arXiv:2302.02672v2 [stat.ML] 3 May 2023

潜变量模型与结构方程模型的可识别性：从线性到非线性

Aapo Hyv¨arinen，[1] Ilyes Khemakhem，[2] Ricardo Monti[2]1)赫尔辛基大学计算机科学系，芬兰2)伦敦大学学院盖茨比计算神经科学中心，英国

2023年5月4日

摘要

多元统计学中的一个老问题是，线性高斯模型通常不可识别，即参数无法被唯一估计。在因子（成分）分析中，因子的正交旋转不可识别；而在线性回归中，效应方向无法识别。对于这类线性模型，已有研究表明，潜变量的非高斯性可以带来可识别性。在因子分析中，这引出了独立成分分析；而对于效应方向的问题，非高斯版本的结构方程模型则能够解决这一问题。最近，我们进一步证明，即使是这类模型的一般非参数非线性版本，也可以进行估计。在这种情况下，仅有非高斯性还不够；但如果我们假设数据是时间序列，或者分布受到某些观测辅助变量的适当调制，那么模型就是可识别的。本文回顾线性与非线性情形下的可识别性理论，并讨论因子分析模型和结构方程模型。

关键词：可识别性；独立成分分析；结构方程模型；因子分析；解缠结；非高斯性

1 引言

本文旨在简洁且相对自成一体地阐述一类称为独立成分分析的潜变量模型，以及一类结构方程模型的可辨识性理论。该理论既有线性版本，也有非线性版本，其中“非线性”指一般的（非参数）非线性。潜变量模型与结构方程模型密切相关，前者的可辨识性理论可用于构建后者的可辨识性理论。本文聚焦于可辨识性理论，力求以

浅显易懂的方式解释基本结果。本文对估计方法和算法着墨甚少。

我们先在本节剩余部分阐述这些不同模型的研究动机。后续各节安排如下。第2节定义可识别性的概念。第3节讨论（线性）独立成分分析（独立成分分析，ICA）模型。第4节讨论线性结构方程模型（结构方程模型，SEM）。接下来转向非线性（非参数）模型：第5节讨论非线性独立成分分析的可识别性，第6节讨论非线性结构方程模型的可识别性。第7节进一步讨论可识别性的用途、未来研究课题以及算法。第8节总结全文。

1.1 线性表征学习与因子分析

早在几十年前，经典因子分析中就已出现潜变量可识别性问题，而经典因子分析构成了我们所有研究的基础。其基本模型如下。假设 $s_i, i = 1, \dots, n$ 是 n标准化且不相关的高斯潜随机变量。因此，$\mathbf{s}$ 的协方差为单位矩阵，这一性质称为“白化性”。假设$\mathbf{A}$ 是一个$m \times n$ 矩阵，并用$\mathbf{n}$ 表示一个 m维的不相关噪声变量向量。我们观测到 m维随机向量$\mathbf{x}$ ，它是一个含噪线性混合，表示为

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \tag{1}$$

目标是恢复向量 $\mathbf{s}$ 中的成分，或者至少恢复矩阵 $\mathbf{A}$。人们已经提出了许多方法（Harman，1967），但从根本上说，这类因子分析受到“因子旋转”不确定性的困扰；换句话说，对于高斯因子，因子分析模型不可识别。这是因为高斯潜向量的任何正交变换，由于其白化性，都会使数据具有完全相同的分布，却会使潜变量（即特征）的取值大不相同。这是一个根本性问题，后文还会多次回到这一点。

因此，参数矩阵$\mathbf{A}$ 和成分 s_i 无法被唯一恢复。这种不可辨识性是个问题，因为拟合此类模型的一个重要目标是发现数据的内在结构，或获得有用的表征。如果模型给出的表征并非唯一，甚至定义不明确，就无法发现数据的内在结构。因此，找到可辨识的模型至关重要。当 $m = n$，即不进行降维时，这个问题尤其突出，而这正是本文的重点。[1]

[1]如果我们只想找到正确的子空间，那么降维本身并不一定会受到这种缺乏唯一性的问题困扰。在降维的情形下，因子分析实际上与主成分分析（PCA）密切相关。不过，本文讨论的是不进行 not 降维的情形；这两者是截然不同的问题。

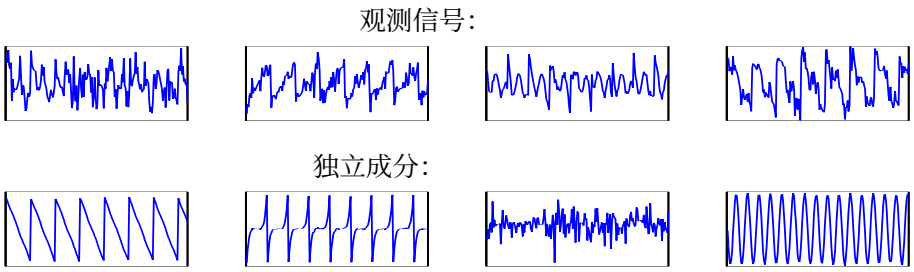


图1：独立成分分析的基本思想。根据上排显示的四个测量信号，独立成分分析能够恢复混合在测量结果中的原始源信号，如下排所示。

这里的一个实际动机是，我们可能希望从线性混合中分离信号，并以“盲”的方式完成，也就是说，只使用最少的信息。图 1 展示了这一点。四个观测信号（视为四维时间序列的各个条目）看起来毫无结构，或带有噪声。我们能否发现它们内在的隐藏结构?

幸运的是，可以通过独立成分分析（独立成分分析）解决这种不可辨识性问题。独立成分分析假设潜变量相互独立且服从非高斯分布。例如，对图1中的信号应用独立成分分析，我们便能将它们“分离”为四个原始源信号；这些信号在观测数据中已严重混合，如图所示。独立成分分析的这种应用通常称为“盲源分离”。如果存在一个尚未确定的因子旋转，就无法进行这样的分离。

下文第3节将回顾独立成分分析，它构成了本文所回顾的大多数理论的基础。

1.2 非线性表征学习与解缠结

近年来，深度学习深刻地改变了机器学习，其本质上意味着从数据中学习（估计）任意的非参数非线性函数。尽管在有监督情形下，相关理论已相当成熟，而这类情形本质上由回归构成，但无监督情形的理论发展则远不充分。所谓无监督，是指我们只能观测到一个可能具有高维度的随机向量 $\mathbf{x}$，而不存在任何定义好的“输出”“标签”或“回归目标”（就像上文的因子分析一样）。事实上，人们普遍认为，对表征或特征进行无监督非线性学习，是当前机器学习面临的最大挑战之一。这通常被称为“解缠结”，不过这个术语含义模糊，定义也不明确。

人们通常认为，无监督深度学习最令人满意的解决方案是基于概率生成模型的估计，因为概率理论往往能为学习提供最优目标，并能立即对各种量（如潜变量）进行概率推断。

这使机器学习理论与统计估计紧密联系起来。于是，可以将寻找良好表征定义为恢复被假定用于生成数据的原始潜变量。无监督深度学习中一些最流行的概率方法基于变分自编码器（VAE；Kingma 和 Welling（2014年））以及生成对抗网络（GAN；Goodfellow 等人（2014年））。这些方法在近似数据的概率密度函数（概率密度函数）并生成新的数据点方面取得了成功。

遗憾的是，在无监督深度学习中使用的大多数模型里，都没有证明能够恢复原始潜变量。事实上，这些模型往往依赖于对高斯、甚至白化的潜变量进行变换。因此，我们再次遇到了线性模型中已经提到的同样严重的不可辨识性问题（因子旋转）。

受线性独立成分分析理论的启发，我们或许可以尝试将其直接推广到非线性情形。这样一个基本框架假设，观测数据由对非高斯独立成分进行可逆非线性（非参数）变换而生成，而我们的目标是仅根据观测数据恢复这些成分。

然而，将独立成分分析推广到一般的非线性混合已被证明问题重重。特别是，如果观测数据 $\mathbf{x}$ 是以独立同分布样本的形式获得的，也就是说，数据中不存在时间结构或类似结构，那么该模型将严重不可辨识（Hyv¨arinen 和 Pajunen，1999年）。这是由于存在另一种非线性模型特有的不可辨识性。

幸运的是，通过利用时间序列中的时间结构（Harmeling 等人，2003年；Sprekeler 等人，2014年；Hyv¨arinen 和 Morioka，2016年、2017年），或类似的“辅助”信息（Hyv¨arinen 等人，2019年；Khemakhem 等人，2020a），可以找到解决非线性独立成分分析中不可识别问题的方法。第5节回顾了此类非线性潜变量模型，尤其是非线性独立成分分析的可辨识性理论。

1.3 因果发现与结构方程模型

因果发现是统计分析的另一个目标。它看起来可能与潜变量模型无关，但我们将在下文看到，两者实际上有着深刻的联系。因果模型在现代科学研究中发挥着基础性作用（Spirtes 等人，2000年；Pearl，2009年；Peters 等人，2017年）。虽然随机对照研究是研究一个变量对另一个变量的影响的金标准，但在许多情形下，这种方法并不可行或不合乎伦理（Spirtes 和 Zhang，2016年）。此外，互联网上公开的大数据集往往力求通用，因此无法有力地依据特定干预。正因如此，开发因果发现 方法，通过这些方法从（可能规模庞大的）被动观测数据中揭示因果结构，既有必要，也很重要。在未明确操纵某些变量的情况下收集的数据通常称为观察性数据，这与实验数据相对照；在实验数据中，某些变量会受到干预，例如随机对照试验。

在讨论一般情形之前，先考察线性双变量情形，对于理解基本思路非常有帮助。此时，因果发现问题本质上意味着通过在两个回归模型之间作出选择来确定“效应方向”；即

$$x_2 = bx_1 + e_2 \tag{2}$$

我们可以将其记为 $x_1 \to x_2$，或者

$$x_1 = bx_2 + e_1 \tag{3}$$

我们可以将其记为 $x_2 \to x_1$。众所周知，如果变量经过标准化且服从高斯分布，那么情况完全对称：两个模型的似然相等，两个模型的解释方差也相等（因此，如上所设，回归系数也相等）。由此我们再次看到，对于高斯变量而言，我们面临一个问题：效应方向不可辨识。不过，正如下文将要看到的，对于非高斯变量，效应方向是可辨识的。

在一般情况下，我们采用结构方程模型[2]（SEM）框架（Bollen，1989年）。从根本上说，结构方程模型定义了一种统计模型，用于描述一组观测变量$\mathbf{x} = (x_1, \ldots, x_n)$之间的相互作用，这些相互作用由一组相互独立的扰动项或噪声变量$\mathbf{e} = (e_1, \ldots, e_n)$来表示。然而，结构方程模型不仅是描述概率分布的统计模型，还是一种能够用于封装因果知识的数学工具（Pearl，2009年）。从这个意义上说，结构方程模型比潜变量模型更强大：它们不仅描述了所有分布的集合，还可以通过改变一个或多个方程（23）中的噪声分布或因果机制来实施干预并回答反事实查询。我们将形式化定义留待后文给出，此处仅通过图2示意结构方程模型。图2展示了如何用图来描述观测变量之间的影响，其中的弧表示哪些变量在因果上影响哪些变量。

只有当结构方程模型定义了一个可辨识的因果模型时，它们才有助于因果发现。对于因果模型，可识别性通常意味着我们能够区分因果关系，或像上面的例子那样找到效应方向。一般来说，这意味着我们能够确定变量（x_1，……，x_n）之间正确的“因果排序”；粗略地说，较早的变量是较后变量的原因。如果不具备可识别性，不同的因果顺序就会产生相同的数据，从而无法对因果关系进行真正的科学分析。

