，則 KKT 通知的局部搜索會產生滿足 KKT 條件 (9) 的解。

The development in this subsection suggests the meta-algorithm in Algorithm 1, which we refer to as KKT-informed local search. An instantiation is described in Section 4.2.

Algorithm 1 KKT-informed local search (KKTS)

Require: Initial set $\mathcal{Z}$ of edge absence constraints. Solve (10).
1: **while** $W^*(\mathcal{Z})$ infeasible **do**
2: Select edge(s) in cycle $((W^*(\mathcal{Z}))_{ij} \neq 0, \left(\nabla h(A^*(\mathcal{Z}))\right)_{ij} > 0)$. Add to $\mathcal{Z}$. Re-solve (10).
3: **end while**
4: **while** $\mathcal{Z}$ reducible **do**
5: Remove one or more unnecessary constraints $(i, j) \in \mathcal{Z}$ for which $\left(\nabla h(A^*(\mathcal{Z}))\right)_{ij} = 0$ (see Lemma 8). Re-solve (10).
6: **end while**

Theorem 9. *If $F(W)$ is separable, KKT-informed local search yields a solution satisfying the KKT conditions* (9).

When combined with Theorem 7 and a convex F(W), Theorem 9 guarantees that KKT-informed local search will result in local minima.
當結合定理 7 和凸 F(W) 時，定理 9 保證 KKT 通知的局部搜索將導致局部最小值。

However, due to the non-convex constraint, the quality of such local minima is highly dependent on the particular instantiation of the meta-algorithm.
然而，由於非凸約束，這種局部最小值的質量高度依賴於元算法的特定實例。

Section 5 shows for example that the choice of initialization plays a large role.
例如，第 5 節顯示初始化的選擇起著重要作用。

4 Algorithms 演算法

For the algorithms in this section, we let the score function F(W) be the sum of a smooth loss function l(W;X) with respect to the data X and an l1 penalty to promote overall sparsity, as in [33]:
對於本節中的算法，我們讓得分函數 F(W) 是關於數據 X 的平滑損失函數 l(W;X) 和 l1 懲罰的總和，以提高整體稀疏性，如 [33] ：

$$F(W) = \ell(W; \mathbf{X}) + \tau \|W\|_1. \qquad (13)$$

4.1 Augmented Lagrangian with absolute value adjacency matrix - 具有絕對值鄰接矩陣的增廣拉格朗日

Formulation (7) naturally suggests an augmented Lagrangian algorithm as in [33] but with h(|W|) instead of h(W。W).
公式 (7) 自然地建議了一種如 [33] 中的增廣拉格朗日算法，但使用 h(|W|) 而不是 h(W。W)。

Using the (W+, W-) representation as in (8), the augmented Lagrangian minimized in each iteration is L(W+, W-, α, ρ), subject to W+ ≥ 0 and W- ≥ 0, where 1 is a vector of ones.
使用 (8) 中的 (W+, W-) 表示，在每次迭代中最小化的增廣拉格朗日為 L(W+, W-, α, ρ)，受 W+ ≥ 0 和 W- ≥ 0 影響，其中 1 是 的向量。

$$L(W^+, W^-, \alpha, \rho) = \ell(W^+ - W^-; \mathbf{X}) + \tau \mathbf{1}^T (W^+ + W^-)\mathbf{1} + \alpha h(W^+ + W^-) + \frac{\rho}{2} h(W^+ + W^-)^2,$$

The gradients with respect to W+, W- are given by ∇w… :
關於 W+、W- 的梯度由 ∇w... 給出：

$$\nabla_{W^\pm} L(W^+, W^-, \alpha, \rho) = \pm \nabla \ell(W^+ - W^-; \mathbf{X}) + \tau \mathbf{1}\mathbf{1}^T + \left(\alpha + \rho h(W^+ + W^-)\right) \nabla h(W^+ + W^-).$$

We otherwise closely follow the algorithm in [33].
否則，我們密切關注 [33] 中的算法。

4.2 KKT-informed local search - KKT 通知的本地搜索

We now describe an instantiation of the KKT-informed local search meta-algorithm in Algorithm 1.
我們現在描述算法 1 中 KKT 通知的本地搜索元算法的實例。

This involves initializing the set Z of edge absence constraints, selecting edges for removal (line 2), reducing unnecessary constraints (line 5), and re-solving (10).
這包括初始化邊缺失約束的集合 Z、選擇要移除的邊（第 2 行）、減少不必要的約束（第 5 行）和重新求解（10）。

We also discuss an additional operation of reversing edges, which is not part of Algorithm 1 but helps in attaining better local minima.
我們還討論了反向邊緣的附加操作，它不是算法 1 的一部分，但有助於獲得更好的局部最小值。

