arXiv:2303.17762v1 [cs.IT] 31 Mar 2023

高斯变量的广义信息瓶颈

Vudtiwat Ngampruetikorn 和 David J. Schwab纽约市立大学研究生中心理论科学倡议，纽约州纽约市 *10016*，美国

信息瓶颈（IB）方法为理解表征学习提供了一个颇具吸引力的框架，但其应用往往受限于计算上的难以处理。对信息瓶颈（IB）方法进行解析刻画不仅具有实际意义，也可能为学习现象带来新的见解。本文研究一种广义信息瓶颈问题，其中原始信息瓶颈方法中的互信息被基于 Rényi 散度和杰弗里斯散度的相关性度量所取代。我们针对高斯相关变量推导出精确的解析IB解。分析揭示了一系列结构转变，与此前在原始信息瓶颈情形中观察到的转变类似。我们进一步发现，尽管求解原始、Rényi 和杰弗里斯信息瓶颈问题通常会得到不同的表征，但这些结构转变发生在相同的临界权衡参数处，而且 Rényi 和杰弗里斯信息瓶颈解在原始信息瓶颈目标下表现良好。我们的结果表明，采用其他相关性度量来构建信息瓶颈（IB）方法，可能为原始信息瓶颈问题提供一种近似求解策略。

I. 信息瓶颈

数据的有效表征是实现泛化学习的关键。阐明什么样的表征才是好的，以及它如何形成，对于理解现代机器学习的成功至关重要。信息瓶颈（IB）方法是一种用于表征学习的信息论表述 [1]——它已被证明是研究这一问题特别有用的概念框架，并加深了我们对监督学习和自监督学习中表征学习的理解 [2–4]。研究这种对表征学习的抽象，有望为学习问题带来新的见解。

要量化一种表征的优劣，必须了解希望从数据中学习什么。信息瓶颈理论利用了这样一个事实：在许多情况下，我们可以通过引入一个额外变量来定义相关信息；例如，该变量可以是分类任务中每张图像的标签。这种相关性的概念使我们能够精确定义最优性——一个IB最优表征 T 能够最大限度地预测相关变量 Y，同时尽可能减少从数据中提取的比特数 X。IB方法将这一原则表述为一个优化问题 [1],

$$\min_{Q_{T|X}} \; I(T;X) - \beta I(T;Y). \tag{1}$$

这里的优化针对编码器进行 $Q_{T|X}$，编码器提供从 X 到 T 的（随机）映射。最大化互信息 $I(T;Y)$ [即式（1）中的第二项]鼓励表征 T 编码更多相关信息，同时最小化 $I(T;X)$ [即式（1）中的第一项] 则抑制其编码无关信息位。参数 $\beta > 0$ 控制着这两个信息项之间的基本权衡。

信息瓶颈方法已在许多应用中证明有效，包括神经编码 [5, 6], 统计物理 [7–9],聚类 [10], 深度学习 [2, 3, 11], 强化学习 [12] 以及学习理论 [13–16]。然而，信息瓶颈问题的非线性特性使其计算代价高昂。尽管基于信息瓶颈原理的可扩展学习方法得以实现，这要归功于互信息的变分界 [11, 17, 18], ，但这类界的选择以及具体实现细节可能会引入很强的归纳与原始目标相竞争的偏差 [19]。

信息瓶颈方法的精确形式通常难以应用于大规模问题，但也存在一些特殊情形。例如，在信息量较低的极限下——也就是说，式（1）中的两项都很小——可以用微扰理论来描述，它提供了一种方法，用于找出能够使每提取一比特信息所获得的相关信息最大化的表征 [20, 21]。不过，也许最重要的特殊情形是源 X和目标 Y 为高斯相关随机变量。在这种情况下，存在精确的解析解 [22]。

尽管信息瓶颈方法最初采用香农互信息来表述，但其基本权衡具有更普遍的适用性：将信息项替换为适当的互依赖度量后，信息瓶颈优化 [式（1）]仍然定义良好。在本工作中，我们研究基于两种重要相关性度量的广义信息瓶颈问题。第一种是基于 Rényi 散度的香农信息参数化推广 [23]。基于 Rényi 的互信息和熵推广是量化量子纠缠的核心工具 [24, 25] ，并已证明是蒙特卡罗模拟中的有力工具 [26–28] 以及实验中的有力工具 [29–31]。我们考虑的第二种互依赖度量基于 Jeffreys散度 [32]。由此得到的 Jeffreys 信息（除一个常数因子外）等于一大类学习算法的泛化差距，这类算法称为 Gibbs算法 [33]。

我们推导出了一种解析IB解，适用于X 与 Y 呈高斯相关的情形，将文献 [22] 中的结果推广到一类信息论互依赖性度量，其中香农信息是其一个极限情形。我们证明，对于Rényi和杰弗里斯两种情形，均可由归一化回归矩阵 $\Sigma_{X|Y}\Sigma_X^{-1}$的特征模态构造最优编码器。我们的解揭示了一系列相变，与高斯IB方法 [22]中观察到的相变类似。在Rényi和杰弗里斯两种情形下，我们发现，尽管最优编码器取决于信息度量，但相变发生在临界权衡参数 $\beta_c^{(i)}$处；这些参数与香农情形下的参数一致，且与Rényi信息的阶数无关。

II. 基于散度的相关性度量

当两个随机变量 X 与 Y 不相关时，它们的联合分布 P_{XY} 等于其边缘分布P_X 和 P_Y的乘积。因此，我们可以通过 P_{XY} 与 $P_X \otimes P_Y$之差来量化 X 与 Y 之间的相互依赖，

$$\Omega(X;Y) \equiv \mathcal{D}(P_{XY} \parallel P_X \otimes P_Y) \geq 0. \tag{2}$$