乍看之下，因果发现问题似乎与潜变量模型估计截然不同，但这里的深刻结果是，在某些情况下，我们可以将结构方程模型的估计简化为潜变量模型（例如独立成分分析）的估计。因此，如果我们开发出一种可辨识的潜变量模型，就可能开发出与之对应且同样可辨识的结构方程模型。事实上，甚至可以在很大程度上共用估计方法。下文第4节

[2]结构方程模型有时在近期的机器学习文献中被称为结构因果模型或函数因果模型。

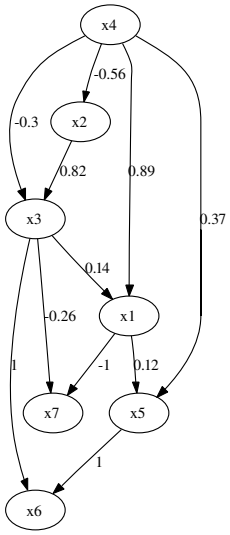


图2：结构方程模型可以用一个有向图（通常是无环的）表示，其中的弧表示因果影响以及统计依赖关系。这里，节点的排列方式使得所有影响都从上到下传递。此处未绘出扰动项或噪声变量，因为每个观测变量都有自己的扰动变量。

将讨论线性情形下的可识别性，而第6节将讨论非线性情形。

2 可识别性的定义

现在我们将给出可识别性的正式定义。可识别性是概率模型的一项重要性质。例如，如果模型不可辨识，那么无论是模型中估计出的参数，还是某些其他潜变量，似乎都没有解释的意义。关于可识别性用途的更详细讨论留待第7节；此处我们仅给出定义和一些示例。

考虑一个概率模型 $\mathcal{P}$，其定义为一组分布 $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$，这些分布的参数 $\boldsymbol{\theta}$ 取值于某个集合 Θ，定义在一组可能的观测值 $\mathcal{X}$（在我们的情形中通常为 $\mathbb{R}^n$）上。若模型 $\mathcal{P}$ 的映射 $\boldsymbol{\theta} \in \Theta \mapsto P_{\boldsymbol{\theta}}(\mathbf{x})$ 是单射（即一一对应），则称该模型为可辨识：

$$(P_{\boldsymbol{\theta}_1}(\mathbf{x}) = P_{\boldsymbol{\theta}_2}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}) \implies \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2. \tag{4}$$

换句话说，如果两个参数 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$ 在观测值集合 $\mathcal{X}$上生成相同的分布，那么它们必然相等。需要注意的是，可识别性是概率模型的一种性质，而不是任何特定估计方法的性质。

举一个简单的理论示例，假设我们抛了一枚可能有偏的硬币 N 次。每次抛硬币只有两种结果：正面，概率为$\theta \in [0,1]$；或反面，概率为 $1-\theta$。我们为这些数据定义两个不同的模型。第一种情况下，我们用参数为 θ 的伯努利模型来描述抛硬币的结果；用 $\text{Ber}_\theta(x)$ 表示给定 $x \in \{H,T\}$ 时抛硬币的概率，其中 $x \in \{H,T\}$ 对正面和反面进行编码。这个模型是可辨识的。为说明这一点，设 (θ_1, θ_2) 满足：对于 $x \in \{H,T\}$，$\text{Ber}_{\theta_1}(x) = \text{Ber}_{\theta_2}(x)$。由于 $\text{Ber}_\theta(H) = \theta$，且在无限次抛硬币的极限下，这个概率可直接由数据得出，因此我们得出 $\theta_1 = \theta_2$，该模型具有可识别性。另一方面，作为一个反例，我们可以设想一种类似潜变量模型的情形：我们观察到的不是一枚真实的硬币，而是计算机模拟的输出。在这种情况下，抛硬币分两步进行：首先抽取一个样本 $z \sim \mathcal{N}(\mu, \sigma^2)$，也就是一个服从均值为 μ 、方差为 σ^2的高斯分布的潜变量；然后，如果 $z \geq 0$，就将 x 赋为正面，否则赋为反面。这实际上是一个参数为 (μ, σ) 的潜变量模型。关键在于，简单计算表明，观测概率仅取决于比值 $\frac{\mu}{\sigma}$；无穷多组参数对 (μ, σ) 都会得到相同的比值，这意味着该模型不可识别。

作为一个与本文目的最相关的基础示例，考虑方程 (1) 中的高斯因子分析，其中成分或因子 s_i 服从高斯分布、彼此不相关且具有单位方差。众所周知，这个模型不可辨识。为简单起见，设矩阵$\mathbf{A}$ 为方阵且正交——用$\mathbf{U}$ 表示它，以突出这些假设——并忽略噪声$\mathbf{n}$。一种直观的不可辨识性解释是，高斯分布完全由协方差（和均值）决定。协方差的数量为 $\approx n^2/2$，这是由对称性导致的，因此我们无法求解混合矩阵中的 n^2 个参数，因为“变量多于方程”。更严谨地说，当协方差矩阵为单位矩阵时（这里的成分正是如此），高斯分布呈现旋转对称性。根据概率变换公式（将$\mathbf{x}$ 变换回$\mathbf{s}$），其概率密度函数为：

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp(\frac{1}{2}\|\mathbf{U}^T\mathbf{x}\|^2)|\det \mathbf{U}^T| = \frac{1}{(2\pi)^{n/2}} \exp(\frac{1}{2}\|\mathbf{x}\|^2) \tag{5}$$

最后一个等号来自$\mathbf{U}$的正交性。现在可以看出，$\mathbf{x}$ 的概率密度函数完全不依赖于$\mathbf{U}$ 。因此，$\mathbf{U}$ 不可辨识。实际上，我们可以用任意正交矩阵旋转因子$\mathbf{s}$ ，并通过用其逆矩阵旋转$\mathbf{U}$ 的列来抵消这一变化，而数据分布将保持不变；这就是所谓的“因子旋转不确定性”。

因此，可识别性的基本定义谈的是参数的可识别性。有时，我们也会稍微宽泛地称之为潜变量的可识别性，但其具体定义并不十分明确。如果考虑式（1）中的无噪声因子分析模型，已知矩阵$\mathbf{A}$ 后，便可利用$\mathbf{A}$ 的（伪）逆映射立即得到成分$\mathbf{s}$ （假设 $m \geq n$，这是常见的情况）。因此，从非严格的意义上说，$\mathbf{A}$ 的可识别性意味着$\mathbf{s}$的可识别性。另一方面，如果

模型中确实存在噪声，那么仅知道$\mathbf{A}$ 并不能得到成分$\mathbf{s}$ ，因为噪声无法被完全消除。在这种情况下，可以将可识别性定义为：后验分布 $p(\mathbf{s}|\mathbf{x},\mathbf{A})$ 能够以适当的方式恢复（Khemakhem 等人，2020a）。不过，这里我们暂不采用这种定义，而是聚焦于参数的可识别性。

还应注意，式（4）中可识别性的严格定义在某些情况下可能有所局限。在实际情境中，我们可能希望采用稍微宽松一些的定义来引入可识别性，例如，参数在一个等价类意义下可识别。比如，独立成分只能在成分任意缩放的意义下可识别，但这通常不被视为问题。

3 线性独立成分分析

独立成分分析是一种统计潜变量模型，与上文回顾的经典因子分析模型密切相关。其基本思想是，假设各成分为非高斯分布，就能打破前文所述的旋转对称性，从而得到可辨识的模型。尤其对于非高斯数据，高阶矩所提供的信息多于协方差中所包含的信息（Hyv¨arinen 和 Oja，2000年；Hyv¨arinen 等，2001年）。基本独立成分分析模型的可识别性自 Comon（1994）以来便已广为人知；事实上，Darmois 和 Skitovich 早在 20 世纪 50 年代就已证明这一点。

3.1 定义与可识别性

基本模型如下（Comon，1994；Jutten 和 H´erault，1991；Hyv¨arinen 和 Oja，2000；Hyv¨arinen 等，2001）。假设 $s_i, i = 1, \ldots, n$ 是 n 相互独立的非高斯潜在随机变量。假设 $\mathbf{A}$ 是一个可逆矩阵。我们观测到随机向量 $\mathbf{x}$ ，它是一个线性混合，表示为

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{6}$$

我们假设 s_i 的均值已中心化为零，且其方差有限并归一化为 1。请注意与因子分析模型的区别：

1. 假设各成分是非高斯的。这是使该模型区别于经典因子分析的根本差异。（虽然“ICA”中的首字母强调独立性，但经典高斯因子分析中的因子也相互独立；就这一差异而言，这一术语略有误导性。）

2. 成分（因子）的数量等于观测变量的数量。为简化数学处理，基本上采用了这一假设：此时，矩阵 $\mathbf{A}$ 可唯一求逆，但这一假设也可以放宽。同时，这强调了这样一个事实：在

独立成分分析中，我们关注的并非降维，而是找到原始成分。实践中，在应用独立成分分析之前，通常会先通过主成分分析（PCA）降低维数。按照因子分析的经典术语，独立成分分析可以被视为一种“因子旋转”，其中主成分被看作因子的估计值。

3. 不存在噪声项。这在一定程度上是由成分数量众多这一点所合理解释的：其中一些成分可能是噪声，另一些则可能是更有意义的成分。由于我们并未降低数据的维数，因此数据的所有方差无论如何都会由这些成分“解释”。

真正从根本上将独立成分分析模型与经典因子分析区分开来，并使可识别性成为可能的，是非高斯性。独立成分分析估计理论进一步凸显了非高斯性的重要性；该理论还直观地证明了可识别性，过程如下。独立成分分析中最基本的定理之一指出，我们可以通过寻找一个使成分非高斯性最大化的可逆变换来估计独立成分分析。事实上，每个成分都对应于某个线性组合$\sum_i w_i x_i$的非高斯性在某些参数 w_i上的最大值。这是因为，粗略地说，根据中心极限定理，若干独立随机变量之和比其中任一原始随机变量更接近高斯分布（严格来说，只有当这些随机变量具有相同分布时，这一点才成立）。线性组合 $\sum_i w_i x_i$ 是 s_i的线性组合，因此当它实际上等于其中一个 s_i时，其非高斯性达到最大值。详情参见 Hyv¨arinen 和 Oja（2000年）；Hyv¨arinen 等人（2001年）。

注意到独立成分分析模型存在两个相当显而易见的不确定性，即模型中不可识别的方面。首先，成分的顺序不可识别，甚至模型也没有定义成分的顺序。其次，每个成分只能估计到线性缩放（以及符号）的程度，因为若将某个成分乘以一个标量常数，同时将 $\mathbf{A}$ 中对应的列除以该常数，数据分布就会保持不变。当按照惯例将各成分的方差定义为 1 时，这种不确定性会部分消除；这也意味着 $\mathbf{s}$ 是白化的。不过，这只是一种约定，成分的真实方差仍不可识别。

3.2 一个简单的可识别性证明

接下来，我们为独立成分分析给出一个简单而严谨的可识别性证明。为求简明，证明的代价是其一般性较著名的 Darmois-Skitovich 理论有所降低。在这个简单情形下，可识别性定理可表述如下：

定理 1 假设独立成分具有有限方差，且独立成分的对数概率密度函数具有连续二阶导数。如果式 *(6)* 中的变量 $x_i, i = 1, \ldots, n$ 相互独立，则$\mathbf{A}$ 的每一行和每一列中恰好各有一个非零元素。

