arXiv:2601.01238v1 [stat.ML] 3 Jan 2026

# 奇异线性模型中的证据斜率与有效维数

Kalyaan Rao 纽约大学，数据科学中心
kmr9800@nyu.edu

贝叶斯模型选择通常依赖拉普拉斯近似或贝叶斯信息准则，它们假设有效模型维数等于参数个数 [1, 2]。奇异学习理论用实对数典范阈值（RLCT）取代这一假设；在过参数化或低秩模型中，实对数典范阈值这一有效维数可能严格小于参数个数 [3, 4]。我们研究了一类简单的线性高斯秩模型，其中实对数典范阈值可通过解析方法求得，精确边际似然也具有闭式形式。在这一设置下，我们从理论和实证两方面表明，拉普拉斯近似/贝叶斯信息准则的误差随 $(d/2 - \lambda) \log n$ 线性增长，而基于实对数典范阈值的校正能恢复正确的斜率，并且对于具有相同数据子空间的过完备重参数化保持不变。我们的结果对奇异模型中的“拉普拉斯近似失效”给出了具体的有限样本刻画，并提出了一种适用于教科书式线性设置的简单实对数典范阈值斜率估计量。

## 1. 引言

贝叶斯证据（边际似然）是许多模型选择程序的基础，而拉普拉斯近似或贝叶斯信息准则是实践中常用的标准近似 [1, 2, 5]。这些近似隐含地假设模型是正则的，即费舍尔信息满秩，且有效维数等于参数数量 $d$。

许多过参数化模型（低秩回归、字典学习、神经网络）都是奇异的：参数存在冗余，似然函数具有平坦方向，而通常的 $d/2$ 惩罚项并不正确。奇异学习理论（SLT）表明，在这类模型中，边际似然实际上由实对数典范阈值（RLCT）决定 $\lambda$，其中

$$\log p(D_n) = \log p(D_n \mid \theta^\star) - \lambda \log n + O(\log \log n),$$

其中 $\theta^\star$ 是使与真实数据生成分布之间的库尔贝克–莱布勒散度最小的任意参数 [3, 4]。因此，用 $d/2$ 替代 $\lambda$ 会导致 $\log p(D_n)$ 中的复杂度项设定错误，并可能使模型比较产生偏差。

我们在一个刻意选取的简单情形中重新审视这一差距：在线性高斯秩模型和带有高斯先验的线性子空间（字典）模型中，实对数典范阈值和精确证据都易于处理；这些模型与概率主成分分析和因子分析密切相关 [6, 7]。

具体而言，我们的贡献如下：

- 对于秩亏线性高斯回归，我们推导出其精确边际似然的闭式形式，并证明实对数典范阈值为 $\lambda(r) = r/2$，其中 $r$ 为内在秩。
- 我们证明，在这些模型中，拉普拉斯近似/贝叶斯信息准则近似的误差主导项为 $\big(d/2 - \lambda(r)\big) \log n = (d - r) \log n/2$；并通过实证表明，简单的基于实对数典范阈值的校正可以消除这种偏差。
- 在线性字典（子空间）模型中，我们证明，对于张成空间相同的过完备重参数化，实对数典范阈值和证据保持不变，而贝叶斯信息准则则不会；通过特征值谱和精确证据的实验，展示了这种表示不变性。

预印本。工作仍在进行中。

- 我们将证据相对于 $\log n$ 的斜率解释为这些教科书式线性场景中实对数典范阈值的实用有限样本估计量，从而提供了一种简单的方法，让我们能从数据中“看见”有效维数。

## 2. 相关工作

拉普拉斯近似与贝叶斯信息准则。拉普拉斯方法和贝叶斯信息准则（BIC）是用于近似正则参数模型边际似然的标准工具 [1, 2, 5]。在满足光滑性和可辨识性假设时，后验分布以速率 $n^{-1/2}$ 集中于某个唯一的最大似然估计或最大后验估计 $\theta_n$附近，且对数证据可展开为

$$\log p(D_n) = \log p(D_n \mid \hat{\theta}_n) - \frac{d}{2}\log n + O(1), \tag{1}$$

其中 $d$是参数维数。贝叶斯信息准则仅保留数据拟合项和 $d/2\log\ n$惩罚项，广泛用于回归、时间序列和潜变量模型中的模型选择。

奇异学习理论。许多现代模型不满足拉普拉斯近似/贝叶斯信息准则所依赖的正则性假设：混合模型、神经网络、低秩和过完备表示都存在冗余参数，且似然中有平坦方向。奇异学习理论以实对数典范阈值（RLCT） $\lambda$取代参数维数 $d$；实对数典范阈值是一种双有理不变量，用于量化库尔贝克–莱布勒极小化点 [3, 4]处似然的局部奇异性。在这类模型中，证据的表现为

$$\log p(D_n) = \log p(D_n \mid \theta^\star) - \lambda \log n + (m-1)\log\log n + O(1), \tag{2}$$