这里，$\mathrm{D}(P \parallel Q)$ 表示一种统计散度，根据定义，该散度非负，且当且仅当 $P = Q$ 时为零。当使用 Kullback–Leibler（KL）散度定义时，上述度量即为香农信息，$I(X;Y) = \mathrm{KL}(P_{XY} \parallel P_X \otimes P_Y)$。

III. Rényi$Q$–信息

我们考虑一种基于 Rényi 散度的相关性度量 [23]。更准确地说，我们将*Rényi* q-信息定义为

$$I_q(X;Y) \equiv \mathcal{R}_q(P_{XY} \parallel P_X \otimes P_Y), \tag{3}$$

其中，$\mathrm{R}q$ 表示阶数为 q 的 Rényi 散度，

$$\mathcal{R}_q(P \parallel Q) = \frac{1}{q-1} \ln \int dQ \left(\frac{dP}{dQ}\right)^q \tag{4}$$

对于 $q \in (0,1) \cup (1,\infty)$。通过在 q 中的连续性，此定义可扩展至 $q = 0, 1$ 和∞。特别地，$\mathrm{R}_1(P\|Q) = \mathrm{KL}(P\|Q)$ [34,定理 5],，因此 $I_1(X;Y) = I(X;Y)$。Rényi散度以及由此得到的 q-信息满足数据处理不等式，因为它们与某个 f-散度 [呈严格递增关系，而 $f(t) = (t^q - 1)/(q-1)$] 具有这一性质，参见例如文献 [35]。

A. 高斯变量

对于高斯相关变量 X和 Y，q-信息可表示为（推导见附录 A）

$$I_q(X;Y) = \frac{-1}{2\bar{q}} \ln \frac{|\Sigma_{X|Y}\Sigma_X^{-1}|^{\bar{q}}}{|I - \bar{q}^2(I - \Sigma_{X|Y}\Sigma_X^{-1})|} \quad \text{with} \quad \bar{q} = 1 - q, \tag{5}$$

其中，I 表示相容维数下的单位矩阵。可以看出，这一信息仅通过归一化回归矩阵 $\Sigma_{X|Y}\Sigma_X^{-1}$ 依赖于协方差矩阵。还需注意，当 $q > 2$ 时，这一信息可能发散，因为 $\Sigma_{X|Y}\Sigma_X^{-1}$ 的特征值介于零和一之间 [22,引理 B.1]。容易验证，香农信息对应于极限 $q \to 1$,

$$I(X;Y) = \lim_{q\to 1} I_q(X;Y) = -\frac{1}{2} \ln |\Sigma_{X|Y}\Sigma_X^{-1}|. \tag{6}$$

此外，我们注意到，对于高斯变量，$I_2(X;Y) = 2I(X;Y)$ 和 $I_q(X;Y)$会随着 q 增大，并在 $q = 0$ 处从零开始。

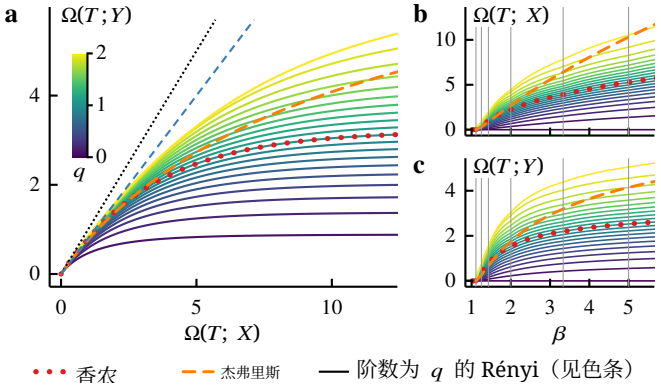


图 1. a 基于广义相关性度量 $\Omega(A;B)$ 的广义信息瓶颈方法的最优前沿，包括香农（$\Omega = I$）、杰弗里斯（$\Omega = J$）和 Rényi（$\Omega = I_q$）信息量（见图例）。对于 Rényi 情形，我们展示了一系列瑞尼阶数 q 的结果（见色标）。需要强调的是，尽管香农信息等价于阶数为一的 Rényi 信息（$q = 1$），杰弗里斯信息并不是 Rényi 信息的特例。相关信息 $\Omega(T;Y)$ 受数据处理不等式（DPI，黑色点线）$\Omega(T;Y) \leq \Omega(T;X)$ 的限制。我们还展示了数据依赖的紧界版本 DPI，即强数据处理不等式（SDPI，蓝色虚线）$\Omega(T;Y) \leq (1 - \lambda\min)\Omega(T;X)$，其中 λmin 是归一化回归矩阵 $\Sigma_{X|Y}\Sigma_X^{-1}$ 的最小特征值。请注意，对于所示的所有信息度量，DPI 和 SDPI 均相同。b-c 提取比特和相关比特 $\Omega(T;X)$ 与 $\Omega(T;Y)$ 分别随权衡参数 β 增大，并在低于临界值 $\beta_c = 1/(1 - \lambda\min)$ 时消失。竖线标示临界权衡参数 [（式 (22) 和 (35)）] 的位置。这里 $\Sigma_{X|Y}\Sigma_X^{-1}$ 的特征值为 $\lambda_i = 0.1$、 0.2、 0.3、 0.5、 0.7、 0.8。

请注意，Rényi 互信息存在其他定义。在物理学文献中，一种常用定义为 $I_q(X;Y) = S_q(X) + S_q(Y) - S_q(X,Y)$，其中 $S_q(X) = (1-q)^{-1} \ln \int dx\, p_X(x)^q$ 是 q 阶 Rényi（微分）熵。然而，对于高斯变量，无论 q 如何，该定义得到的 Rényi 信息都等于香农信息；因此，由此得到的 Rényi 信息瓶颈问题与原始信息瓶颈问题完全相同。