证明：[3] 与 ICA理论中的常见做法一样，我们可以不失一般性地假设$\mathbf{A}$ 是正交的。这是因为$\mathbf{s}$ 是白化的，而且我们还可以将$\mathbf{x}$ 白化作为预处理；由于$\mathbf{x}$ 的协方差等于$\mathbf{A}\mathbf{A}^T$，因此$\mathbf{A}$ 是正交的。

$\mathbf{x}$ 的概率密度函数由概率密度变换公式得到：

$$p(\mathbf{x}) = \prod_{i=1}^{n} p_i(\sum_{j=1}^{n} a_{ji}x_j)|\det \mathbf{A}^T| \tag{7}$$

其中， p_i 表示第 i个成分的概率密度函数。取对数，并注意到正交矩阵的行列式等于 ± 1，我们得到

$$\log p(\mathbf{x}) = \sum_{i=1}^{n} \log p_i(\sum_{j=1}^{n} a_{ji}x_j) \tag{8}$$

接下来，由于假设 x_i 相互独立，$\log p(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i)$，其中 f_i是一些函数。这意味着任意二阶交叉偏导数都必须为零：

$$\frac{\partial^2 \log p(\mathbf{x})}{\partial x_k \partial x_l} = \sum_{i=1}^{n} a_{ki}a_{li}(\log p_i)''(\sum_{j=1}^{n} a_{ji}x_j) = 0, \quad \text{for all } k \neq l \tag{9}$$

可以将这组方程汇总并写成矩阵形式：

$$\mathbf{A}^T \mathrm{diag}_i[(\log p_i)''(\sum_{j=1}^{n} a_{ji}x_j)]\mathbf{A} = \mathrm{diag}_i[c_i(\mathbf{x};\mathbf{A})] \tag{10}$$

其中， $c_i(\mathbf{x};\mathbf{A})$是关于$\mathbf{x}$ 和$\mathbf{A}$ 的一些未知标量值函数，对应于情形 $k = l$。该方程必须对所有$\mathbf{x}$成立。

仅对于高斯密度，$\log p_i$ 是二次函数，其二阶导数为常数。我们假设其具有非高斯性，因此 $(\log p_i)''$ 不是常数。根据连续性，$(\log p_i)''$ 实际上会取无穷多个不同的值，因为它会取实数轴上某个区间内的所有值。这意味着，对于每个 i ，我们都能找到一个点 y_i ，使得各项 $d_i = (\log p_i)''(y_i)$ 两两不同，即 $d_j \neq d_k$ 对于 $j \neq k$。（即使其中一个成分是高斯分布，这也是可能的。）由于 $\mathbf{A}$ 可逆，我们可以找到相应的 $\mathbf{x}$ ，使得$y_i = \sum_{j=1}^{n} a_{ji}x_j$ 对所有 i 均成立。下文中，我们考虑 (10)，并将 $\mathbf{x}$ 固定为该值，因此对角元素 d_i 和 c_i 也是固定的。

[3]或者，我们也可以在傅里叶域中进行同样的证明，即使用特征函数$\hat{}p_i$。此时， $p(\mathbf{x})$和 $p_i(s_i)$将由特征函数取代，并且雅可比矩阵在第一个方程中消失。因此，我们可以用假设p_i的特征函数（二阶导数连续，记为$\hat{}p_i$）来替代光滑概率密度函数的假设。这种假设与成分的矩结构有关：它只比假设成分具有有限方差略微严格一些。整个证明只需稍作修改，即可适用于特征函数。因此，我们得到一个更一般、但稍微复杂一些的证明。

现在，(10) 左端的方程实际上是一个特征值分解（EVD），因为 **A** 是正交的。重要的是，对角元素（特征值）互不相同；根据线性代数中的一个著名结果，这意味着该特征值分解除特征值的排列顺序和特征向量的符号外是唯一的。右端也可以解释为一个特征值分解，其特征向量的每一行和每一列都恰有一个非零元素。由于两边特征值分解的“特征向量”必须在排列意义下相匹配，**A** 的特征向量的非零元素也必须具有相同的性质。 □

请注意，如果这些成分服从高斯分布，那么式（10）左侧的对角矩阵将等于单位矩阵，因此无论**A**取何值，整个等式左侧都等于单位矩阵。此时，任何正交**A** 都满足式（10），因为很容易证明右侧也必然等于单位矩阵。这说明高斯变量是不允许的。事实上，特征值分解的唯一性仅在特征值互不相同时成立。

还需注意，矩阵**A** 并未被证明是单位矩阵。**A** 的形式允许对成分进行排列，也允许将每个成分乘以一个标量常数。正如上文所指出的，这些是根本性的不确定性，在任何类似因子分析的模型中都无法消除。

3.3 替代方法

盲源分离领域的文献实际上将非高斯性视为能够估计线性混合模型的诸多原理之一，或许也是其中最重要的一项。对于时间序列数据，可识别性可由两种截然不同的统计结构实现：成分的自相关（Tong 等，1991；Belouchrani 等，1997），或其非平稳性（Matsuoka 等，1995；Pham 和 Cardoso，2001）。Cardoso（2001）将这两者与非高斯性合称为“通往 [线性] 独立成分分析的三条捷径”。下文将进一步探讨这些原理在非线性独立成分分析中的应用。

当数据为非负值，尤其是数据集中在较小的正值附近时，还可以采用另一种方法。这便是最初称为正矩阵分解（Paatero 和 Tapper，1994年）的模型，Lee 和 Seung（1999年）则以非负矩阵分解（NMF）之名使其广为流行。Donoho 和 Stodden（2004）分析了该模型的可识别性。Plumbley（2003年）和 Hoyer（2004）探讨了非负矩阵分解与独立成分分析的结合。与此相关，Hyttinen 等（2022年）研究了观测值为二元值的情形。

我们还要指出，独立成分分析模型与字典学习和稀疏编码非常相似。这种相似性在这些方法早期的研究中已广为人知（Olshausen 和 Field，1997年），但近来似乎已被遗忘。独立成分分析与字典学习的概率形式之间，唯一真正的区别在于：字典学习中的成分数量较多（$n > m$），并且存在噪声。Eriksson 和 Koivunen（2004年）探讨了这类情形的一些可识别性结果。

4 线性结构方程模型

接下来，我们考虑线性结构方程模型，并将其作为因果发现的基本框架。我们将看到，如何基于线性独立成分分析理论证明其可识别性。

4.1 模型定义

线性结构方程模型由一组如下形式的方程构成：

$$x_i = \sum_{j=1}^{n} b_{ij} x_j + e_i, \quad i = 1, \ldots, n, \tag{11}$$

其中， x_i 为观测变量， e_i 为称作扰动项、外部影响或简称噪声变量的潜变量。其基本思想是，变量 x_i 由那些系数 b_{ij} 非零的 x_j 所导致。

结构方程模型通常与一个有向无环图（DAG） $\mathcal{G}$ 相关联，该图称为因果图。 $\mathcal{G}$ 中的每个节点都对应一个观测变量 x_i，当且仅当 b_{ij} 非零时，才存在一条从 x_j 指向 x_i 的边。这里我们实际上采用了这一语境中非常典型的无环性假设。这意味着，当矩阵 $\mathbf{B}$ 被解释为一个（加权）图的连接矩阵时，由此得到的图没有环：无法从某个节点出发，沿着边前进后又回到起始节点。有向无环图的一个有趣性质是，可以对节点（变量）排序，使所有连接都按该排序“向前”延伸；这称为“因果顺序”，但它不一定唯一。图2展示了一个表示结构方程模型的有向无环图示例，并按照因果排序进行了直观排列。

现在的问题是根据对$\mathbf{x}$的观测值来估计参数 b_{ij} 。众所周知，这个问题通常是病态的。具体而言，对于高斯数据，模型是不可识别的，正如上文针对两个变量的情形所示。一种广泛采用的解决方法是利用先验知识，将大多数 b_{ij} 设为零；例如，可以通过随机对照试验获得这类知识。然而，使用此类先验知识，更不用说进行干预，都与我们关注的发现这一目标大相径庭。

4.2 可识别性

结构方程模型可识别性的关键在于它与独立成分分析的密切关系。潜变量模型与结构方程模型之间的这种紧密联系是本文的核心主题；我们稍后将看到，它甚至适用于非线性情形。

用$\mathbf{B}$ 表示一个汇集所有系数 b_{ij}的矩阵。我们可以将结构方程模型表示为

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \tag{12}$$

其中，向量$\mathbf{e}$ 收集了外部影响。现在，根据基础线性代数，这意味着

$$\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e} \tag{13}$$

换句话说，结构方程模型意味着数据遵循潜变量模型。事实上，这不过是一个混合矩阵为$\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$的独立成分分析模型，但需要满足适当的假设。具体而言，假设 e_i 相互独立且非高斯。那么，方程（13）中的模型恰好是一个独立成分分析模型，并且根据定义，其成分数量与观测变量数量相同（成分对应于扰动项）。

因此，似乎我们可以估计出关于$\mathbf{x}$的独立成分分析模型，并将得到的$\mathbf{A}$ 通过简单地$\mathbf{B} = \mathbf{I} - \mathbf{A}^{-1}$变换回$\mathbf{B}$ 。然而，这里有一个严重的问题：独立成分分析无法估计各成分的顺序， e_i。正如第3节所指出的，成分之间并不存在固有的顺序。这与结构方程模型形成鲜明对比；在结构方程模型中，我们知道 e_i 是变量 x_i 的外部影响（对于相同的索引 i）。因此，在变换回$\mathbf{B}$之前，我们需要找到独立成分的正确顺序，也就是混合矩阵$\mathbf{A}$各列的顺序。

一般而言，无法恢复正确的顺序。然而，如上所述，在结构方程模型和因果发现理论中，通常会假设图是有向无环图（DAG）。基于无环性假设，可以找到各成分的正确顺序。为说明如何做到这一点，考虑二维情形。假设我们已估计出混合矩阵的逆$\mathbf{W} = \mathbf{A}^{-1}$，其中成分的顺序是任意的。根据公式（13），我们看到应有$\mathbf{W} = \mathbf{I} - \mathbf{B}$。根据定义，$\mathbf{B}$ 的无环性在这个二维情形下意味着它恰好有一个非零元素，且该元素位于非对角线上（除非图是退化的，即$\mathbf{B}$ 的元素全为零）。因此，$\mathbf{W}$ 恰好有一个零元素。将任意一个非零元素记为 $*$，实际的$\mathbf{W}$可以是如下形式，例如

$$\mathbf{W} = \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \tag{14}$$

现在，如果由于成分顺序的不确定性，在估计时将$\mathbf{W}$ 的行顺序弄错，很容易看出零元素会落到对角线上。这与定义相矛盾，因为按定义，对角元素全为 1（即使允许进行任意缩放也是如此）。因此，在两种不同的排序中，可以依据无环性找出正确的排序。一般情形由 Shimizu 等人（2006）作了解释；请注意，无环性是充分条件，但不是必要条件（Lacerda 等人，2008）。

还有一个较小的细节：独立成分分析和结构方程模型中混合矩阵的归一化方式不同：在独立成分分析中，通常将各成分的方差定义为 1；而在结构方程模型中，则通过$\mathbf{I} - \mathbf{B}$ 的对角元素全为 1，实现相关的归一化。不过，这只是一种归一化约定，对可识别性的影响很小。

因此，我们看到，在独立性和非高斯性（如独立成分分析中的情形）以及新增的无环性假设下，结构方程模型是可辨识的。由此得到的模型称为 LiNGAM，即线性非高斯无环模型。关于这一基本模型的详细内容，请参阅 Shimizu 等人（2006）；关于包含一些较新进展的更深入论述，请参阅 Shimizu（2014）。