其中， $\theta^\star$ 是任意一个 KL 极小化点， $\lambda > 0$ 是实对数典范阈值， $m \in \mathbb{N}$ 是其重数。对于正则模型，$\lambda = d/2$ 和 $m = 1$可还原经典的拉普拉斯近似/贝叶斯信息准则展开；而在奇异模型中，通常有$\lambda < d/2$。某些混合模型和秩约束模型的解析实对数正则阈值已被推导出来 [3, 8, 9]，但除简单模型族外，闭式结果仍然罕见。

经验平坦性与有效维数。若干研究基于海森矩阵谱、费舍尔信息或 PAC-贝叶斯界，提出了模型“平坦性”或有效维数的经验替代指标 [10–13]。这些代理指标通常与过参数化网络中的泛化相关，但属于启发式指标，且并未直接建立在渐近边际似然的基础上。与奇异学习理论更为接近的是，一些近期研究尝试根据训练曲线或后验样本估计类似实对数典范阈值的指数，但通常无法获得真实实对数正则阈值或精确证据 [14, 15]。我们的工作有所不同：我们将研究限定在线性高斯族中，在该族中实对数典范阈值和精确边际似然均可解析求解，并将这些量直接与拉普拉斯近似/贝叶斯信息准则进行比较。

低秩与过参数化模型。低秩回归、因子分析和过完备字典是典型的模型示例：它们利用低维子空间进行表示，但所用参数可能很多 [6, 7, 16]。这类模型的模型选择通常依据参数的原始数量，采用 BIC 型惩罚项。然而，从奇异学习理论的角度看，有效复杂度应取决于内在秩（即子空间的维数），而非用于表示该子空间的坐标数量。我们研究的线性高斯模型提供了一种情境，可以明确区分这两者，并用边际似然对其进行量化。

## 3. 背景与问题设定

### 3.1. 贝叶斯证据与拉普拉斯近似

令 $D_n = \{(x_i,\ y_i)\}_{i=1}^n$ 为数据， $p(y \mid x,\ \theta)$ 为似然， $\pi(\theta)$ 为参数 $\theta \in \mathbb{R}^d$ 的先验分布。该模型的边际似然（证据）为

$$Z_n := p(D_n) = \int p(D_n \mid \theta)\,\pi(\theta)\,d\theta = \int \prod_{i=1}^{n} p(y_i \mid x_i, \theta)\,\pi(\theta)\,d\theta. \tag{3}$$

贝叶斯模型比较通常更倾向于选择 $Z_n$ 较大的模型。

ˆ 在正则模型中，当 $n \to \infty$ 时，后验分布会集中在对数后验的唯一极大值点 $\theta_n$ 附近。对 $\theta_n$ 进行二阶泰勒展开，可得拉普拉斯近似 [5]

$$\log Z_n \approx \log p(D_n \mid \hat{\theta}_n) + \log \pi(\hat{\theta}_n) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det H_n, \tag{4}$$

ˆ 其中 $H_n$ 是在 $\theta_n$ 处负对数后验的海森矩阵。在通常的正则性条件下，$H_n = nI(\theta^\star) + o_p(n)$，其中 $I(\theta^\star)$ 是 KL 极小化点 $\theta^\star$ 处的费舍尔信息，且 $\det H_n \asymp n^d$，因此

$$\log Z_n = \log p(D_n \mid \hat{\theta}_n) - \frac{d}{2}\log n + O_p(1). \tag{5}$$

贝叶斯信息准则舍弃 $\log \pi(\hat{\theta}_n)$ 和常数项，只保留这一 $d/2\log n$ 惩罚项 [1 ].