Initializing Z We allow any matrix W to serve as an initial solution.
初始化 Z 我們允許任何矩陣 W 作為初始解。

To define the set Z, we set to zero elements in W that are smaller than a threshold ! in absolute value.
為了定義集合 Z，我們將 W 中小於閾值的元素設為零！ 絕對值。

We then let Z = f(i; j) : Wij = 0g.
然後我們讓 Z = f(i; j) : Wij = 0g。

Selecting edges for removal (line 2) There are many possible ways of selecting edges to break cycles.
選擇要移除的邊（第 2 行） 有多種可能的方法可以選擇邊以中斷循環。

Here we consider an approach of minimizing the Lagrangian F(W)+ h(|W|) of (7) subject to the existing constraints Wij = 0 for (i, j) ∈ Z.
在這裡，我們考慮一種在現有約束 Wij = 0 下（i, j）∈ Z 最小化（7）的拉格朗日 F(W)+ h(|W|) 的方法。

The Lagrangian thus trades off minimizing the score function against reducing infeasibility.
因此，拉格朗日函數在最小化得分函數與降低不可行性之間進行權衡。

For α = 0, the minimizer is the existing solution W*(Z), and as α increases, weights Wij will be set to zero to decrease the infeasibility penalty h(|W|).
對於 α = 0，最小化器是現有解 W*(Z)，隨著 α 的增加，權重 Wij 將設置為零以減少不可行性懲罰 h(|W|)。

We implement a computationally simple version of the above idea.
我們實現了上述想法的計算簡單版本。

First, h(A) = h(|W|) in the Lagrangian is linearized around A*(Z) = |W*(Z)| as h(A) ≈h(A*(Z))+[∇h(A*(Z)), A – A*(Z)].
首先，拉格朗日函數中的 h(A) = h(|W|) 圍繞 A*(Z) = |W*(Z)| 線性化 如 h(A) ≈h(A*(Z))+[∇h(A*(Z)), A – A*(Z)]。

$$h(A) \approx h\big(A^*(\mathcal{Z})\big) + \big\langle \nabla h\big(A^*(\mathcal{Z})\big), A - A^*(\mathcal{Z}) \big\rangle .$$

After dropping constant terms and expanding the inner product, the constrained, linearized Lagrangian to be minimized is as follows:
在删除常数项並展開內積之後，要最小化的受約束的線性化拉格朗日如下：

$$\min_{W} \; F(W) + \alpha \sum_{(i,j):i\neq j} \big(\nabla h(A^*(\mathcal{Z}))\big)_{ij} |W_{ij}| \quad \text{s.t.} \quad W_{ij} = 0, \quad (i,j) \in \mathcal{Z}. \qquad (14)$$

Problem (14) is a score minimization problem with a weighted l1 penalty and zero-value constraints, i.e. the corresponding parameters are simply absent.
問題 (14) 是一個帶有加權 l1 懲罰和零值約束的分數最小化問題，即相應的參數根本不存在。

Furthermore, in the common case where F(W) is separable column-wise, (14) is also separable.
此外，在 F(W) 可按列分離的常見情況下，(14) 也是可分離的。

Second, α is increased from zero only until a single existing edge (i, j) (with (W*(Z))ij $\neq$ 0) belonging to a cycle (∇h(Z))ij > 0) is set to zero.
其次，α 僅從零增加，直到屬於循環 (∇h(Z))ij > 0) 的單個現有邊 (i, j)（其中 (W*(Z))ij $\neq$ 0）設置為零 。

This involves following the solution path of (14) defined by from W*(Z) at α = 0 until the first additional edge is removed.
這涉及遵循由 α = 0 處的 W*(Z) 定義的 (14) 的求解路徑，直到移除第一條附加邊。

If l(W;X) in (13) is the least-squares loss, the solution path is piecewise linear and we have implemented a modified version of the LARS algorithm [11] to efficiently track the path.
如果 (13) 中的 l(W;X) 是最小二乘損失，則求解路徑是分段線性的，我們已經實現了 LARS 算法 [11] 的修改版本以有效跟踪路徑。

The modification accounts for the non-uniformity of the weights (∇h(A*(Z)))ij , some of which may be zero, in the l1 penalty in (14).
修改解釋了權重 (∇h(A*(Z)))ij 的不均勻性，其中一些可能為零，在（14）的 l1 懲罰中。

It is described further in Appendix B.3.
在附錄 B.3 中進一步描述。

Reducing unnecessary constraints (line 5)
減少不必要的約束（第 5 行）

We also refer to this step as restoring edges (“restore” because these edges were likely present in an earlier iteration when W was denser), in analogy with the previous step which removes edges.
我們也將此步驟稱為恢復邊緣（ “恢復” ，因為當 W 更密集時，這些邊緣可能出現在較早的迭代中），類似於前一步驟刪除邊緣。