关于 LiNGAM 的估计，同样可以对所得结构方程模型估计作出非常明确的解释，即将其理解为非高斯性最大化（Hyv¨arinen 和 Smith，2013），这与独立成分分析估计中的情形类似。

4.3 替代方法

例如，Peters 和 B¨uhlmann（2014年；Jakobsen 等人，2022年）提出了其他可辨识的结构方程模型框架。这些框架不再假设非高斯性，而是引入了更受因果推断文献启发的其他限制。因果发现中一个尤为重要的问题是，我们可能无法观测到所有相关变量；例如，Hoyer 等人（2008年）和 Tashiro 等人（2014年）研究了存在这类隐藏“混杂因素”时的因果发现。Monti 和 Hyv¨arinen（2018）提出了结合源分离与 LiNGAM 的方法。最后，我们指出，上文第3.3小节回顾的线性混合模型替代估计方法，也可用于估计线性结构方程模型。事实上，Zhang 和 Hyv¨arinen（2010）运用一种相关思路，提出了另一种结合源分离与因果发现的方法。

5 非线性独立成分分析

非线性独立成分分析是无监督学习中的一个基础问题，近年来受到广泛关注。它为表征学习和“解缠结”提供了一种有原则的方法，尤其是利用深度神经网络时。非线性独立成分分析试图通过推广第3节中的线性独立成分分析框架，从多维数据中找出非线性成分。与大多数无监督表征学习方法的本质区别在于，这种方法首先定义一个生成模型，使原始潜变量能够被恢复，即模型在设计上便是可辨识的。

5.1 可识别性问题

我们先介绍一个简单的非线性独立成分分析模型，结果发现该模型不可识别。与上文一样，用$\mathbf{x} = (x_1, \ldots, x_n)$表示一个观测到的 n维随机向量。我们假设它由 n 个相互独立的潜变量生成，这些变量称为独立成分， s_i。对非线性独立成分分析问题，一个直接的定义是：假设观测数据是对潜变量$\mathbf{s} = (s_1, \ldots, s_n)$进行任意（但平滑且可逆的）变换$\mathbf{f}$后得到的，如下所示：

$$\mathbf{x} = \mathbf{f}(\mathbf{s}) \tag{15}$$

因此，目标是仅根据对$\mathbf{x}$ 的观测值，恢复逆映射$\mathbf{f}^{-1}$ 以及独立成分 s_i 。

非线性独立成分分析的研究一直受到阻碍，原因在于，与线性独立成分分析的情形形成鲜明对比的是，这类简单的非线性独立成分分析方法并不可辨识。简单来说，对于任意 x_1，x_2，总能找到一个函数$g(x_1, x_2)$，使其与 x_1相互独立。Darmois早在1952年就给出了这样的构造，从而展示了非线性独立成分分析的“不可能性”。他仅使用了条件累积分布函数

$$g(\xi_1, \xi_2) = P(x_2 < \xi_2 | x_1 = \xi_1) \tag{16}$$

以构造一个新的变量 $z = g(x_1, x_2)$，结果它与x_1独立。（Hyv¨arinen 和 Pajunen（1999年）给出了一个略微推广的版本，而 Locatello 等人（2019年）则给出了一个略有不同的版本。）问题在于，先进行初始变换 $\tilde{x}_1 = h(x_1, x_2)$ 后，我们同样可以应用这一方法。因此我们可以看到，任何 此类变换 $\tilde{x}_1$ 都可以被视为一个独立成分，因为我们总能找到一种数据分解方式，即 $\tilde{x}_1$ 以及通过上述构造定义的变量。此外，我们还可以轻易地将这两个变量的边缘分布变换为任意形式；因此，特别是，我们可以精确复现原始独立成分的分布 ，而无需恢复这些成分 本身。（这一推理与上文给出的可识别性定义略有不同，因为这里考虑的是潜变量的分布，但二者是等价的。）由此可见，独立性这一假设不足以确保可识别性，无论我们对独立成分的分布作出何种假设（例如非高斯性） s_i。

不可辨识性在实践中当然是一个重大问题，因为独立成分分析的大部分效用都建立在模型可辨识这一事实上；换一种说法，就是“源信号可以分离”。我们还应注意到，颇为奇怪的是，许多无监督深度学习模型，例如变分自编码器和生成对抗网络，都假设潜变量模型中的 s_i 甚至服从高斯分布，因此重现了经典的高斯因子旋转问题，使不可辨识性变得更加严重。

5.2 可辨识模型的定义

不过，需要指出的是，上述 Darmois 反例假设数据是从一个随机向量中独立同分布采样得到的；具体而言，观测值（数据点）彼此统计独立（这与各分量之间的独立性不可混淆）。一个值得探索的方向是放宽独立同分布采样这一假设。一个基本情形是考察时间序列以及独立成分的时间结构 。因此，我们假设

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) \tag{17}$$

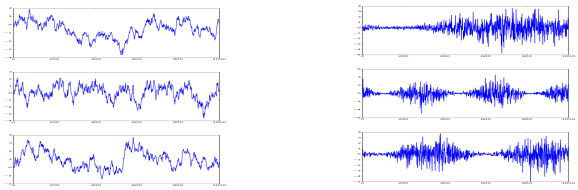
其中 t 是时间索引。作为初步尝试，我们可以假设源$s_i(t)$具有非零自相关，这在线性情形下由来已久（Tong 等，1991；Belouchrani 等，1997）。在机器学习文献中，

此类模型常以特征的“时间相干性”、“时间稳定性”或“慢特征”为名得到广泛应用。Harmeling 等人（2003）提出，我们可以尝试寻找非线性变换，使其即使跨越不同时间滞后，并经过非线性标量变换后，仍具有最大程度的不相关性。Sprekeler 等人（2014）表明，如果所有源的自相关函数各不相同，一种密切相关的方法便能够实现源分离。因此，Sprekeler 等人（2014）的研究可能是首个针对具有一般非线性关系的非线性独立成分分析所作的可识别性证明；然而，该方法受限于一个条件：源必须具有不同的统计性质，而这在许多情况下并不现实。需要注意的是，正如线性独立成分分析中的线性缩放不可识别一样，非线性独立成分分析中的单调函数缩放在根本上也是不可识别的；在这两种情况下，每个分量的整体符号及其排列顺序同样不可识别。

Hyv¨arinen 和 Morioka（2017年）针对作为平稳时间序列的成分提出了严谨且一般性的可识别性处理方法：各成分彼此独立，但每个成分都具有时间依赖关系。假设成分 $s_i(t)$ 根据某些技术定义具有充分的时间依赖关系和充分的非高斯性，该模型已被证明是可辨识的，除成分的单调逐点变换之外不存在其他不确定性。H¨alv¨a 等人（2021年）对证明进行了改进，部分基于 Schell 和 Oberhauser（2023年）的工作；Klindt 等人（2020年）提出了替代方法。

直观地说，时间依赖关系能够提供更多信息，因为此时成分在任意滞后下都相互独立，即 $s_i(t)$ 和 $s_j(t-\tau)$ 对任意滞后 τ 都相互独立。随后，这种独立性可以被施加到成分的任意估计上。与上文 Schell 和 Oberhauser（2023年）所述 Darmois 构造中考虑的独立同分布情形相比，这会带来更多约束。因此，从直观上看，该模型变得可辨识是合理的。

在线性盲源分离中，先前使用过的另一种时间结构是非平稳性（ Matsuoka 等，1995；Pham 和 Cardoso，2001），尤其是方差的非平稳性。方差的这种非平稳性似乎广泛存在于许多类型的真实数据中，例如脑电图和脑磁图（Brookes 等，2011）、自然视频（Hyv¨arinen 等，2009），并且与金融时间序列中波动率的变化密切相关。Hyv¨arinen 和 Morioka（2016）将这一原理扩展到了非线性情形；Khemakhem 等人（2020a）迄今基于这一原理给出了最强的可识别性结果。他们假设一个分段非平稳模型，其中每个分段中的成分都服从指数族，而指数族的参数则随分段而变化。此外，他们还假设分段方式已知。在保证非平稳性足够强的技术性假设下，他们证明了该模型的可识别性。尽管这里的假设看起来相当严格，但如下文所示，这一结果可以通过多种方式扩展。请注意，在这一理论中，通常只能恢复成分的某种非线性逐点函数，但该函数不一定是单调的；例如，在最简单的模型中，唯一能够恢复的是



自相关　　非平稳性

图3：非线性独立成分分析中使用的两种基本时间结构。左：自相关源，呈现时间依赖关系的基本形式。右：非平稳源，尤其表现为方差非平稳。如果独立成分具有其中任一种结构，模型便可实现可辨识。此外，这些结构还可以通过多种方式组合或推广。

成分的平方。

同样，我们可以用简单直观的方式解释这些可识别性结果。根据独立性这一基本假设，在任意时刻，各成分彼此独立，即对于任意 t， $s_i(t)$ 和 $s_j(t)$ 必须相互独立。关键在于，非平稳性意味着各片段内的分布不同；因此，独立性约束的数量基本上会乘以片段数。这样，与上文 Darmois 构造中讨论的独立同分布情形相比，我们得到了更多约束。

图3展示了此处使用的两种时间结构的基本形式。这种时间结构的二分法有助于直观理解时间依赖关系。不过，如果假设表面上的非平稳性源自隐马尔可夫模型，那么非平稳性也可以归入时间依赖关系这一类（H¨alv¨a 和 Hyv¨arinen，2020年）。这引出了 H¨alv¨a 等人（2021年）提出的一般理论，该理论表明，对于非常一般的时间依赖关系（包括源自隐马尔可夫模型的时间依赖关系），模型均具有可识别性。Morioka等人（2021年）提出了一种结合这两种时间结构的相关模型。

推广可识别性的另一种方式是认识到，基于非平稳性的证明仅假设在给定片段索引的条件下，成分的分布会发生变化，而不对片段内部的时间结构作任何限制。因此，我们可以用任何能够对成分分布进行条件化的观测到的辅助变量 来替代这个片段索引。它可以是生物医学场景中的受试者索引、国际研究中的国家索引，也可以是任何能够以适当方式改变底层统计特性的指标。由此，我们基于观测到这类额外辅助信息，得到另一组可识别性结果（Hyv¨arinen 等人，2019年）。Khemakhem等人（2020a,b）进一步推广了这些结果；另见Zimmermann 等人（2021年）。下文将看到，这在

因果发现中尤其有用。

结果的另一种推广是考虑带噪混合模型，即像式 1 中的经典因子分析那样，在混合过程中加入噪声。这种加性噪声具有一些算法上的优势，事实上，相关的无监督模型以及 Khemakhem 等人（2020a）也广泛采用了这种方法。虽然通常假定其分布已知，但 H¨alv¨a 等人（2021年）证明，在相当一般的条件下，噪声分布是可识别的。

5.3 可识别性证明概要

接下来，我们将按照 Hyv¨arinen 和 Morioka（2016年）的方法，概述一个简单的可识别性证明。这只是当前已有理论的一个特例，但有助于说明基本原理。关于基于 Hyv¨arinen 和 Morioka（2016年）的最复杂的现有证明，请参阅 Khemakhem 等人（2020a,b）；因此，他们的证明是对下文证明的推广。还需指出，H¨alv¨a 等人（2021年）基于 Hyv¨arinen 和 Morioka（2017年）采用一种截然不同的方法给出了证明，这种方法让人联想到上文给出的线性独立成分分析可识别性证明。Hyv¨arinen 等人（2019年）还提出了另一种截然不同的证明方法。