IV. 高斯变量的 Rényi 信息瓶颈

在原始信息瓶颈目标函数[式（1）]中将互信息 q替换为 q-信息后，得到

$$\mathcal{L}_q[Q_{T|X}] = I_q(T;X) - \beta I_q(T;Y) \tag{7}$$

其中，X 表示源数据，Y 表示目标变量，T 表示 X的表征。损失函数随编码器 $Q_{T|X}$ 而变化，该编码器提供从 X 到T的随机映射。一般而言，q-信息项的阶数不必相同，但只有在 $q = q'$时，才能保证数据处理不等式 $I_q(T;X) \geq I_{q'}(T;Y)$。

我们针对 X 和 Y 服从高斯相关的情形，考察一族带噪线性编码器，

$$T = AX + \xi \quad \text{with} \quad \xi \sim N(0, \Sigma_\xi). \tag{8}$$

由于 Rényi散度在随机变量经过可逆变换后保持不变 [参见公式（4）], 我们可以对 T进行变换，使得 Σ_ξ 变为单位矩阵，同时不改变信息量。在以下分析中，我们不失一般性地令 $\Sigma_\xi = I$。也就是说，编码器变为一个高斯信道，其参数仅由矩阵 A表示，

$$T \mid X \sim N(AX, I). \tag{9}$$

为了计算式（7）中的信息，我们首先将X 从上述方程中边缘化，得到

$$T \sim N(A\mu_X, I + A\Sigma_X A^\mathsf{T}) \tag{10}$$

$$T \mid Y \sim N(A\mu_{X|Y}, I + A\Sigma_{X|Y} A^\mathsf{T}), \tag{11}$$

其中我们使用 $X \sim N(\mu_X,\ \Sigma_X)$和 $X \mid Y \sim N(\mu_{X|Y},\ \Sigma_{X|Y})$。将上述方程中的协方差矩阵代入式（5）可得

$$I_q(T;X) = \frac{-1}{2\bar{q}} \ln \frac{|I + A\Sigma_X A^\mathsf{T}|^q}{|I + (1-\bar{q}^2)A\Sigma_X A^\mathsf{T}|} \tag{12}$$

$$I_q(T;Y) = \frac{-1}{2\bar{q}} \ln \frac{|I + A\Sigma_{X|Y} A^\mathsf{T}|^{\bar{q}} |I + A\Sigma_X A^\mathsf{T}|^q}{|I + A[I - \bar{q}^2(I - \Sigma_{X|Y}\Sigma_X^{-1})]\Sigma_X A^\mathsf{T}|}. \tag{13}$$

按照参考文献 [22], 的分析，我们定义混合矩阵W，使得

$$A = WV, \tag{14}$$

其中 V 是归一化回归矩阵 $\Sigma_{X|Y}\Sigma_X^{-1}$的左（行）特征向量矩阵，即

$$V\Sigma_{X|Y}\Sigma_X^{-1} = \Lambda V \quad \text{with} \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots). \tag{15}$$

注意到 $V\Sigma_X^{1/2}$ 是正交的，因此 $V\Sigma_X V^\mathsf{T}$ 是对角矩阵 [22,引理 B.1], , 即

$$V\Sigma_X V^\mathsf{T} = R \quad \text{with} \quad R = \mathrm{diag}(r_1, r_2, \cdots). \tag{16}$$

将式（12-13）用W、 Λ 和 R 表示，可得

$$I_q(T;X) = \frac{-1}{2\bar{q}} \ln \frac{|I + WRW^\mathsf{T}|^q}{|I + (1-\bar{q}^2)WRW^\mathsf{T}|} \tag{17}$$

$$I_q(T;Y) = \frac{-1}{2\bar{q}} \ln \frac{|I + W\Lambda RW^\mathsf{T}|^{\bar{q}} |I + WRW^\mathsf{T}|^q}{|I + W[I - \bar{q}^2(I-\Lambda)]RW^\mathsf{T}|}. \tag{18}$$

将方程（17-18）代入方程（7），并令其关于混合矩阵W 的一阶导数为零，可得一阶条件

$$\begin{aligned} &\frac{q}{I + RW^\mathsf{T}W} - \frac{1-\bar{q}^2}{I + (1-\bar{q}^2)RW^\mathsf{T}W} \\ &= \beta\left(\frac{q}{I + RW^\mathsf{T}W} + \frac{\bar{q}}{I + \Lambda RW^\mathsf{T}W}\Lambda \right. \\ &\qquad \left. - \frac{1}{I + [I - \bar{q}^2(I-\Lambda)]RW^\mathsf{T}W}[I - \bar{q}^2(I-\Lambda)]\right). \end{aligned} \tag{19}$$

在推导上述结果时，我们使用恒等式 $d\ln|I+WCW^\mathsf{T}|/dW = 2W$ $I + CW^\mathsf{T}W^{-1}C$，该恒等式适用于任意相容的方阵 C，并假设 R 和 W 均可逆。

我们寻求如下形式的解

$$RW^\mathsf{T}W = \mathrm{diag}(r_1 w_1^2, r_2 w_2^2, \cdots) \equiv \mathrm{diag}(u_1, u_2, \cdots). \tag{20}$$

将此假设代入方程（19），得到

$$\frac{1}{\beta} = g_q(u_i, \lambda_i) = \frac{1-\lambda_i}{1+u_i\lambda_i} \frac{1 + \frac{\bar{q}(1+\bar{q})(1-\lambda_i)u_i}{1+(1-\bar{q}^2(1-\lambda_i))u_i}}{1 + \frac{\bar{q}(1+\bar{q})u_i}{1+(1-\bar{q}^2)u_i}}. \tag{21}$$