When there are multiple unnecessary constraints, the order in which they are removed can matter because the removal of constraints and re-optimization of (10) can make previously unnecessary constraints necessary.
當有多個不必要的約束時，刪除它們的順序可能很重要，因為約束的刪除和（10）的重新優化可能會使以前不必要的約束成為必要。

Because of this, even though Lemma 8 allows for multiple unnecessary constraints (i1, j), …, (iJ , j) to be removed at a time, we opt to do so more gradually, only one at a time.
因此，即使引理 8 允許一次刪除多個不必要的約束 (i1, j), ⋯, (iJ, j)，我們還是選擇逐步刪除，一次只刪除一個。

To decide among multiple unnecessary constraints (i, j), we greedily choose one for which the absolute partial derivative of the loss function, |(∇l(W;X))ij |, is largest.
為了在多個不必要的約束 (i, j) 中做出決定，我們貪婪地選擇損失函數的絕對偏導數 |(∇l(W;X))ij | 最大的約束。

Since |(∇l(W;X))ij| is the marginal rate of decrease of the loss as the constraint Wij = 0 is relaxed, this strategy gives the largest marginal rate of decrease.
因為 |(∇l(W;X))ij| 是當約束 Wij = 0 放鬆時損失的邊際減少率，該策略給出了最大的邊際減少率。

We note also that if |(∇l(W;X))ij | ≤ τ, relaxing the constraint Wij = 0 does not change its value because Wij = 0 already satisfies the optimality conditions for minimizing F(W).
我們還注意到如果 |(∇l(W;X))ij | ≤ τ，放寬約束 Wij = 0 不會改變其值，因為 Wij = 0 已經滿足最小化 F(W) 的最優條件。

Reversing edges 反轉邊緣

In addition to removing and restoring edges, we consider reversing edges, which involves two operations:
除了去除和恢復邊，我們還考慮反轉邊，這涉及兩個操作：

adding (i, j) to Z to remove an existing edge (W*(Z))ij ≠ 0, and removing (j, i) from Z (which must have been a necessary constraint if W*(Z) is feasible, to avoid a 2-cycle) to introduce the opposite edge.
將 (i, j) 添加到 Z 以刪除現有邊 (W*(Z))ij ≠ 0，並從 Z 中刪除 (j, i)（如果 W*(Z) 可行，這必須是必要的約束， 以避免 2 循環）引入相反的邊緣。

In contrast to removing edges, which generally increases F(W) but decreases h(A), and restoring edges, which decreases F(W) and is guaranteed by Lemma 8 not to increase h(A), reversing edges does not necessarily decrease F(W) or h(A).
與去除邊緣相比，通常會增加 F(W) 但減少 h(A)，恢復邊緣會減少 F(W) 並由引理 8 保證不增加 h(A)，反向邊緣不一定會減少 F(W) 或 h(A)。

We therefore accept an edge reversal only if it decreases one of F(W), h(A) relative to the original direction and does not increase the other, and otherwise reject the reversal.
因此，只有當它相對於原始方向減小 F(W)、h(A) 之一併且不增加另一個時，我們才接受邊緣反轉，否則拒絕反轉。

There are many possible variations in when to perform edge reversals within Algorithm 1.
在算法 1 中何時執行邊緣反轉有許多可能的變化。

In our implementation, we restrict reversals to the second while-loop and alternate between restoring one edge (reducing Z by one) and attempting all possible reversals given the current state.
在我們的實現中，我們將反轉限制為第二個 while 循環，並在恢復一條邊（將 Z 減少 1）和嘗試給定當前狀態的所有可能反轉之間交替。

When there are multiple reversal candidates, similar to restoring edges, we evaluate the loss partial derivatives $|(\nabla l(W;X))_{ji}|$, this time associated with introducing the reverse edges (j, i), and proceed in order of decreasing $|(\nabla l(W;X))_{ji}|$.
當有多個反向候選時，類似於恢復邊，我們評估損失偏導數$|(\nabla l(W;X))_{ji}|$，這次與引入反向邊（j，i）相關，並按以下順序進行 遞減$|(\nabla l(W;X))_{ji}|$。

The edge reversal operation is made much more efficient by keeping a memory of previously attempted reversals that do not have to be attempted again for some time.
通過保留先前嘗試的反轉的記憶，而不必在一段時間內再次嘗試，從而使邊緣反轉操作更加高效。