我们考虑一个非平稳模型，在每个片段 $\tau = 0, \dots T$ 中，第 i 个分量服从一阶指数族：

$$\log p_\tau(s_i) = b_i(s_i) + \lambda_{\tau,i} q_i(s_i) \tag{18}$$

其中，q_i, b_i 是分量 i 的充分统计量和基测度，而 $\lambda_{\tau,i}$ 是分量 i 在片段 τ 中的参数。（证明也可以针对时间索引 t 而非片段 τ 来展开，无需做任何修改。）现在，计算非线性独立成分分析模型下片段 τ 中一个数据点 $\mathbf{x}$ 的对数概率密度函数。根据概率变换公式，其对数概率密度函数为

$$\log p_\tau(\mathbf{x}) = \sum_{i=1}^{n} b_i(g_i(\mathbf{x})) + \lambda_{\tau,i} q_i(g_i(\mathbf{x})) + \log |\det \mathbf{J}\mathbf{g}(\mathbf{x})| \tag{19}$$

为简洁起见，我们省略了 $\mathbf{x}$ 中的索引 t；$\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}) \dots, g_n(\mathbf{x}))^T$ 是（真实）混合函数 $\mathbf{f}$ 的逆函数；因此，根据定义，$s_i = g_i(\mathbf{x})$。$\mathbf{J}$ 表示雅可比矩阵。

假设现在有两个不同的模型，它们给出了相同的数据分布，因此违反了可识别性。也就是说，还存在另一个逆函数 $\tilde{\mathbf{g}}$、充分统计量和基测度 $\tilde{q}$，b 以及参数 λ，使得对于所有τ，

$$\begin{aligned}
&\sum_{i=1}^{n} b_i(g_i(\mathbf{x})) + \lambda_{\tau,i} q_i(g_i(\mathbf{x})) + \log |\det \mathbf{J}\mathbf{g}(\mathbf{x})| \\
&\qquad = \sum_{i=1}^{n} \tilde{b}_i(g_i(\mathbf{x})) + \tilde{\lambda}_{\tau,i} \tilde{q}_i(\tilde{g}_i(\mathbf{x})) + \log |\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x})|
\end{aligned} \tag{20}$$

现在，将这个方程两边对应于 $\tau = 0$ 得到的项相减。我们得到

$$\sum_{i=1}^{n} \alpha_{\tau,i} q_i(g_i(\mathbf{x})) = \sum_{i=1}^{n} \tilde{\alpha}_{\tau,i} \tilde{q}_i(\tilde{g}_i(\mathbf{x})) \tag{21}$$

其中 $\alpha_{\tau,\ i} = \lambda_{\tau,\ i} - \lambda_{0,\ i}$。值得注意的是，我们消除了雅可比项。

现在，只需汇总所有 $\tau = 1, \ldots\ldots, T$ 对应的公式（21）。假设 α 的矩阵是满秩的，这意味着我们有足够多的片段（其数量大于数据的维数），且这些片段的分布具有足够的变化。这样，我们就可以求出 $\tilde{q}_i(\tilde{g}_i(\mathbf{x}))$，它们可表示为 $q_i(g_i(\mathbf{x}))$ 的线性变换。假设各成分边缘独立（除在每个片段中条件独立之外），则可通过线性独立成分分析消除这种线性不确定性。因此，我们看到，对于某个指标排列$i,\ j$，有 $\tilde{q}_i(\tilde{g}_i(\mathbf{x})) = q_j(g_j(\mathbf{x})) = q_j(s_j)$，从而可以恢复各成分，但只能确定到由充分统计量给出的逐点非线性变换。（Khemakhem 等人（2020a）详细刻画了这种逐点变换；实际上，他们并不需要边缘独立性条件。）□

5.4 通过约束非线性实现的替代方法

使非线性独立成分分析可辨识的一种替代方法是约束混合函数。一种自然的做法是，假设混合函数接近线性（Zhang 和 Chan，2008年）。近期研究致力于对混合函数的雅可比矩阵 $\mathbf{Jf}$ 施加约束，以实现类似的效果。

经典的保角映射刘维尔定理与此密切相关。该定理考虑雅可比矩阵近似正交的情形，即

$$\mathbf{Jf}(\mathbf{x})^T \mathbf{Jf}(\mathbf{x}) = \alpha(\mathbf{x})\mathbf{I} \tag{22}$$

对于标量值函数 α。该定理指出，任何足够光滑的此类函数$\mathbf{f}$ 都必须属于一个非常受限的函数类（称为 M¨obius 变换）。该定理的一个简单推论（见附录 A 中的陈述和证明）指出，如果 $\alpha \equiv 1$，即雅可比矩阵是正交的，那么函数$\mathbf{f}$ 实际上必然是仿射函数：$\mathbf{f(x)} = \mathbf{Ux} + \mathbf{b}$，其中$\mathbf{U}$是某个正交矩阵。因此，这一理论有力地说明了对雅可比矩阵施加何种约束才有意义。将雅可比矩阵约束为正交并无意义，因为这不允许存在任何非线性混合函数。因此，一个引人深思的问题是，探究对式（22）中的约束作出某些放宽。希望这类放宽能够容许足够广泛的非线性函数，同时仍保证可识别性（Gresele 等人，2021年；Zimmermann 等人，2021年；Buchholz 等人，2022年）。

近期提出了许多进一步限制混合的研究方法。Kivva 等人（2022年）证明了在混合函数为分段仿射函数时的可识别性。Moran 等人（2021年）则限制了受单个独立成分影响的

观测变量的数量，从而使混合函数具有一种特殊的“稀疏性”。在一个略有不同的背景下，Donoho 和 Grimes（2003年）在降维情形下考虑了雅可比矩阵正交性的约束（但其中的 f 不可逆），Horan 等人（2021年）则将该约束应用于结合降维的非线性独立成分分析。此外，Willetts 和 Paige（2021年）通过解决一个次要任务，自动从观测值中学习辅助变量，即使没有时间结构也是如此；而 Gresele 等人（2020b）使用数据的另一种“视图”，而非辅助变量。

6 非线性结构方程模型

正如线性独立成分分析理论有助于解决第4节中的线性结构方程模型问题，非线性独立成分分析理论也有助于解决非线性结构方程模型问题。接下来，我们将重点讨论双变量因果发现问题。这对应于利用两个变量的观测值恢复因果结构，我们将这两个变量记为 x_1 和 x_2。双变量因果发现是更一般的因果发现问题的一个简化特例，但仍然是一项具有挑战性的任务。

6.1 问题定义

我们首先定义任意维数的完全非线性结构方程模型：

$$x_j = f_j(\mathbf{PA}_j, e_j), \quad j = 1, \ldots, n, \tag{23}$$

对于任意非线性函数 f_i。变量 $\mathbf{PA}_j \subseteq \{x_1, \ldots, x_n\} \setminus x_j$ 是关联图中变量 x_j 的父节点。这里的关键在于，“父节点”会导致变量 x_j；一般而言，因果发现问题可以表述为确定哪些变量分别是哪些变量的父节点。

为了实现可识别性，因果发现算法通常采用两种技术之一。第一种方法是对定义结构方程模型的函数 f_j 施加约束（23）。在极端情况下，我们会得到线性模型，也可能包含非高斯外部影响，如上文所述。与独立成分分析的情形不同，研究者也提出了一些相当有意义的受限非线性结构方程模型，下面将对此进行回顾。尽管如此，最有意思的情况或许是非线性不受限制，同时对外部影响施加一些统计假设；这正是本文的主要议题。

6.2 可辨识的弱非线性因果模型

在本小节中，我们将回顾一些最值得注意的、基于对非线性施加强限制的可辨识非线性因果模型。

加性噪声模型（ANM）。 Hoyer 等人（2009年）提出了加性噪声模型，其中结构方程模型的形式为

$$x_j = f_j(\mathbf{PA}_j) + e_j,$$

并且噪声变量$\mathbf{e}$ 相互独立，且独立于$\mathbf{x}$。他们的理论可识别性结果关注两个变量的情形x_1 和 x_2。该结果指出，若 x_1 导致 x_2，记作 $x_1 \to x_2$，那么对于某个函数 g 和独立于 x_2的噪声 $\tilde{e}$ ，我们无法写出 $x_1 = g(x_2) + \tilde{e}$ 。本质上，这个结构方程模型对于 x_1 和x_2 是不对称的，因此只能描述自然界中的因果关系。换言之，它是可识别的。Peters 等人（2014年）将这一可识别性结果推广到了两个以上变量的情形。

后非线性模型（PNL）。Zhang 和 Hyv¨arinen（2009年）提出了后非线性模型，该模型通过添加一个后续可逆映射 g_j对 ANM 进行了推广：

$$x_j = g_j(f_j(\mathbf{PA}_j) + e_j).$$

仍假设噪声变量$\mathbf{e}$ 相互独立，且独立于各个原因。作者证明，双变量后非线性模型在大多数情况下是可识别的，并列举了模型不可识别的五种特殊情况。这一可辨识性理论推广了 ANM 的理论；当 g_j 为恒等映射时，PNL 模型是 ANM 的一个特殊情形。请注意，如果我们知道g_j，就可以通过对效应应用映射 g_j的逆映射，将 PNL 模型化为 ANM；变换后的变量 g_j^{-1} (x_j)是原始效应 x_j的确定性函数。

因果自回归流（CAREFL）。Khemakhem 等人（2021年）指出，结构方程模型与机器学习中的一类模型密切相关，这类模型称为（仿射自回归）归一化流（Rezende 和 Mohamed，2015；Huang 等人，2018）。该理论引出了关于观测值 $\mathbf{x}$ 的结构方程模型的以下定义：

$$x_j = e^{\alpha_j(\mathbf{PA}_j)} z_j + \beta_j(\mathbf{PA}_j), \quad j = 1, 2 \tag{24}$$

其中， z_1, z_2 是统计独立的潜在噪声变量，而 α_j(**PA**j) 和 β_j(**PA**j) 是标量值函数；当不存在父节点时，它们被定义为常数（相对于 $\mathbf{x}$）。这个仿射因果模型通过在结构方程模型中为噪声变量 z_j 添加依赖于原因的系数，从而推广了加性噪声模型（Hoyer 等人，2009年）；因此，扰动项 $e^{\alpha_j(\mathbf{PA}j)} z_j$ 与原因 β_j(**PA**j) 并不独立，这与大多数模型不同。（这里采用 Khemakhem 等人（2021年）的记法，其中用 z 代替 e，右端的两项顺序互换，并用希腊字母表示函数；但与加性噪声模型的真正区别在于噪声的调制。）Khemakhem 等人（2021年）在非常一般的条件下证明了该模型的可识别性。Immer 等人（2022年）以及 Strobl 和 Lasko（2022年）的近期工作修改并推广了可识别性条件。

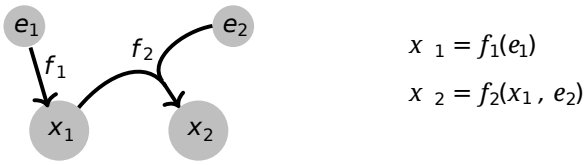


图4：与方程（25）和（26）中的结构方程模型相关的有向无环图 $\mathcal{G}$ 的可视化。相关结构方程显示在右侧。请注意，与图2不同，图中也显示了扰动变量。

6.3 一般非线性情况下的可辨识性理论

接下来，我们介绍一种针对一般非线性的非线性结构方程模型估计方法；该方法也由此得出了一个可识别性结果。我们假设观测到双变量数据 $\mathbf{x}(t) \in \mathbb{R}^2$，其中 t 为所有观测值提供索引（例如，t 可以是时间索引，但这并非必要）。