## 3.2. 奇异学习理论与实对数典范阈值

在奇异模型中，多个参数值可能表示同一个函数，费舍尔信息可能是秩亏的，而且后验分布无法用单个高斯分布很好地近似。奇异学习理论通过实对数典范阈值（RLCT）$\lambda$ 及其重数 $m$ [3, 4]，刻画这种情况下的证据。令 $\theta^\star$ 表示任意一个使模型与真实数据生成分布 $q(x,\ y)$ 之间的库尔贝克–莱布勒散度最小化的参数。在温和条件下，可得渐近展开

$$\log Z_n = \log p(D_n \mid \theta^\star) - \lambda \log n + (m-1)\log\log n + O(1). \tag{6}$$

这里，$\lambda > 0$ 是一个有理数，取决于 KL 极小化点集合周围似然的局部代数结构；$m \in \mathbb{N}$ 是其重数。在一个正则的 $d$ 维模型中，可得 $\lambda = d/2$ 和 $m = 1$，此时展开式简化为拉普拉斯近似/贝叶斯信息准则公式。在奇异模型中，通常有 $\lambda < d/2$，因此用 $d/2$ 替换 $\lambda$ 会使模型受到过度惩罚，在对数证据中多出 $(d/2 - \lambda)\log n$。

在这项工作中，我们不使用代数几何的全部工具。相反，我们利用这样一个事实：对于线性高斯模型，可以通过初等线性代数求得实对数典范阈值和精确证据，并将这些量直接与其拉普拉斯/BIC 对应项进行比较。

## 3.3. 线性高斯秩模型与字典

现在，我们介绍所研究的具体模型并确定符号记法。

秩-$r$ 线性回归。设 $x_i \in \mathbb{R}^p$、$y_i \in \mathbb{R}$，并考虑模型

$$y_i = x_i^\top B\theta + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{7}$$

其中 $B \in \mathbb{R}^{p\times d}$、$\theta \in \mathbb{R}^d$ 和 $\sigma^2 > 0$ 已知。我们将行向量为 $x_i^\top$ 的设计矩阵记为 $X_n \in \mathbb{R}^{n\times p}$，并将 $A_n := X_n B \in \mathbb{R}^{n\times d}$ 记为有效设计矩阵。给定 $\theta$，似然为

$$p(D_n \mid \theta) = \mathcal{N}\big(y;\, A_n\theta,\ \sigma^2 I_n\big), \tag{8}$$

其中 $y = (y_1、\ldots\ldots、\ y_n)^\top$。我们设高斯先验分布为 $\theta \sim \mathcal{N}(0,\ \tau^2 I_d)$，其中 $\tau^2 > 0$。如果 $\mathrm{rank}(B) = d$ 且 $X_n$ 满列秩，那么这是一个正则的 $d$ 维回归模型。如果 $\mathrm{rank}(B) = r < d$，那么 $A_n$ 的秩至多为 $r$，且 $\theta$ 中有 $d - r$ 个方向不会影响似然；该模型在环境坐标中是奇异的。

线性子空间（字典）模型。我们还考虑一种简单的线性高斯潜因子模型，并采用表征学习 [6, 7, 16] 中使用的“字典”这一术语来表述。令 $y_i \in \mathbb{R}^p$ 为观测值，令 $D \in \mathbb{R}^{p\times d}$ 为一个矩阵，其列 $d_1, \ldots, d_d$ 张成一个子空间 $V \subseteq \mathbb{R}^p$。我们生成

$$z_i \sim \mathcal{N}(0, \tau^2 I_d), \qquad y_i = Dz_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_p). \tag{9}$$

因此，每个 $y_i$ 都位于线性子空间 $V =$ 张成空间$\{d_1, \ldots, d_d\}$ 附近。我们强调，我们不对 $z_i$ 施加稀疏性约束；这与概率主成分分析或因子分析中的线性高斯模型相同，其中 $D$ 起到该子空间“字典”的作用。

该表示的内在维度为 $r := \dim V \leq d$。如果 $r < d$，该字典就是过完备的：列数多于子空间维度，许多不同的系数向量 $z_i$ 会得到相同的 $Dz_i$。我们关注的是比较共享相同张成空间 $V$ 的“最小”与“过完备”参数化（因此它们在 $y_i$上的分布族相同），并理解在贝叶斯信息准则与基于实对数典范阈值的视角下，它们的证据会如何变化。

精确与近似证据。在这两种设置中，高斯分布先验和似然意味着边际似然 $Z_n$ 具有闭式形式，可通过解析地积分掉 $\theta$ （或 $z_i$）得到。我们将这个闭式值视为精确证据。我们将其与以下结果进行比较：(i) 一种朴素的拉普拉斯近似/贝叶斯信息准则近似，使用环境参数维数 $d$；(ii) 一种 RLCT 感知近似，以第 $\lambda$ 节推导的解析 RLCT$r$ 替代 $d/2$ 。

# 4. 秩-$r$ 线性回归中的证据斜率

本节将推导秩-$r$ 回归模型中边际似然的闭式表达式，确定相应的实对数典范阈值，并说明拉普拉斯近似/贝叶斯信息准则的近似误差与秩亏 $d - r$之间的关系。

## 4.1. 闭式边际似然

对于固定的设计矩阵 $A_n \in \mathbb{R}^{n\times d}$和高斯先验 $\theta \sim \mathcal{N}(0, \tau^2 I_d)$，线性高斯模型具有闭式边际似然。