When the reversal of edge (i, j) is attempted, it is recorded in the memory, and if the reversal is accepted, reversal of (j, i) is also added to the memory as it would revert to the previous inferior state.
當嘗試反轉邊 (i, j) 時，將其記錄在內存中，如果反轉被接受，則 (j, i) 的反轉也會添加到內存中，因為它會恢復到先前的劣等狀態。

The memory for (i, j) is cleared when either column i or j is updated since this may change the value of reversing (i, j).
當更新列 i 或 j 時，(i, j) 的內存會被清除，因為這可能會改變反轉 (i, j) 的值。

Re-solving (10) (lines 2, 5)
重新求解（10）（第 2、5 行）

Removing, restoring, and reversing edges all involve re-solving (10) after adding to Z, reducing Z, or both in the case of reversals.
重新求解 (10)（第 2、5 行） 移除、恢復和反轉邊都涉及在添加到 Z 後重新求解（10）、減少 Z 或在反轉的情況下兩者兼而有之。

When l(W;X) in (13) is the least-squares loss, these re-optimizations can be done efficiently using the LARS

algorithm.
當 (13) 中的 l(W;X) 是最小二乘損失時，可以使用 LARS 算法有效地完成這些重新優化。

In the case of adding (i, j) to Z, an increasing penalty is imposed on |Wij |, while in the case of removing (i, j) from Z, a penalty equivalent to the constraint Wij = 0 is inferred and then decreased to zero.
在將 (i, j) 添加到 Z 的情況下，對 |Wij | 施加遞增的懲罰，而在從 Z 中刪除 (i, j) 的情況下，推斷出等效於約束 Wij = 0 的懲罰，然後 減少到零。

Further details are in Appendix B.
更詳細的信息在附錄 B 中。

5 Experiments 實驗

We compare the structure learning performance of the following base algorithms: NOTEARS [33], the FGS implementation [22] of GES [7], MMHC [29], PC [28], augmented Lagrangian with absolute value adjacency matrix A = |W| (Section 4.1, abbreviated 'Abs'), and KKT-informed local search (Section 4.2, KKTS) initialized with the unconstrained solution (Z = {(i, i), i ∈ V} just to avoid self-loops).
我們比較了以下基本算法的結構學習性能：NOTEARS [33]、GES [7] 的 FGS 實現 [22]、MMHC [29]、PC [28]、具有絕對值鄰接矩陣的增廣拉格朗日矩陣 A = |W | （第 4.1 節，縮寫為 "Abs"），以及使用無約束解（Z = {(i, i), i ∈ V} 只是為了避免自循環）初始化的 KKT 通知局部搜索（第 4.2 節，KKTS）。

We also experimented with CAM [6] but defer those results to Appendix C.5 as we found them less competitive in the tested settings.
我們還對 CAM [6] 進行了試驗，但將這些結果推遲到附錄 C.5 中，因為我們發現它們在測試設置中的競爭力較低。

In addition, we use each of the above base algorithms to initialize KKTS (denoted by appending '-KKTS' and excepting KKTS itself).
此外，我們使用上述每種基本算法來初始化 KKTS（由附加 "-KKTS" 表示，KKTS 本身除外）。

Algorithm parameter settings are detailed in Appendix C.1.
算法參數設置詳見附錄 C.1。

Of note are the default termination tolerance on h, Є = 10^(-10), and the threshold on W, ω = 0:3 following [33], applied after NOTEARS, Abs, and KKTS as well as to initialize Z before KKTS.
值得注意的是 h 上的默認終止容差，Є = 10^(-10)，以及 W 上的閾值，ω = 0:3 跟隨 [33]，在 NOTEARS、Abs 和 KKTS 之後應用以及在之前初始化 Z KKTS。

The experimental setup is similar to [33].
實驗設置類似於[33]。

In brief, random Erdös-Rényi or scale-free graphs are generated with kd expected edges (denoted ERk or SFk), and uniform random weights W are assigned to the edges.
簡而言之，隨機 Erdös-Rényi 或無標度圖是用 kd 條預期邊（表示為 ERk 或 SFk）生成的，並且均勻的隨機權重 W 被分配給這些邊。

Data X ∈ R^(nxd) is then generated by taking n i.i.d. samples from the linear SEM X = (W^T)X + z, where z is either Gaussian, Gumbel, or exponential noise.
然後通過取 n i.i.d 生成數據 X ∈ R^(nxd)。 來自線性 SEM 的樣本 X = (W^T)X + z，其中 z 是高斯、Gumbel 或指數噪聲。

100 trials are performed for each graph type-noise type combination, which is an order of magnitude larger than in e.g. [33, 31] and reduces the standard errors of the estimated means.
對每個圖形類型-噪聲類型組合執行 100 次試驗，這比例如 [33, 31] 並減少了估計均值的標準誤差。