重要的是，我们进一步假设在一组不同的环境条件或片段 $C \in \mathcal{C}$ 下可以获得数据。因此，每个 $\mathbf{x}(t)$ 都被分配到一个这样的环境条件 $C(t) \in \mathcal{C}$。在因果发现中，这些环境条件可能源于不同的干预，也可能是所观测系统内部动力学发生变化所致。为了使所提出的方法与第5节回顾的非线性独立成分分析术语保持一致，我们指出，可以将每种环境条件视为基于非平稳性的非线性独立成分分析中的一个独立片段。如上所述，片段应作非常宽泛的理解，正如基于辅助变量的非线性独立成分分析理论中所述（Hyv¨arinen 等人，2019年；Khemakhem 等人，2020a）。

接下来，我们概述 Monti 等人（2019）提出的双变量数据因果发现方法，他们将其称为利用非平稳性的非线性结构方程模型估计（NonSENS）。不失一般性，我们假设 $x_1 \to x_2$，并据此解释其基本逻辑；相应的结构方程模型形式如下：

$$x_1(t) = f_1(e_1(t)), \tag{25}$$
$$x_2(t) = f_2(x_1(t), e_2(t)), \tag{26}$$

其中，e_1，e_2 是潜在扰动项，假设其分布会随环境条件变化，从而产生非平稳性。方程（25）和（26）对应的有向无环图如图4所示。这里的关键思想是，双变量结构方程模型中的潜在扰动项**e**，可以被视为对应于非线性独立成分分析模型中的独立源**s**，这与上文基于 Shimizu等人（2006）所回顾的原始 LiNGAM理论颇为相似。

所提出的 NonSENS 算法由一个两步程序组成。首先，它尝试通过基于非平稳性的非线性独立成分分析来恢复潜在扰动。

我们注意到，由各种环境条件 $C \in \mathcal{C}$引入的非平稳性意味着，Hyv¨arinen 和 Morioka（2016年）的方法非常适合恢复源变量 e_1 和 e_2。获得估计的成分后，我们可以利用观测数据与估计源之间统计独立性的相关知识来推断因果结构，这一推断基于 Monti 等人（2019年）指出的成分的一项有趣的独立性性质。用 $x \perp\!\!\!\perp y$ 表示 x 和 y的统计独立性，用 $\not\perp\!\!\!\perp$表示不独立。我们有：

命题1 假设真实的因果结构遵循方程（*25*）和（*26*），如图*4*所示。由此可得，$x_1 \perp\!\!\!\perp e_2$，而 $x_1 \not\perp\!\!\!\perp e_1$ 和$x_2 \not\perp\!\!\!\perp e_1$，以及 $x_2 \not\perp\!\!\!\perp e_2$。

该命题揭示了观测值**x** 与真实潜在源**e**之间的关系：原因与其中一个扰动项独立，而效应与所有扰动项都不独立。在实践中，非线性独立成分分析返回的估计潜在源，其排列顺序是随机的。因此，我们必须对变量 x_1 和 x_2分别与每个估计潜在源$\hat{e}_1$ 和$\hat{e}_2$进行独立性检验。这总共需要进行四次不同的独立性检验。

因此，所提出的方法可概述如下：通过非线性独立成分分析估计潜变量（的逐点变换）后，我们进行四次统计独立性检验；对于其中一个方向，如果有证据表明三项检验拒绝独立性零假设，而仅有一项检验未能拒绝零假设，则得出存在因果效应的结论。在这种情况下，将未拒绝零假设所对应的观测变量识别为原因变量。

最后需要指出的是，以上基于 Monti 等人（2019）所述的命题和方法，实际上构成了一种构造性可识别性证明：在任意非线性结构方程模型中，效应方向是可辨识的；其基本假设是，数据被划分为若干条件，并且这些条件满足基于非平稳性的非线性独立成分分析方法的假设。

7 讨论

本节将讨论可识别性在实际应用中的效用、未来研究的一些问题，以及算法开发的问题。

7.1 可识别性的效用

7.1.1 可解释性；寻找因果方向

当模型的参数（或潜变量）能够直接提供关于所分析现象的有用信息时，可识别性的价值显而易见。这在因果分析中尤为明显，

因为因果分析的核心就是找出哪些变量导致哪些结果；如果模型不可辨识，通常就根本无法进行分析。同样，在线性独立成分分析中，得到的成分往往对应着有意义的现象，例如盲源分离的情况（图 1）。这一方法已在神经科学等领域得到广泛应用，例如，分离出的源可能对应于大脑中的活动源（Hyv¨arinen 等，2010）。

7.1.2 特征提取；半监督与迁移学习

人们对可识别性在非线性独立成分分析中的实用价值了解较少。在深度学习中，成分通常很难解释，因为神经网络所实现的非线性难以理解或可视化。因此，非线性独立成分分析的可识别性的主要实用价值可能体现在其他方面。事实上，通过无监督深度学习学习到的表征通常会被用于某种后续目的；例如，特征会被用于进一步的分类任务。因此，问题在于，成分的可识别性是否对这类任务有用。

首先需要指出的是，即使无法解释成分的计算过程，单个成分仍可能对分类和决策有用。例如，在生物医学场景中，单个成分可能充当生物标志物，而发现此类生物标志物显然需要可识别性。在更一般的场景下，Zhu 等（2023）表明，非线性独立成分分析对于在半监督场景下分析神经科学数据非常有用；这意味着，首先从一个大型通用数据集中学习成分，然后将其用于较小的数据集，通过简单的线性分类器解决特定的分类任务。类似地，Khemakhem 等（2020b）提出了一些理论，说明可辨识表示对于迁移学习尤其有用。迁移学习是一项相关任务：我们拥有来自许多不同数据集（例如生物医学场景中的受试者）的数据，然后希望将模型泛化到新的受试者。尽管这些进展指出了非线性独立成分分析的经验实用价值，但它们尚未完全证明非线性独立成分分析的可识别性的实用价值；关于这一主题还需要进一步研究。

7.2 未来研究的问题

7.2.1 将因果分析与特征学习相结合

本文中的方法要么是寻找某些隐藏因子，要么是进行因果发现。一个非常有趣的课题是将两者结合起来，使模型能够学习隐藏因子及其之间的因果连接 。这正是一些人所称的“因果表示学习”。在线性情形下，Zhang 和 Hyv¨arinen（2009年）以及 Monti 和 Hyv¨arinen（2018）提出了此类方法。非线性情形下，这是深度学习领域当前备受关注的课题，例如可参见 Lachapelle 等人（2022年）以及 Morioka 和 Hyv¨arinen（2023年）的一些进展。

7.2.2 相关成分；含混杂因素的因果发现

事实上，潜变量之间存在因果依赖的模型，是潜变量彼此不独立的模型中的一种特殊情形。Khemakhem 等（2020b）提出了针对这一情形的通用框架，他们扩展了非线性独立成分分析，称之为独立调制成分分析（IMCA），在这种方法中，允许成分彼此依赖。其核心思想是，假定这些依赖关系具有平稳性，或者说独立于辅助变量。允许存在此类依赖关系可能会在许多情境中发挥作用，也是一个极具前景的未来研究课题。

IMCA 框架在因果发现中的另一项潜在应用，是允许存在混杂因素。如前所述，混杂因素是同时影响因变量和自变量的隐藏变量，会在因果图中造成虚假关联。我们可以建立 IMCA 与含混杂因素的结构方程模型（SEM）之间的联系，因为扰动项之间的依赖关系可以视为由未观测到的混杂因素引起。IMCA 的可识别性意味着此类 SEM 的因果方向同样可辨识。然而，由于 NonSENS 基于独立性检验，上文所述的基于检验的估计方法在此处无法使用。或许仍可采用一种非约束型方法，例如似然比度量（Hyv¨arinen 和 Smith，2013；Monti 等人，2019年）。

7.2.3 中间层的可识别性

如果非线性独立成分分析中的非线性函数由神经网络建模，那么我们估计的远不止单个非线性函数。事实上，神经网络的中间层常被用作后续分类任务的有效特征。在某些应用中，它们甚至可能优于最终层学习到的表征（Alain 和 Bengio，2018；Chen 等人，2020年；Mikolov 等人，2013）。由此引出一个引人深思的问题：能否将非线性独立成分分析所学习表征的可识别性结果推广到前面的层？Khemakhem 等（2020b）表明，某些此类神经网络架构实际上是可辨识的，他们采用了一种归纳方法，在网络中向前“传递可识别性”。因此，一个潜在的研究方向是证明，神经网络中位于最终层之前的中间层能够继承最终层的可识别性保证。将此类证明推广到卷积网络尤其值得关注，因为卷积网络常用于图像学习，并且具有坚实的数学理论基础（Wiatowski 和 B¨olcskei，2017年）。

7.3 估计算法

最后，我们简要讨论一下有哪些估计方法可用。基本上，我们只需要用于估计独立成分分析（包括非线性独立成分分析）的方法，因为结构方程模型的估计可以归结为独立成分分析。在机器学习中，我们通常将估计方法区分为两个部分：1）

估计原理（例如最大似然），通常会导出一个目标函数；以及 2）计算算法，通常用于优化该目标函数。对于线性独立成分分析，我们提出了 FastICA（Hyv¨arinen，1999年），该方法得到广泛应用，最初也被 Shimizu 等人（2006）用于求解结构方程模型的估计，尽管随后 Shimizu 等人（2011）以及 Hyv¨arinen 和 Smith（2013）等人又开发了其他结构方程模型方法。

对于非线性独立成分分析，我们提出了几种方案；深入介绍请参见 Hyv¨arinen 等人（2023年）。首先，我们提出了“自监督”方法（Hyv¨arinen 和 Morioka，2016年、2017年；Hyv¨arinen 等人，2019年），这是一类目前在机器学习领域备受关注的方法。随后，我们提出了最大似然估计，首先研究的是带有加性噪声的模型（Khemakhem 等人，2020a；H¨alv¨a 等人，2021年），这类模型支持变分近似。本文所用基本无噪声模型中的最大似然估计起初看起来很简单，但实际上存在严重的计算问题；我们已在很大程度上解决了这些问题（Gresele 等人，2020a）。无论如何，开发更好的算法无疑也是未来研究的一个有趣课题。

有了估计算法之后，就可以考察其有限样本表现。对于线性独立成分分析，已有多位作者分析了渐近方差（统计效率），包括 Cardoso 和 Laheld（1996年）、Pham 和 Garrat（1997年）、Hyv¨arinen（1997年）以及 Tichavsky 等人（2006年）；这些结果或多或少也可以直接应用于线性结构方程模型的估计。关于抗离群值稳健性的分析也已在一定程度上开展，涵盖线性和非线性情形（Hyv¨arinen，1997年；Sasaki 等人，2020年）。对于非线性情形，据我们所知，除了通过模拟进行的分析外，还没有关于统计效率的分析。这些显然都是未来研究中值得关注的问题。

8 结论

我们从线性高斯因子分析模型谈起。早在20世纪50年代，甚至更早，人们就已知道该模型不可识别。独立成分分析解决了可识别性问题：它是一种非高斯因子分析模型，能够以严格的意义实现可辨识，即恢复出实际生成数据的成分，如盲源分离的情形。线性独立成分分析还可用于估计线性结构方程模型，从而实现可识别的因果发现。独立成分分析可以推广到非线性情形，但这绝非易事：要实现可识别性，还需要新的假设。本文主要考察了时间依赖关系和非平稳性（后者采用了非常广义的定义）。限制非线性混合也可能奏效，但迄今提出的成功模型寥寥无几。与线性情形一样，我们最后还看到，非线性独立成分分析同样能够实现非线性结构方程模型的可识别性与估计。