我们看到， $\Sigma_{X|Y}\Sigma_X^{-1}$的特征模态所产生的贡献彼此解耦，而每个模态对应的约化混合权重$u_i = r_i w_i^2$仅取决于该模态的特征值 λ_i、信息瓶颈权衡参数 β 以及 Rényi 信息的阶数 q。对于 $\lambda \in (0,1)$ 和 $q \in [0,2]$，函数$g_q(u,\lambda)$ 关于 u严格递减（当 $u \geq 0$时），并且在 $u \to \infty$ 时趋于零（见附录C）。因此，当 $1/\beta < g_q(0,\lambda_i)$ 时，方程（21）恰有一个正解 $u_i > 0$。也就是说，只有当 β 超过临界值时，特征值为 λ_i 的特征模态才会对 Rényi信息瓶颈编码器产生贡献

$$\beta_c^{(i)} = \frac{1}{g_q(0,\lambda_i)} = \frac{1}{1-\lambda_i}. \tag{22}$$

注意， $\beta_c^{(i)}$不依赖于 q。要得到 u_i，我们可以直接求解方程（21），或者使用等价三次方程的根的解析公式（省略特征模态下标）

$$0 = au^3 + bu^2 + cu + d, \tag{23}$$

其中系数由下式给出

$$\begin{aligned} a &= \lambda(1+\bar{q})(1-(1-\lambda)\bar{q}^2) \\ b &= \lambda(2+2\bar{q}+\lambda\bar{q}^2) + d(1+\bar{q})(1-\lambda\bar{q}-(1-\lambda)\bar{q}^2) \\ c &= \lambda(1+\bar{q}+\bar{q}^2) + d(2+(1-\lambda)\bar{q}-\bar{q}^2) \\ d &= 1 - \beta(1-\lambda). \end{aligned}$$

尽管上述计算无法唯一确定混合矩阵W 我们仍可通过取 $W =$对角$(w_1,\ w_2,\ \cdots)$来得到一个有效的信息瓶颈编码器，其中 $w_i = \sqrt{u_i/r_i}$因为 Rényi信息瓶颈损失仅依赖于W 的对角元素 $W^\mathsf{T}W$ 为此，将式（20）代入式（17-18）并写出

$$I_q(T;X) = \frac{-1}{2\bar{q}} \sum_i^{\beta>\beta_c^{(i)}} \ln \frac{(1+u_i)^q}{1+(1-\bar{q}^2)u_i} \tag{24}$$

$$I_q(T;Y) = \frac{-1}{2\bar{q}} \sum_i^{\beta>\beta_c^{(i)}} \ln \frac{(1+\lambda_i u_i)^{\bar{q}}(1+u_i)^q}{1+[1-\bar{q}^2(1-\lambda_i)]u_i}, \tag{25}$$

其中求和仅针对那些对信息瓶颈编码器有贡献的特征模态，即满足 $u_i > 0$。我们在图1中描绘了 Rényi 信息瓶颈的最优前沿。

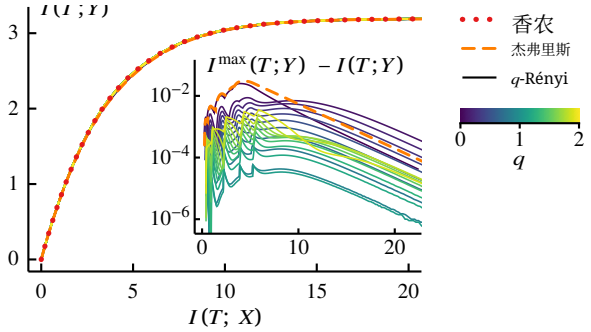


图 2. 求解 Rényi 和杰弗里斯信息瓶颈问题，可得到与香农信息瓶颈最优表征接近的表征。在香农信息平面上绘制时，香农（点线）、杰弗里斯（虚线）和 Rényi（实线）信息瓶颈问题的解几乎无法区分。对于 Rényi 情形，我们展示了不同 Rényi 阶数 q 范围内的结果（见颜色条）。插图：我们展示了可达到的最大相关香农信息量与编码后的相关香农信息量之间的差值，分别为 $I\max(T;Y)$ 和 $I(T;Y)$，并将其表示为提取的香农信息 $I(T;X)$ 的函数。在低信息量和高信息量极限下，该差值趋于零，即 $I(T;X)\to 0$ 和 $I(T;X)\to\infty$。卫星峰的出现，是因为尽管香农、杰弗里斯和 Rényi 信息瓶颈问题的解都在相同的一组临界权衡参数处发生结构转变，但这些转变发生时的 $I(T;X)$ 值各不相同。这里，$\Sigma_{X|Y}\Sigma_X^{-1}$的特征值为 $\lambda_i = 0.1$，0.2，0.3，0.5，0.7，0.8。

为完成对 Rényi 信息瓶颈的分析，我们指出，参考文献 [22] 中的解析解是我们结果的一个极限情形。当 $q\to 1$时，式（21）可写为

$$\frac{1}{\beta} = g_{q=1}(u_i,\lambda_i) = \frac{1-\lambda_i}{1+u_i\lambda_i} \implies u_i^{(q=1)} = \frac{\beta(1-\lambda_i)-1}{\lambda_i}. \tag{26}$$

回想起 $u_i = r_i w_i^2$，我们立即看到，该解与参考文献 [22, 引理 4.1]中的解相同。

V. 高斯变量的杰弗里斯信息瓶颈

上一节中的方法同样适用于基于其他统计散度的信息瓶颈问题。本节中，我们考虑杰弗里斯信息瓶颈，其损失函数定义为

$$\mathcal{L}_J[Q_{T|X}] = J(T;X) - \beta J(T;Y). \tag{27}$$

这里 $J(X;Y)$ 表示杰弗里斯信息，它是一种互依赖度量，定义为

$$J(X;Y) \equiv \mathrm{D_J}(P_{XY} \parallel P_X \otimes P_Y), \tag{28}$$

其中 $\mathrm{D_J}$ 表示杰弗里斯散度 [32],

$$\mathrm{D_J}(P \parallel Q) = \frac{1}{2}[\mathrm{KL}(P \parallel Q) + \mathrm{KL}(Q \parallel P)]. \tag{29}$$