Figure 1 shows structural Hamming distances (SHD) with respect to the true graph and running times for three graph-noise combinations and n = 1000.
圖 1 顯示了結構漢明距離 (SHD) 相對於真實圖和三個圖噪聲組合的運行時間和 n = 1000。

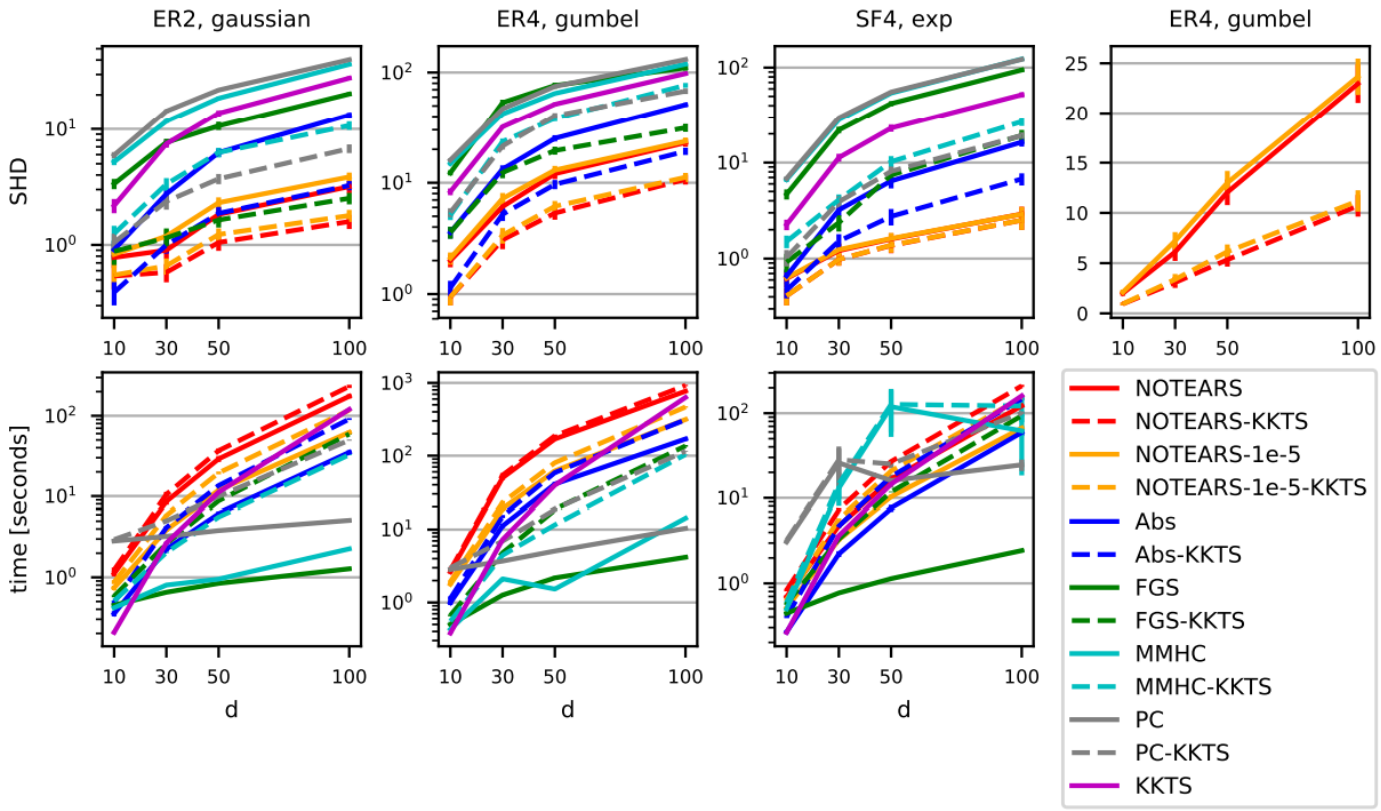
Figure 2 shows the same metrics and combinations for the more challenging setting n = 2d, with largely similar patterns.
圖 2 顯示了更具挑戰性的設置 n = 2d 的相同指標和組合，具有基本相似的模式。

Other graphnoise combinations, results in tabular form, and computing environment details are in Appendix C.
其他圖形噪聲組合、表格形式的結果和計算環境詳細信息在附錄 C 中。

Figure 1: Structural Hamming distances (SHD) with respect to true graph and solution times for n = 1000. Error bars indicate standard errors over 100 trials.
圖 1：結構漢明距離 (SHD) 相對於真實圖和 n = 1000 的求解時間。誤差條表示 100 次試驗的標準誤差。