参考文献

Alain, G. 和 Bengio, Y.（2018）。使用线性分类器探针理解中间层。*arXiv*预印本 *arXiv:1610.01644*。Belouchrani, A., Meraim, K. A., Cardoso, J.-F. 和 Moulines, E.（1997）。一种基于二阶统计量的盲源分离技术。*IEEE*《信号处理汇刊》，45(2):434–444。Bollen, K. A.（1989）。含潜变量的结构方程。John Wiley & Sons。Brookes, M., Woolrich, M., Luckhoo, H., Price, D., Hale, J., Stephenson, M., Barnes, G., Smith, S. 和 Morris, P.（2011）。利用脑磁图研究静息态网络的电生理基础。美国国家科学院院刊，108:16783–16788。Buchholz, S., Besserve, M. 和 Sch¨olkopf, B.（2022）。用于可辨识非线性独立成分分析的函数类。*arXiv*预印本 *arXiv:2208.06406*。Cardoso, J.-F.（2001）。实现独立成分分析的三条简便途径：对比与几何。载于独立成分分析与盲信号分离国际研讨会论文集（*ICA2001*），加利福尼亚州圣迭戈。Cardoso, J.-F. 和 Laheld, B. H.（1996）。等变自适应源分离。*IEEE*《信号处理汇刊》，44(12):3017–3030。Chen, T., Kornblith, S., Norouzi, M. 和 Hinton, G.（2020）。视觉表征对比学习的一个简单框架。*2002.05709*。Comon, P.（1994）。独立成分分析——一个新概念？《信号处理》，36:287–314。Donoho, D. L. 和 Grimes, C.（2003）。Hessian特征映射：面向高维数据的局部线性嵌入技术。美国国家科学院院刊，100(10):5591–5596。Donoho, D. L. 和 Stodden, V.（2004）。非负矩阵分解何时能将数据正确分解为各个部分？载于神经信息处理系统进展*16*（*NIPS*2003*论文集）。麻省理工学院出版社。Eriksson, J. 和 Koivunen, V.（2004）。线性独立成分分析模型的可识别性、可分离性与唯一性。*IEEE*《信号处理快报》，11(7):601–604。Flanders, H.（1966年）。保角映射的刘维尔定理。《数学与力学杂志》，15(1):157–161。Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. 和 Bengio, Y.（2014年）。生成对抗网络。载于神经信息处理系统进展，第2672–2680页。

Gresele, L., Fissore, G., Javaloy, A., Sch¨olkopf, B., and Hyv¨arinen, A.（2020a）。无监督深度学习中雅可比项的相对梯度优化。载于神经信息处理系统进展（*NeurIPS2020*），线上。Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Sch¨olkopf, B.（2020b）。不完整罗塞塔石碑问题：多视图非线性独立成分分析的可识别性结果。载于人工智能中的不确定性，第 217–227 页。PMLR。Gresele, L., Von K¨ugelgen, J., Stimper, V., Sch¨olkopf, B., and Besserve, M.（2021）。独立机制分析，一种新概念？神经信息处理系统进展，34:28233–28248。H¨alv¨a, H., Corff, S. L., Leh´ericy, L., So, J., Zhu, Y., Gassiat, E., and Hyv¨arinen, A.（2021）。利用结构化非线性独立成分分析，从噪声数据中解缠结出可识别特征。载于神经信息处理系统进展（*NeurIPS2021*），线上。H¨alv¨a, H. and Hyv¨arinen, A.（2020）。隐马尔可夫非线性独立成分分析：从非平稳时间序列中进行无监督学习。载于第 *36* 届人工智能中的不确定性会议论文集（*UAI2020*），加拿大多伦多（线上）。Harman, H. H.（1967）。现代因子分析。芝加哥大学出版社，第 2 版。Harmeling, S., Ziehe, A., Kawanabe, M., and M¨uller, K.-R.（2003）。基于核的非线性盲源分离。*Neural Computation*，15(5):1089–1124。Horan, D., Richardson, E., and Weiss, Y.（2021）。何时可以实现无监督解缠结？神经信息处理系统进展，34:5150–5161。Hoyer, P. O.（2004）。带稀疏性约束的非负矩阵分解。机器学习研究杂志，5:1457–1469。Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., and Sch¨olkopf, B.（2009）。基于加性噪声模型的非线性因果发现。载于神经信息处理系统进展，第 21 卷，第 689–696 页。麻省理工学院出版社。Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M.（2008）。利用含隐变量的线性非高斯因果模型估计因果效应。国际近似推理杂志，49:362–378。Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A.（2018）。神经自回归流。载于国际机器学习会议，第 2078–2087 页。PMLR。

Hyttinen, A., Barin-Pacela, V., and Hyv¨arinen, A.（2022）。二元独立成分分析：一种基于非平稳性的方法。载于第 *38* 届人工智能中的不确定性会议论文集（*UAI2022*），第 874–884 页，荷兰埃因霍温。Hyv¨arinen, A.（1997）。独立成分分析的单单元对比函数：一种统计分析。载于信号处理神经网络 *VII*（*IEEE* 信号处理神经网络研讨会论文集），第 388–397 页，美国佛罗里达州阿米莉亚岛。Hyv¨arinen, A.（1999）。用于独立成分分析的快速稳健不动点算法。*IEEE* 神经网络汇刊，10(3):626–634。Hyv¨arinen, A., Hurri, J., and Hoyer, P. O.（2009）。自然图像统计。施普林格出版社。Hyv¨arinen, A., Karhunen, J., and Oja, E.（2001）。独立成分分析。Wiley Interscience。Hyv¨arinen, A., Khemakhem, I., and Morioka, H.（2023）。用于无监督深度学习中有原则解缠结的非线性独立成分分析。*arXiv* 预印本 *arXiv:2303.16535*。Hyv¨arinen, A. and Morioka, H.（2016）。通过时间对比学习和非线性独立成分分析进行无监督特征提取。载于神经信息处理系统进展（*NIPS2016*），西班牙巴塞罗那。Hyv¨arinen, A. and Morioka, H.（2017）。对时间依赖平稳源进行非线性独立成分分析。载于人工智能与统计会议论文集（*AISTATS2017*），美国佛罗里达州劳德代尔堡。Hyv¨arinen, A. and Oja, E.（2000）。独立成分分析：算法与应用。《神经网络》，13(4-5):411–430。Hyv¨arinen, A. and Pajunen, P.（1999）。非线性独立成分分析：存在性与唯一性结果。《神经网络》，12(3):429–439。Hyv¨arinen, A., Ramkumar, P., Parkkonen, L., and Hari, R.（2010）。利用短时傅里叶变换对自发脑电图/脑磁图进行独立成分分析。*NeuroImage*，49(1):257–271。Hyv¨arinen, A., Sasaki, H., and Turner, R.（2019）。使用辅助变量和广义对比学习进行非线性独立成分分析。载于人工智能与统计会议论文集（*AISTATS2019*），日本冲绳。Hyv¨arinen, A. and Smith, S. M.（2013）。用于估计非高斯结构方程模型的成对似然比。机器学习研究杂志，14:111–152。

Immer, A., Schultheiss, C., Vogt, J. E., Sch¨olkopf, B., B¨uhlmann, P., and Marx, A.（2022年）。关于因果位置-尺度噪声模型的可识别性与估计。*arXiv*预印本 *arXiv:2210.09054*。Jakobsen, M. E., Shah, R. D., B¨uhl mann, P., and Peters, J.（2022年）。有向树的结构学习。机器学习研究杂志，23:159。Jutten, C. and H´erault, J.（1991年）。源盲分离，第一部分：一种基于神经拟态架构的自适应算法。《信号处理》，24:1–10。Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyv¨arinen, A.（2020a）。变分自编码器与非线性独立成分分析：一个统一框架。载于人工智能与统计会议论文集（*AISTATS2020*）。Khemakhem, I., Monti, R. P., Kingma, D. P., and Hyv¨ arinen, A.（2020b）。ICE-BeeM：基于非线性独立成分分析的可辨识条件能量深度模型。载于神经信息处理系统进展（*NeurIPS2020*），线上。Khemakhem, I., Monti, R. P., Leech, R., and Hyv¨arinen, A.（2021年）。因果自回归流。载于人工智能与统计会议论文集（*AISTATS2021*），线上。Kingma, D. P. and Welling, M.（2014年）。自编码变分贝叶斯。载于国际表征学习会议论文集（*ICLR2014*），班夫，加拿大。Kivva, B., Rajendran, G., Ravikumar, P., and Aragam, B.（2022年）。没有辅助信息时混合先验下深度生成模型的可识别性。*arXiv*预印本 *arXiv:2206.10044*。Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D.（2020年）。利用时间稀疏编码实现自然数据的非线性解缠结。*arXiv*预印本 *arXiv:2007.10930*。Kumar, A. and Poole, B.（2020年）。关于 β- 变分自编码器中的隐式正则化。载于国际机器学习会议，第 5480–5490 页。PMLR。Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O.（2008年）。通过独立成分分析发现循环因果模型。载于第*24*届人工智能中的不确定性会议论文集（*UAI2008*），赫尔辛基，芬兰。Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S.（2022年）。通过机制稀疏正则化实现解缠结：非线性独立成分分析的新原则。载于因果学习与推理会议，第428–484页。PMLR。Lee, D. D. and Seung, H. S.（1999年）。通过非负矩阵分解学习物体的组成部分。《自然》，401:788–791。

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Sch¨olkopf, B., and Bachem, O.（2019年）。对解缠表示无监督学习中常见假设的挑战。载于国际机器学习会议，第4114–4124页。PMLR。Matsuoka, K., Ohya, M., and Kawamoto, M.（1995）。用于非平稳信号盲分离的神经网络。《神经网络》，8(3):411–419。Mikolov, T., Chen, K., Corrado, G., and Dean, J.（2013）。向量空间中词表示的高效估计。*arXiv*预印本 *arXiv:1301.3781*。Monti, R. P. and Hyv¨arinen, A.（2018）。一种从高维数据中学习潜在因子及其连接关系的统一概率模型。载于第*34*届人工智能中的不确定性会议论文集（*UAI2018*），蒙特雷，加利福尼亚州。Monti, R. P., Zhang, K., and Hyv¨arinen, A.（2019年）。利用非线性独立成分分析，对具有一般非线性关系的因果关系进行发现。载于第*35*届人工智能中的不确定性会议论文集（*UAI2019*），特拉维夫，以色列。Moran, G. E., Sridhar, D., Wang, Y., and Blei, D. M.（2021年）。通过稀疏解码实现可辨识变分自编码器。*arXiv*预印本 *arXiv:2110.10804*。Morioka, H., H¨alv¨a, H., and Hyv¨arinen, A.（2021年）。非线性向量自回归过程的独立创新分析。载于人工智能与统计会议论文集（*AISTATS2021*），线上。Morioka, H. and Hyv¨arinen, A.（2023）。连接对比学习：结合因果发现与表征学习以处理多模态数据。载于人工智能与统计会议论文集（*AISTATS2023*），巴伦西亚，西班牙。Nevanlinna, R.（1960年）。关于可微映射。解析函数，第3–9页。Olshausen, B. A. and Field, D. J.（1997年）。使用超完备基集进行稀疏编码：V1采用的一种策略?《视觉研究》，37:3311–3325。Paatero, P. and Tapper, U.（1994）。正矩阵分解：一种非负因子模型，可最优利用数据值的误差估计。《环境计量学》，5:111–126。Pearl, J.（2009年）。《因果关系》。剑桥大学出版社。Peters, J. and B¨uhlmann, P.（2014年）。相等误差方差高斯结构方程模型的可识别性。《生物计量学》，101(1):219–228。Peters, J., Janzing, D., and Sch¨olkopf, B.（2017年）。因果推断要素：基础与学习算法。麻省理工学院出版社。