对于高斯相关随机变量，杰弗里斯信息具有简单的形式（推导见附录B）

$$J(X;Y) = \frac{1}{2}\,\mathrm{tr}\left(\Sigma_X\Sigma_{X|Y}^{-1} - I\right). \tag{30}$$

使用式（9）中的线性编码器，式（27）中的信息项为

$$J(T;X) = \frac{1}{2}\,\mathrm{tr}\left(WRW^\mathsf{T}\right) \tag{31}$$

$$J(T;Y) = \frac{1}{2}\,\mathrm{tr}\left((I+WRW^\mathsf{T})\frac{1}{I+W\Lambda RW^\mathsf{T}} - I\right). \tag{32}$$

其中W Λ 和 R 在方程（14-16）中定义。

为求解信息瓶颈优化，我们对损失函数关于混合矩阵W 求导，并令所得导数为零，得到

$$I = \beta\frac{1}{I+\Lambda RW^\mathsf{T}W}\left(I - (I+RW^\mathsf{T}W)\frac{1}{I+\Lambda RW^\mathsf{T}W}\Lambda\right), \tag{33}$$

其中用到了以下恒等式 $\partial A^{-1}/\partial a = -A^{-1}(\partial A/\partial a)A^{-1}$ 以及 $\partial\,\mathrm{tr}(CABA^\mathsf{T})/\partial A = 2CAB$（适用于对称矩阵 B和 C）。可见，只要令 $W^\mathsf{T}W$ 为对角矩阵，该方程便可求解。将式（20）代入上述方程并求解 u_i，得到

$$u_i = \frac{\sqrt{\beta(1-\lambda_i)}-1}{\lambda_i}. \tag{34}$$

由于 $u_i = r_i w_i^2 \geq 0$，可见该解仅在权衡参数 β 大于临界值时有效

$$\beta_c^{(i)} = \frac{1}{1-\lambda_i}. \tag{35}$$

我们注意到，该临界值与 Rényi 信息瓶颈中的临界值相同[式（22）]。最后，将式（34）代入式（31-32），得到

$$J(T;X) = \frac{1}{2}\sum_i^{\beta>\beta_c^{(i)}}\frac{\sqrt{\beta(1-\lambda_i)}-1}{\lambda_i} \tag{36}$$

$$J(T;Y) = \frac{1}{2}\sum_i^{\beta>\beta_c^{(i)}}\frac{1-\lambda_i}{\lambda_i}\frac{\sqrt{\beta(1-\lambda_i)}-1}{\sqrt{\beta(1-\lambda_i)}}. \tag{37}$$

其中，求和仅限于对编码器有贡献的模态，即满足以下条件的模态 $\beta_c^{(i)} < \beta$。图 1展示了一个根据上述方程计算得到的杰弗里斯信息瓶颈最优前沿示例。我们强调，尽管香农信息等价于 $q = 1$时的 Rényi 信息，但 Jeffreys 信息并非 Rényi 信息的特例。

VI. 讨论与结论

在图 2中，我们在香农信息平面上绘出了原始、Rényi 和杰弗里斯信息瓶颈问题的解。可以看到，这些解都非常接近由香农信息瓶颈解所刻画的最优前沿。这一结果表明，对于采用香农信息以外的其他相关性度量定义的信息瓶颈问题进行建模和求解，或许能为原始信息瓶颈问题提供一种近似求解策略。为了更清楚地展示原始、Rényi 和杰弗里斯信息瓶颈问题解之间的差异，

插图显示，与香农信息瓶颈的最优表示相比，Rényi 和杰弗里斯信息瓶颈的最优表示所编码的相关香农信息少了多少。可以看到，在中等信息量处差异最大，而在低信息量和高信息量极限下，差异消失。此外，香农信息差距呈现出卫星峰，这是由信息瓶颈解的结构性转变造成的。我们注意到，尽管这些转变发生在相同的临界权衡参数 [方程（22）和（35）],但它们通常对应于不同的提取香农信息值。

总之，我们研究了广义信息瓶颈问题，其中互信息被基于 Rényi 和杰弗里斯散度的互依赖度量所取代。对于高斯相关随机变量，我们得到了精确的解析解，推广了文献 [22] 的结果。我们表明，信息瓶颈中相关性与压缩之间的基本权衡，同样适用于香农信息以外的相关性度量。我们的分析揭示了最优表示中的结构转变，这与原始信息瓶颈方法 [22] 中的情况类似。有趣的是，尽管原始、Rényi 和 Jeffreys 信息瓶颈问题的解各不相同，其临界权衡参数却相同。

我们预计，本研究将应用于相关组分物理学；该领域依赖熵和信息的 Rényi 推广来量化纠缠。此外，我们对 Jeffreys 信息瓶颈的刻画，可能有助于理解 Gibbs学习算法的泛化性质；这类算法的泛化差距与拟合模型和训练数据之间的 Jeffreys 信息成正比。最后，我们指出，对于条件信息瓶颈问题，如果将压缩项 $I(T;\,X)$ 替换为 $I(T;\,X\,|\,\mathcal{Y})$，那么由于 Rényi 信息和 Jeffreys 信息不满足链式法则，广义信息度量下的问题将变得非平凡。也就是说，在马尔可夫约束 T–X–$\mathcal{Y}$下，对于香农信息，有 $I(T;\,X\,|\,Y) = I(T;\,X) - I(T;\,\mathcal{Y})$；但一般而言，$\Omega(T;\,X\,|\,Y) \neq \Omega(T;\,X) - \Omega(T;\,\mathcal{Y})$。我们分析中的逻辑步骤很容易推广到条件信息瓶颈问题。