Red lines overlap with orange in the SF4 SHD plot.
在 SF4 SHD 圖中，紅線與橙色重疊。

The upper right panel focuses on combinations with NOTEARS using a linear vertical scale.
右上角的面板側重於使用線性垂直刻度與 NOTEARS 的組合。



We focus first on the base algorithms (solid lines), of which NOTEARS is clearly the best in terms of SHD.(#1)
我們首先關注基本算法（實線），其中 NOTEARS 在 SHD 方面顯然是最好的。

Abs is next and better than FGS, MMHC, and PC.
Abs 緊隨其後，並且優於 FGS、MMHC 和 PC。

We hypothesize that the smoothness of the quadratic adjacency A = W。W used by NOTEARS is better able to overcome non-convexity than the non-smooth A = |W| of Abs, which tends to force parameters Wij to zero, perhaps too soon.
我們假設 NOTEARS 使用的二次鄰接的平滑度 A=W。W 比非平滑的 A=|W|更能克服非凸性。 的 Abs，這往往會迫使參數 Wij 為零，也許為時過早。

The non-convexity is further reflected in the inferior performance of (pure) KKTS, which only takes local steps starting from the unconstrained solution.
非凸性進一步反映在（純）KKTS 的較差性能上，它僅從無約束解開始採取局部步驟。

We now turn to the '-KKTS' combinations (dashed lines). It is seen that KKTS, and the theoretical understanding it embodies, improve the SHD of all base algorithms (including CAM in Appendix C.5).
我們現在轉向 "-KKTS" 組合（虛線）。 可以看出，KKTS 及其所體現的理論理解改進了所有基本算法（包括附錄 C.5 中的 CAM）的 SHD。

The improvement is by at least a factor of 2, except when the SHD is already low (e.g. NOTEARS on SF4), and moreover is consistent across dimensions d.
改進至少是 2 倍，除非 SHD 已經很低（例如 SF4 上的 NOTEARS），而且在維度 d 上是一致的。

An ablation study in Appendix C.4 shows that both reducing unnecessary constraints and reversing edges contribute to the improvement.
附錄 C.4 中的消融研究表明，減少不必要的約束和反轉邊緣都有助於改進。

In the case of NOTEARS-KKTS, while Proposition 3 asserts that NOTEARS cannot yield an exactly feasible solution, let alone a KKT point, Figure 1 confirms that it yields high-quality nearly feasible solutions.
在 NOTEARS-KKTS 的情況下，雖然命題 3 斷言 NOTEARS 不能產生完全可行的解決方案，更不用說 KKT 點了，但圖 1 證實它產生了高質量的幾乎可行的解決方案。

NOTEARS is therefore well-suited as an initialization for KKTS, and combining them apparently results in new state-of-the-art accuracy.
因此，NOTEARS 非常適合作為 KKTS 的初始化，並且將它們結合起來顯然會產生新的最先進的準確性。

Furthermore, in an attempt to achieve feasibility, NOTEARS uses more augmented Lagrangian iterations and very large penalty parameters α and ρ .
此外，為了實現可行性，NOTEARS 使用更多的增強拉格朗日迭代和非常大的懲罰參數 α 和 ρ 。

The latter causes the augmented Lagrangian (6) to be poorly conditioned and optimization solvers for it to take longer to converge.
後者導致增廣拉格朗日 (6) 條件不佳，優化求解器需要更長的時間才能收斂。

Thus, to reduce solution time as well as satisfy KKT conditions, we terminate NOTEARS early with a higher h tolerance of $\epsilon = 10^{-5}$ before running KKTS.
因此，為了減少求解時間並滿足 KKT 條件，我們在運行 KKTS 之前以更高的 h 容差 $\epsilon = 10^{-5}$ 提前終止 NOTEARS。

Figure 1 shows that this results in nearly the same SHD improvement over NOTEARS while also taking considerably less time (except for SF4).
圖 1 顯示，這導致與 NOTEARS 幾乎相同的 SHD 改進，同時花費的時間也少得多（SF4 除外）。

Abs-KKTS similarly outperforms NOTEARS on ER graphs and takes even less time.
Abs-KKTS 在 ER 圖上同樣優於 NOTEARS，並且花費的時間更少。

####
1
The SHDs for NOTEARS and FGS in Figure 1 are much better than those reported in [33], by almost an order of magnitude in some cases.
圖 1 中 NOTEARS 和 FGS 的 SHD 比 [33] 中報告的要好得多，在某些情況下幾乎是一個數量級。

Part of the improvement is due to code updates for NOTEARS but the rest we cannot explain.