Peters, J., Mooij, J. M., Janzing, D., 和 Sch¨olkopf, B.（2014）。使用连续加性噪声模型进行因果发现。机器学习研究杂志，15：2009–2053。Pham, D.-T. 和 Cardoso, J.-F.（2001）。非平稳源瞬时混合的盲分离。*IEEE*信号处理汇刊，49(9)：1837–1848。Pham, D.-T. 和 Garrat, P.（1997）。通过准最大似然方法对独立源混合进行盲分离。*IEEE*信号处理汇刊，45(7)：1712–1725。Plumbley, M. D.（2003）。非负独立成分分析算法。*IEEE* 神经网络汇刊，14(3)：534–543。Rezende, D. J. 和 Mohamed, S.（2015）。使用归一化流进行变分推断。*arXiv*预印本 *arXiv:1505.05770*。Sasaki, H., Takenouchi, T., Monti, R. P., 和 Hyv¨arinen, A.（2020）。存在离群值时的稳健对比学习与非线性独立成分分析。载于第*36*届人工智能不确定性会议论文集（*UAI2020*），加拿大多伦多（线上）。Schell, A. 和 Oberhauser, H.（2023）。面向离散时间和连续时间信号的非线性独立成分分析。统计年刊。Shimizu, S.（2014）。LiNGAM：用于估计因果结构的非高斯方法。行为计量学，41(1)：65–98。Shimizu, S., Hoyer, P. O., Hyv¨arinen, A., 和 Kerminen, A.（2006）。用于因果发现的线性非高斯无环模型。机器学习研究杂志，7：2003–2030。Shimizu, S., Inazumi, T., Sogawa, Y., Hyv¨arinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., 和 Bollen, K.（2011）。DirectLiNGAM：学习线性非高斯结构方程模型的直接方法。机器学习研究杂志，12：1225–1248。Spirtes, P., Glymour, C., Scheines, R., Heckerman, D., Meek, C., 和 Richard-son, T.（2000）。因果、预测与搜索。麻省理工学院出版社。Spirtes, P. 和 Zhang, K.（2016）。因果发现与推断：概念及近期方法学进展。应用信息学。Sprekeler, H., Zito, T., 和 Wiskott, L.（2014）。将慢特征分析扩展至非线性盲源分离。机器学习研究杂志，15(1)：921–947。Strobl, E. V. 和 Lasko, T. A.（2022）。使用异方差噪声模型识别患者特异性根本原因。*arXiv*预印本 *arXiv:2205.13085*。

Tashiro, T., Shimizu, S., Hyv¨arinen, A., 和 Washio, T.（2014）。ParceLiNGAM：一种能够抵御潜在混杂因素的因果排序方法。神经计算，26：57–83。Tichavsky, P., Koldovsky, Z., 和 Oja, E.（2006）。FastICA算法的性能分析及线性独立成分分析的Cramér–Rao界。*IEEE*信号处理汇刊，54(4)：1189–1203。Tong, L., Liu, R.-W., Soon, V. C., 和 Huang, Y.-F.（1991）。盲辨识的不确定性与可识别性。*IEEE*电路与系统汇刊，38：499–509。Wei, Y., Shi, Y., Liu, X., Ji, Z., Gao, Y., Wu, Z., 和 Zuo, W.（2021）。用于图像生成无监督解缠的正交雅可比矩阵正则化。载于*IEEE/CVF*国际计算机视觉会议论文集，第6721–6730页。Wiatowski, T. 和 B¨olcskei, H.（2017）。深度卷积神经网络特征提取的数学理论。*arXiv*预印本 *arXiv:1512.06293*。Willetts, M. 和 Paige, B.（2021）。我不需要 u：无需辅助信息的可辨识非线性独立成分分析。*arXiv*预印本 *arXiv:2106.05238*。Zhang, K. 和 Chan, L.（2008）。非线性独立成分分析的最小非线性失真原理。机器学习研究杂志，9：2455–2487。Zhang, K. 和 Hyv¨arinen, A.（2009）。后非线性因果模型的可识别性。载于第*25*届人工智能不确定性会议论文集（*UAI2009*），第647–655页，加拿大蒙特利尔。Zhang, K. 和 Hyv¨arinen, A.（2010）。脑磁图和脑电图的源分离与高阶因果分析。载于第*26*届人工智能不确定性会议论文集（*UAI2010*），加利福尼亚州卡塔利娜岛。Zhu, Y., Parviainen, T., Heinil¨a, E., Parkkonen, L., 和 Hyv¨arinen, A.（2023）。利用非线性独立成分分析对自发脑磁图数据进行无监督表征学习。$NeuroImage$，274(120142)。Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., 和 Brendel, W.（2021）。对比学习逆转数据生成过程。载于国际机器学习会议，第12979–12990页。PMLR。

A 刘维尔定理与非线性独立成分分析

使非线性独立成分分析可辨识的一种可能研究方向，是对变换的雅可比矩阵施加适当条件。这些条件通常与雅可比矩阵的正交性有关，也称为局部等距（Gresele 等人，2021年；Zimmermann 等人，2021年；Buchholz 等人，2022年）。在本附录中，我们考虑实空间内的可逆映射集合，即维数不变的映射，这也是独立成分分析理论中的典型情形。我们证明，局部等距的函数 $\mathbb{R}^n \to \mathbb{R}^n$ 必然是仿射函数，因此无法为非线性独立成分分析理论提供有意义的基础。为此，我们证明了刘维尔保角映射定理的一个变体。尽管这一结果可谓众所周知，我们仍为该变体给出一个非常简单的证明，而这类证明在文献中难觅踪迹。

A.1 定义

我们先给出一些定义：

定义 1 设**f** 是从 U $\mathbb{R}^n$ 的一个开子集到 $\mathbb{R}^n$ 的一个开子集 V 的可微映射。若满足以下条件，则称该映射为保角映射

$$\mathbf{Jf}(\mathbf{x})^T \mathbf{Jf}(\mathbf{x}) = c(\mathbf{x})\mathbf{I} \tag{27}$$

其中，**Jf** 是**f**的雅可比矩阵（偏导数矩阵），且 $c(\mathbf{x})$ 取正标量值。

从几何角度看，这一定义意味着映射保持角度不变，但缩放可以随 $c(\mathbf{x})$ 改变。一个重要的特殊情形是雅可比矩阵为正交矩阵：

定义 2 若在定义 *1* 中， $c(\mathbf{x}) \equiv 1$，则称该映射为局部等距。

从几何角度解释，这种映射必须同时保持角度和缩放不变。这与 Donoho 和 Grimes（2003年）采用的定义有关，Horan 等人（2021年）则以他们的定义为基础。不过，这些作者考虑的是映射**f**降低维数的情形，因此我们的结果相当不同。事实上，局部等距通常是在流形的语境下定义的，此类流形的维数低于空间本身，因此我们的定义是更传统定义的一个特例。雅可比矩阵的正交性也曾被更启发式地用作正则化项（Wei 等人，2021年；Kumar 和 Poole，2020年）。

为简洁起见，我们进一步引入以下术语：

定义 3 若函数 **f** 具有如下形式，则称其为正交仿射 函数：

$$\mathbf{f}(\mathbf{x}) = \mathbf{U}\mathbf{x} + \mathbf{b} \tag{28}$$

其中 **b** 为某个常向量，**U**为正交矩阵。

在某些语境中，这样的函数也称为“刚体运动”。

A.2 局部等距函数是正交仿射的

我们的分析基于刘维尔于1850年提出的一个关于保角映射的著名定理。为使表述简洁，我们仅给出该定理的一个变体，并请读者参阅相关文献了解原定理（Flanders，1966年；Nevanlinna，1960年）。我们的定理变体如下：

定理2 假设$\mathbf{f}$ 是局部等距的（定义*2*），并且属于 $\mathcal{C}^2$ （即具有连续二阶导数），且定义在维数为 $n \geq 2$ 的实空间中。那么$\mathbf{f}$ 是正交仿射的（定义*3*）。

证明： 此证明与 Flanders（1966）给出的刘维尔定理证明的前半部分密切相关；另一个广为人知的证明见 Nevanlinna（1960）。

为简化记号，我们省略自变量$\mathbf{x}$，并用$\mathbf{w} \cdot \mathbf{v}$ 表示点积。用$\mathbf{J}_i$ 表示$\mathbf{f}$的雅可比矩阵的第 i 列，即关于 x_i 的偏导数构成的向量。考虑由雅可比矩阵的正交性定义的等距矩阵方程中的第 i, j 个元素：

$$\mathbf{J}_i \cdot \mathbf{J}_j = \delta_{i=j} \tag{29}$$

对等式两边关于 x_k 求导：

$$\mathbf{J}_i \cdot \mathbf{J}_j^k + \mathbf{J}_j \cdot \mathbf{J}_i^k = 0 \tag{30}$$

其中$\mathbf{J}_i^k$是$\mathbf{J}_i$各元素关于 x_k 的偏导数构成的向量。因此，我们得到以下反对称性条件：

$$\mathbf{J}_i \cdot \mathbf{J}_j^k = -\mathbf{J}_j \cdot \mathbf{J}_i^k \tag{31}$$

另一方面，由于求导的次序可以交换，我们有对称性条件：

$$\mathbf{J}_i \cdot \mathbf{J}_j^k = \mathbf{J}_i \cdot \mathbf{J}_k^j \tag{32}$$

显然，既对称又反对称的矩阵必定为零。这里我们实际处理的是一个张量，但接下来将证明同一原理也适用于这样的张量；这一结果有时称为辫子引理。任取指标 i、j、k，并交替应用 (31) 和 (32)，共三次。我们得到

$$\mathbf{J}_i \cdot \mathbf{J}_j^k = -\mathbf{J}_j \cdot \mathbf{J}_i^k = -\mathbf{J}_j \cdot \mathbf{J}_k^i = \mathbf{J}_k \cdot \mathbf{J}_j^i = \mathbf{J}_k \cdot \mathbf{J}_i^j = -\mathbf{J}_i \cdot \mathbf{J}_k^j = -\mathbf{J}_i \cdot \mathbf{J}_j^k \tag{33}$$

首项与末项相等，这表明它们必定为零。这对所有 i 都成立，而 $\mathbf{J}_i$ 构成一个正交基，因此 $\mathbf{J}_j^k$ 为零。于是，所有二阶偏导数在每一点都消失，$\mathbf{f}$ 必定是仿射的。根据等距的定义，它还必定是正交仿射的。 □

我们的定理扩展了刘维尔理论，因为我们的定理适用于 $n \geq 2$，而刘维尔定理假设 $n \geq 3$。另一方面，

我们的结果是一个特例，因为刘维尔定理假设保角性，而我们假设局部等距。鉴于刘维尔定理，要证明一个针对 $n \geq 3$ 给出所需结果的推论会非常容易，但我们更愿意证明一个并非严格推论、同时又构成扩展的定理，而不借助这类高级理论。事实上，在我们的情形下可以给出一个非常简单的证明，而刘维尔定理似乎只有非常复杂的证明方法。

这意味着，若假设非线性独立成分分析模型（或任何深度潜变量模型）中的（维数保持）混合是局部等距的，那么该模型就会直接退化为线性独立成分分析模型。（请注意，这里我们没有用到任何成分的统计性质。）从这一角度看，局部等距这一假设过于严格。