致谢

本研究部分经费由美国国家科学基金会通过生物功能物理中心（PHY-1734030）、西蒙斯基金会和斯隆基金会提供。

附录 A：高斯变量的 Rényi 信息

在本附录中，我们推导高斯相关变量的Rényi互信息。根据式（3-4）的定义，我们写出连续随机变量的Rényi互信息，

$$I_q(X;Y) = \frac{1}{q-1}\ln\int dxdy\, p_X(x)p_Y(y)\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right)^q. \tag{A1}$$

其中，p_X、p_Y 和 p_{XY} 分别表示 X、Y和（X，$\mathcal{Y}$）的概率密度函数。我们考虑高斯相关随机变量

$$\begin{bmatrix}X\\Y\end{bmatrix} \sim N(\mu,\Sigma) \;\text{ with }\; \mu=\begin{bmatrix}\mu_X\\\mu_Y\end{bmatrix} \;\text{ and }\; \Sigma=\begin{bmatrix}\Sigma_X & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_Y\end{bmatrix}. \tag{A2}$$

在这种情况下，联合概率密度为

$$p_{XY}(x,y) = \frac{\exp\left\{-([\,{}^{x}_{y}\,]-\mu)^{\mathsf{T}}\Sigma^{-1}([\,{}^{x}_{y}\,]-\mu)\right\}}{|2\pi\Sigma|^{1/2}} \tag{A3}$$

边缘分布的乘积等于一个将 Σ_{XY} 和 Σ_{YX} 设为零的联合分布，即

$$p_X(x)p_Y(y) = \frac{\exp\left\{-([\,{}^{x}_{y}\,]-\mu)^{\mathsf{T}}\bar{\Sigma}^{-1}([\,{}^{x}_{y}\,]-\mu)\right\}}{|2\pi\bar{\Sigma}|^{1/2}} \tag{A4}$$

其中 $\bar{\Sigma} = \left[\begin{smallmatrix}\Sigma_X & \\ & \Sigma_Y\end{smallmatrix}\right]$。将上述密度代入公式（A1），并对$x$ 和 y进行相应的高斯积分，得到

$$I_q(X;Y) = \frac{1}{q-1}\ln\frac{|q\Sigma^{-1}+(1-q)\bar{\Sigma}^{-1}|^{-1/2}}{|\Sigma|^{q/2}|\bar{\Sigma}|^{(1-q)/2}}. \tag{A5}$$

协方差矩阵的行列式为

$$|\Sigma| = |\Sigma_Y|\times|\Sigma_{X|Y}| \quad\text{and}\quad |\bar{\Sigma}| = |\Sigma_Y|\times|\Sigma_X|, \tag{A6}$$

其中 $\Sigma_{X|Y} = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}$，我们使用恒等式

$$\begin{vmatrix}A & B\\ C & D\end{vmatrix} = |D|\times|A-BD^{-1}C|. \tag{A7}$$

现在考虑式（A5）中的分子，

$$\begin{aligned}\left|q\Sigma^{-1}+(1-q)\bar{\Sigma}^{-1}\right| &= \left|\Sigma^{-1}\left(q\bar{\Sigma}+(1-q)\Sigma\right)\bar{\Sigma}^{-1}\right|\\ &= \frac{1}{|\Sigma|\times|\bar{\Sigma}|}\begin{vmatrix}\Sigma_X & (1-q)\Sigma_{XY}\\ (1-q)\Sigma_{YX} & \Sigma_Y\end{vmatrix}\\ &= \frac{\left|I-(1-q)^2(I-\Sigma_{X|Y}\Sigma_X^{-1})\right|}{|\Sigma_Y|\times|\Sigma_{X|Y}|},\end{aligned} \tag{A8}$$

其中，最后一个等号由式（A6-A7）得出。最后，我们写出高斯变量的 Rényi 信息

$$I_q(X;Y) = \frac{1/2}{q-1}\ln\frac{|\Sigma_{X|Y}\Sigma_X^{-1}|^{1-q}}{\left|I-(1-q)^2(I-\Sigma_{X|Y}\Sigma_X^{-1})\right|}. \tag{A9}$$

此表达式与式（5）（其中 $\bar{q}=1-q$）相同。

附录 B：高斯变量的 Jeffreys 信息

Jeffreys 信息定义为

$$J(X;Y) \equiv \mathrm{D_J}(P_{XY}\parallel P_X\otimes P_Y), \tag{B1}$$

其中 $\mathrm{D_J}$ 是 Jeffreys散度 [32],

$$\mathrm{D_J}(P\parallel Q) = \frac{1}{2}[\mathrm{KL}(P\parallel Q)+\mathrm{KL}(Q\parallel P)]. \tag{B2}$$

对于高斯相关的 X 和 Y，Jeffreys 信息可直接由两个多元高斯分布之间的 KL 散度得到

$$\mathrm{KL}(N(\mu_0,\Sigma_0)\,\|\,N(\mu_1,\Sigma_1)) = \frac{1}{2}\Big(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0 - I) + (\mu_1-\mu_0)^{\mathsf{T}}\Sigma_1^{-1}(\mu_1-\mu_0) + \ln\frac{|\Sigma_1|}{|\Sigma_0|}\Big). \quad \text{(B3)}$$

对于 X 和 Y 由式（A2）描述的情形，我们有 $P_{XY} = N(\mu,\ \Sigma)$ 和 $P_X \otimes P_Y = N(\mu,\ \bar{\Sigma})$，其中 $\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}$ 和 $\bar{\Sigma} = \begin{bmatrix} \Sigma_X & \cdot \\ \cdot & \Sigma_Y \end{bmatrix}$。因此，我们有

$$\mathrm{KL}(P_{XY}\,\|\,P_X\otimes P_Y) = \frac{1}{2}\left(\mathrm{tr}(\bar{\Sigma}^{-1}\Sigma - I) + \ln\frac{|\bar{\Sigma}|}{|\Sigma|}\right) \quad \text{(B4)}$$

$$\mathrm{KL}(P_X\otimes P_Y\,\|\,P_{XY}) = \frac{1}{2}\left(\mathrm{tr}(\Sigma^{-1}\bar{\Sigma} - I) + \ln\frac{|\Sigma|}{|\bar{\Sigma}|}\right). \quad \text{(B5)}$$