部分改進是由於 NOTEARS 的代碼更新，但其餘的我們無法解釋。

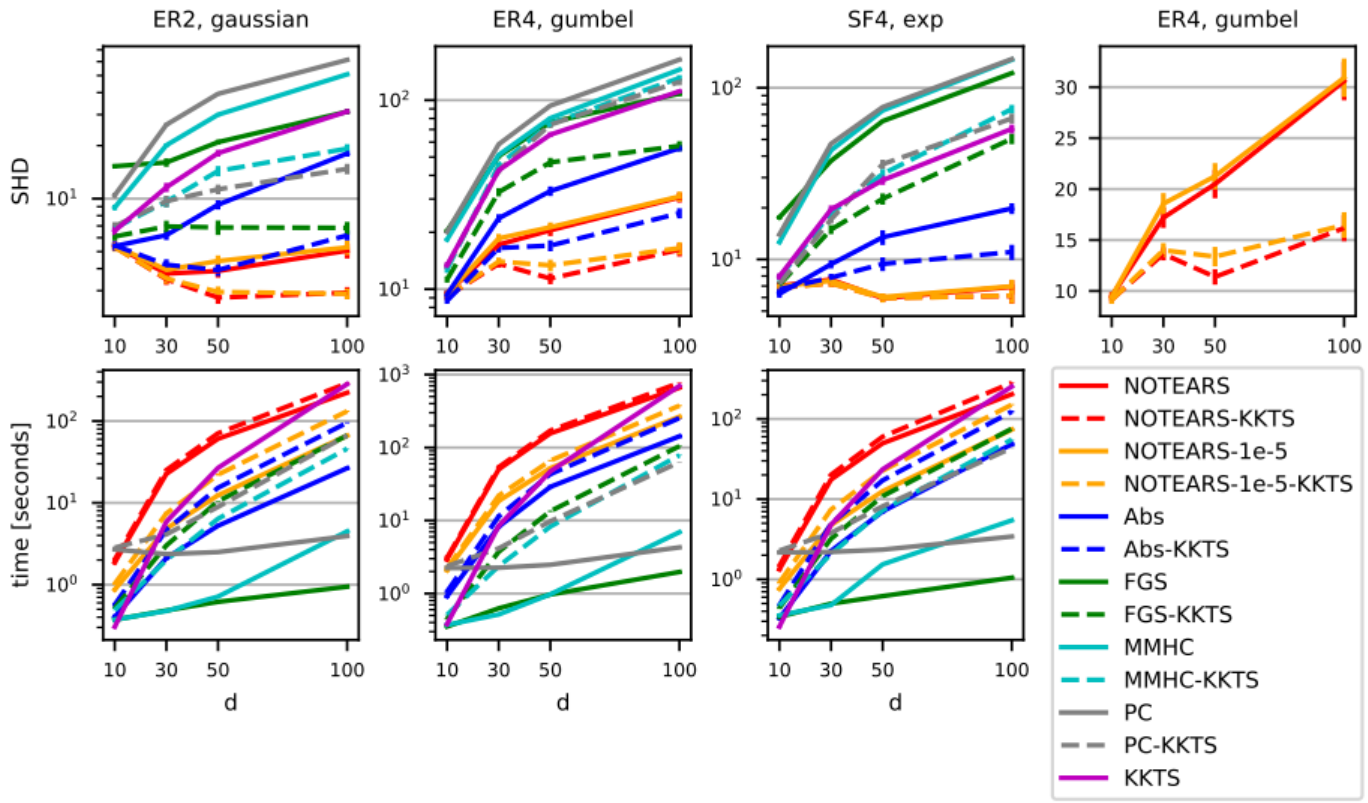
We also show in Appendix C.3 that subtracting the mean from X improves the SHD by a noticeable factor for some noise types.
我們還在附錄 C.3 中表明，對於某些噪聲類型，從 X 中減去平均值可以顯著提高 SHD。

All results in Figure 1 are obtained with zero-mean X.
圖 1 中的所有結果都是使用零均值 X 獲得的。

Figure 2: Structural Hamming distances (SHD) with respect to true graph and solution times for n = 2d.
圖 2：結構漢明距離 (SHD) 相對於真實圖和 n = 2d 的求解時間。

Red lines overlap with orange in the SF4 SHD plot.
在 SF4 SHD 圖中，紅線與橙色重疊。

The upper right panel focuses on combinations with NOTEARS using a linear vertical scale.
右上角的面板側重於使用線性垂直刻度與 NOTEARS 的組合。



6 Conclusion and future work - 結論和未來工作

We have re-examined a recently proposed continuous optimization framework for learning Bayesian networks.
我們重新審視了最近提出的用於學習貝葉斯網絡的持續優化框架。

Our most important contributions are as follows:
我們最重要的貢獻如下：

(1) better understanding of the NOTEARS formulation and algorithm of [33];
(1) 更好地理解[33]的 NOTEARS 公式和算法；

(2) analysis and understanding of the KKT optimality conditions for an equivalent reformulation (for which they do indeed hold);
(2) 對等價重構的 KKT 最優條件的分析和理解（它們確實成立）；

(3) a local search algorithm informed by the KKT conditions that significantly and universally improves the accuracy of NOTEARS and other algorithms.
(3) 一種由 KKT 條件通知的局部搜索算法，可顯著且普遍地提高 NOTEARS 和其他算法的準確性。

A clear next step is to generalize the theory and algorithms to the case in which each edge in the graph corresponds to multiple parameters.
明確的下一步是將理論和算法推廣到圖中的每條邊對應多個參數的情況。

One motivation is to allow nonlinear models; a nonlinear extension of the absolute value case of Section 3.2 could parallel the recent nonparametric extension [34] for the quadratic case.
一個動機是允許非線性模型； 3.2 節中絕對值情況的非線性擴展可以與最近的二次情況的非參數擴展 [34] 平行。

Another reason for having multiple parameters is to accommodate nonbinary categorical variables, which are typically encoded into multiple binary variables on the input side, or predicted using e.g. multi-logit regression [14] on the output side.
具有多個參數的另一個原因是容納非二進制分類變量，這些變量通常在輸入端編碼為多個二進制變量，或使用例如預測 輸出端的多邏輯回歸[14]。

Other future directions include improving the efficiency of algorithms for solving (4), (7) and exploring alternative acyclicity characterizations from Section 2.
其他未來方向包括提高求解 (4)、(7) 算法的效率以及探索第 2 節中的替代非循環特徵。

Broader Impact 更廣泛的影響

Bayesian networks are fundamentally about modeling the joint probability distribution of data, in a parsimonious and comprehensible manner.
貝葉斯網絡從根本上講是以簡約和易於理解的方式對數據的聯合概率分佈進行建模。

This work therefore contributes mostly to layer 0 (“foundational research”) in the “Impact Stack” of [3], particularly with regard to the theoretical aspects.

因此，這項工作主要對 [3] 的“影響堆棧”中的第 0 層（“基礎研究”）做出了貢獻，特別是在理論方面。

If one views Bayesian network structure learning as a “ML technique” rather than a“foundational technique”, then the algorithmic contribution also falls into layer 1.
如果將貝葉斯網絡結構學習視為一種“ML 技術”而不是“基礎技術”，那麼算法貢獻也屬於第 1 層。

We thus confine our discussion of broader impacts mostly to layers 0 and 1, i.e. “tractable” impacts according to [3], as it is difficult and perhaps inappropriate to speculate further.
因此，我們將我們對更廣泛影響的討論主要限於第 0 層和第 1 層，即根據 [3] 的“易處理”影響，因為進一步推測是困難且可能不合適的。

The predominant contribution of this work is to theoretical understanding of the optimization problem that is score-based structure learning, and specifically a continuous formulation thereof.
這項工作的主要貢獻是對基於分數的結構學習的優化問題的理論理解，特別是其連續公式化。

This understanding has resulted in improvements in accuracy (as measured by structural Hamming distance), and we expect that further improvements will be made in future work.
這種理解導致了準確性的提高（以結構漢明距離衡量），我們預計在未來的工作中將有進一步的改進。

We also believe that this understanding may lead to advances in computational efficiency as well, beyond the simple measure of terminating the NOTEARS algorithm early when it has no hope of reaching feasibility, or observing that the absolute value version (Abs) converges more quickly.
我們還認為，這種理解也可能導致計算效率的提高，而不僅僅是在沒有希望達到可行性時提前終止 NOTEARS 算法的簡單措施，或者觀察到絕對值版本 (Abs) 收斂得更快。

For example, new optimization algorithms may be proposed for problems (4) and/or (7) that take better advantage of their properties.
例如，可以針對問題 (4) 和/或 (7) 提出新的優化算法，以更好地利用它們的特性。

As the accuracy and scalability of Bayesian network structure learning continue to increase, we hope that it becomes an even more commonly used technique for modeling data than it is now.
隨著貝葉斯網絡結構學習的準確性和可擴展性不斷提高，我們希望它成為一種比現在更常用的數據建模技術。

We are particularly interested in its use as the first step in causal structure discovery, which may then facilitate other causal inference tasks.
我們對它作為因果結構發現的第一步特別感興趣，這可能會促進其他因果推理任務。

We recognize however that errors in structure learning may compound into potentially more serious

downstream errors.
然而，我們認識到結構學習中的錯誤可能會復合成潛在的更嚴重的下游錯誤。

This is an issue calling for further study.
這是一個需要進一步研究的問題。