我们看到，对称化后，对数项消失了[式（B2）]。将 $\bar{\Sigma}^{-1} = \begin{bmatrix} \Sigma_X^{-1} & \cdot \\ \cdot & \Sigma_Y^{-1} \end{bmatrix}$ 以及式（A6）中的行列式公式代入式（B4），得到

$$\mathrm{KL}(P_{XY}\,\|\,P_X\otimes P_Y) = \frac{1}{2}\ln\frac{|\bar{\Sigma}|}{|\Sigma|} = -\frac{1}{2}\ln|\Sigma_{X|Y}\Sigma_X^{-1}| \quad \text{(B6)}$$

这正是通常的互信息，正如预期的那样。为了计算式（B5）中的迹，我们写出协方差矩阵的逆矩阵，

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{X|Y}^{-1} & -\Sigma_{X|Y}^{-1}\Sigma_{XY}\Sigma_Y^{-1} \\ -\Sigma_Y^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1} & \Sigma_{Y|X}^{-1} \end{pmatrix}. \quad \text{(B7)}$$

因此我们有

$$\begin{aligned}\mathrm{tr}(\Sigma^{-1}\bar{\Sigma} - I) &= \mathrm{tr}(\Sigma^{-1}(\bar{\Sigma}-\Sigma)) \\ &= \mathrm{tr}\left(\begin{bmatrix} \Sigma_{X|Y}^{-1} & -\Sigma_{X|Y}^{-1}\Sigma_{XY}\Sigma_Y^{-1} \\ -\Sigma_Y^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1} & \Sigma_{Y|X}^{-1} \end{bmatrix}\begin{bmatrix} \cdot & -\Sigma_{XY} \\ -\Sigma_{YX} & \cdot \end{bmatrix}\right) \\ &= \mathrm{tr}(\Sigma_{X|Y}^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}) + \mathrm{tr}(\Sigma_Y^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1}\Sigma_{XY}) \\ &= 2\,\mathrm{tr}(\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_{X|Y}^{-1}) \\ &= 2\,\mathrm{tr}(\Sigma_X\Sigma_{X|Y}^{-1} - I), \end{aligned} \quad \text{(B8)}$$

其中，最后一个等号由恒等式 $\Sigma_{X|Y} = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}$ 得出。将上述结果代入式（B5），得到

$$\mathrm{KL}(P_X\otimes P_Y\,\|\,P_{XY}) = \mathrm{tr}(\Sigma_X\Sigma_{X|Y}^{-1} - I) + \frac{1}{2}\ln|\Sigma_{X|Y}\Sigma_X^{-1}|. \quad \text{(B9)}$$

最后利用式（B6）消去对数项，得到

$$\begin{aligned} J(X;Y) &= \frac{1}{2}[\mathrm{KL}(P_X\otimes P_Y\,\|\,P_{XY}) + \mathrm{KL}(P_{XY}\,\|\,P_X\otimes P_Y)] \\ &= \frac{1}{2}\,\mathrm{tr}\left(\Sigma_X\Sigma_{X|Y}^{-1} - I\right). \end{aligned} \quad \text{(B10)}$$

附录C：补充图

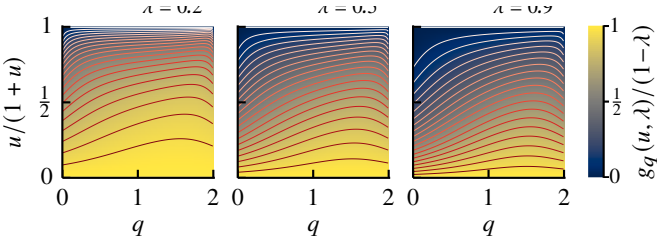


图 C1。函数 $g_q(u,\lambda)$ [式（21）]随 u 增大而递减，从 $u=0$ 处的 $1-\lambda$ 开始，并在 $u\to\infty$ 时趋近于零。因此，当 $\beta > 1/(1-\lambda)$ 时，方程 $\beta^{-1} = g_q(u,\lambda)$ 总有唯一的正解。我们只考虑 $0\le q\le 2$，因为高斯变量的 Rényi 信息在 $q>2$ [时可能发散，见式（5）]。

[1] N. Tishby、F. C. N. Pereira 和 W. Bialek，《信息瓶颈方法》，载于 第 *37* 届 *Allerton* 通信、控制与计算会议，B. Hajek 和 R. S. Sreenivas 编（伊利诺伊大学，1999 年），第 368–377 页。[2] A. Achille 和 S. Soatto，《信息丢弃：通过噪声计算学习最优表示》，IEEE 模式分析与机器智能汇刊 **40**，2897（2018 年）。[3] A. Achille 和 S. Soatto，《深度表征中不变性与解缠的涌现》，机器学习研究期刊 **19**，1（2018 年）。[4] Y. Tian、C. Sun、B. Poole、D. Krishnan、C. Schmid 和 P. Isola，《什么样的视图有利于对比学习？》，载于神经信息处理系统进展，第 33 卷，由 H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, 和 H. Lin（Curran Associates, Inc., 2020），第6827–6839页。[5] S. E. Palmer、O. Marre、M. J. Berry 和 W. Bialek，《感觉神经元群体中的预测信息》，美国国家科学院院刊 **112**，6908（2015）。[6] S. Wang、I. Segev、A. Borst 和 S. Palmer，《果蝇早期视觉系统中的最大效率预测或可支持规避性飞行动作》，《PLOS计算生物学》**17**，e1008965（2021）。[7] S. Still、D. A. Sivak、A. J. Bell 和 G. E. Crooks，《预测的热力学》，《物理评论快报》**109**，120604（2012）。

[8] A. Gordon、A. Banerjee、M. Koch-Janusz 和 Z. Ringel，《重整化群与信息论中的相关性》，《物理评论快报》126，240601（2021）。[9] A. G. Kline 和 S. E. Palmer，《高斯信息瓶颈与非微扰重整化群》，《新物理学杂志》24，033007（2022）。[10] D. Strouse 和 D. J. Schwab，《信息瓶颈与几何聚类》，《神经计算》31，596（2019）。[11] A. A. Alemi、I. Fischer、J. V. Dillon 和 K. Murphy，《深度变分信息瓶颈》，载于《国际学习表征会议》（2017）。[12] A. Goyal、R. Islam、D. Strouse、Z. Ahmed、H. Larochelle、M. Botvinick、S. Levine 和 Y. Bengio，《通过信息瓶颈实现迁移与探索》，载于《国际学习表征会议》（2019）。[13] W. Bialek、I. Nemenman 和 N. Tishby，《可预测性、复杂性与学习》，《神经计算》13，2409（2001）。[14] O. Shamir 、 S. Sabato 和 N. Tishby，《基于信息瓶颈的学习与泛化》，《理论计算机科学》411，2696（2010），算法学习理论（ALT 2008）。[15] W. Bialek、S. E. Palmer 和 D. J. Schwab，《是什么使我们能够在自然界中学习概率分布？》（2020），arXiv:2008.12279 [cond-mat.stat-mech]。[16] V. Ngampruetikorn 和 D. J. Schwab，《高维回归的信息瓶颈理论：相关性、效率与最优性》，载于《神经信息处理系统进展》第35卷，由 S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh 编（Curran Associates, Inc.，2022），第9784–9796页。[17] M. Chalk、O. Marre 和 G. Tkacik，《具有变分信息瓶颈的相关稀疏编码》，载于《神经信息处理系统进展》第29卷，由 D. D. Lee、M. Sugiyama、U. V. Luxburg、I. Guyon 和 R. Garnett 编（Curran Associates, Inc., 2016），第1957–1965页。[18] B. Poole、S. Ozair、A. Van Den Oord、A. Alemi 和 G. Tucker，《关于互信息的变分界》，载于《第36届国际机器学习会议论文集》，《机器学习研究论文集》第97卷，由 K. Chaudhuri 和 R. Salakhutdinov 编（PMLR，2019），第5171–5180页。[19] M. Tschannen、J. Djolonga、P. K. Rubenstein、S. Gelly 和 M. Lucic，《关于表征学习中的互信息最大化》，载于《国际学习表征会议》（2020）。[20] T. Wu、I. Fischer、I. L. Chuang 和 M. Tegmark，《信息瓶颈的可学习性》，熵21，924（2019）。[21] V. Ngampruetikorn 和 D. J. Schwab，《信息瓶颈的微扰理论》，载于《神经信息处理系统进展》第34卷，由 M. Ranzato、A. Beygelzimer 编，Y. Dauphin、P. Liang 和 J. W. Vaughan（Curran Associates, Inc., 2021）第21008–21018页。[22] G. Chechik、A. Globerson、N. Tishby 和 Y. Weiss，《高斯变量的信息瓶颈》，机器学习研究期刊 6，165（2005）。[23] A. Rényi，《论熵与信息的度量》，载于《第四届伯克利数学统计与概率研讨会论文集》第1卷，J. Neyman 编（1961），第547–561页。[24] R. Horodecki、P. Horodecki、M. Horodecki 和 K. Horodecki，《量子纠缠》，《现代物理评论》81，865（2009）。[25] J. Eisert、M. Cramer 和 M. B. Plenio，《专题综述：纠缠熵的面积定律》，《现代物理评论》82，277（2010）。[26] M. B. Hastings、I. González、A. B. Kallin 和 R. G. Melko，《在量子蒙特卡罗模拟中测量 Rényi纠缠熵》，《物理评论快报》104，157201（2010）。[27] R. R. P. Singh、M. B. Hastings、A. B. Kallin 和 R. G. Melko，《互信息的有限温度临界行为》，《物理评论快报》106，135701（2011）。[28] C. M. Herdman、P. N. Roy、R. G. Melko 和 A. D. Maestro，《超流体 ^{4}He中的纠缠面积定律》，《自然·物理》13，556（2017）。[29] R. Islam、R. Ma、P. M. Preiss、M. Eric Tai、A. Lukin、M. Rispoli 和 M. Greiner，《在量子多体系统中测量纠缠熵》，《自然》528，77（2015）。[30] A. Bergschneider、V. M. Klinkhamer、J. H. Becher、R. Klemt、L. Palm、G. Zürn、S. Jochim 和 P. M. Preiss，《通过位置和动量关联实验表征双粒子纠缠》，《自然·物理》15，640（2019）。[31] T. Brydges、A. Elben、P. Jurcevic、B. Vermersch、C. Maier、B. P. Lanyon、P. Zoller、R. Blatt 和 C. F. Roos，《通过随机测量探测 Rényi 纠缠熵》，《科学》364，260（2019）。[32] H. Jeffreys，《估计问题中先验概率的一种不变形式》，《伦敦皇家学会会刊：A辑，数学与物理科学》186，453（1946）。[33] G. Aminian、Y. Bu、L. Toni、M. Rodrigues 和 G. Wornell，《对Gibbs算法泛化误差的精确刻画》，载于《神经信息处理系统进展》第34卷，M. Ranzato、A. Beygelzimer、Y. Dauphin、P. Liang 和 J. W. Vaughan 编（Curran Associates, Inc., 2021），第8106–8118页。[34] T. van Erven 和 P. HarremoÃ«s，《Rényi散度与Kullback–Leibler散度》，《IEEE信息论汇刊》60，3797（2014）。[35] F. Liese 和 I. Vajda，《论统计学和信息论中的散度与信息量》，《IEEE信息论汇刊》52，4394（2006）。