命题1（线性高斯回归中的精确证据）。设 $y \in \mathbb{R}^n$ 和 $A_n \in \mathbb{R}^{n\times d}$ 为固定值，并考虑

$$y \mid \theta \sim \mathcal{N}(A_n\theta, \sigma^2 I_n), \qquad \theta \sim \mathcal{N}(0, \tau^2 I_d).$$

则边际似然 $Z_n = p(y \mid A_n)$ 为

$$\log Z_n = -\frac{1}{2}\Big(n\log(2\pi) + n\log\sigma^2 + \log\det(I_d + \alpha S_n) + \sigma^{-2}\big(y^\top y - \alpha\, y^\top A_n(I_d + \alpha S_n)^{-1}A_n^\top y\big)\Big),$$

其中 $S_n := A_n^\top A_n$ 且 $\alpha := \tau^2/\sigma^2$。

证明。 根据贝叶斯定理， $\theta$ 的后验分布为高斯分布，其精度矩阵为

$$\Lambda_n = \sigma^{-2}A_n^\top A_n + \tau^{-2}I_d = \sigma^{-2}(S_n + \alpha^{-1}I_d),$$

均值为 $\mu_n = \sigma^{-2}\Lambda_n^{-1}A_n^\top y$。写出联合密度 $p(y,\ \theta)$，并对 $\theta$ 积分，相当于对 $\theta$ 配方。利用标准高斯恒等式可得

$$p(y) = \int p(y \mid \theta)\,\pi(\theta)\,d\theta = (2\pi)^{-n/2}\sigma^{-n}\det(I_d+\alpha S_n)^{-1/2}\exp\Big(-\tfrac{1}{2\sigma^2}\big(y^\top y - \alpha\, y^\top A_n(I_d+\alpha S_n)^{-1}A_n^\top y\big)\Big),$$

对其取对数即可得到所述表达式。 □

特别地，证据仅通过格拉姆矩阵 $S_n = A_n^\top A_n$ 依赖于 $A_n$ ，其特征值揭示了该模型的有效秩。

## 4.2. 实对数典范阈值与有效秩

现在，我们专门讨论上述秩为$r$ 的情形，其中 $A_n = X_nB$ 的秩为$B$，且 $= r \leq d$和 $x_i$ 是具有非退化协方差 $\Sigma_x$的独立同分布变量。

令 $\Sigma_A := \mathbb{E}[xx^\top]BB^\top = \Sigma_x BB^\top$。在对 $x_i$ 的温和条件下（例如次高斯），经验格拉姆矩阵满足

$$\frac{1}{n}S_n = \frac{1}{n}A_n^\top A_n = \frac{1}{n}B^\top X_n^\top X_n B \to B^\top\Sigma_x B$$

几乎 必然如 $n \to \infty$。因此， $S_n$ 具有 $r$ 个阶为 $n$ 的特征值，以及 $d - r$ 个阶为 1.

命题 2(秩为 -实对数典范阈值$r$ 回归)。 在秩为 $r$ 的线性回归模型中，采用高斯先验分布和噪声时，实对数典范阈值为

$$\lambda(r) = \frac{r}{2}.$$

证明概要。 在 KL 极小化点 $\theta^\star$ 处的费舍尔信息与 $B^\top \Sigma_x B$成正比，其秩为 $r$。局部而言，对数似然在 $\theta^\star$ 附近表现为在 $r$个方向上的非退化二次型，在其余 $d-r$ 个方向上则是平坦的。在奇异学习理论中，这种“二次型乘以平坦项”结构的实对数典范阈值为 $r/2$ [3, 4]，反映了 $r$ 个弯曲方向各自对指数贡献 $1/2$ 。在我们的模型中，可以直接从命题1中的对数 $Z_n$ 精确表达式看出这一点：对 $S_n$ 进行谱分解并以 $n$归一化后，对数 $Z_n$ 中的主导 $n \to \infty$ 项包含$-\frac{1}{2}\sum_{j=1}^{r} \log n = -\frac{r}{2}\log\, n$，而其余 $d-r$ 个方向仅贡献 $O(1)$ 项。 □

4.3. 拉普拉斯近似/贝叶斯信息准则误差与证据斜率

现在，我们将精确证据与拉普拉斯/BIC近似进行比较。对于采用平坦（或弱信息）先验分布的高斯分布模型，在最大似然估计量处应用拉普拉斯方法可得

$$\log Z_n^{\mathrm{Lap}} = \log p(D_n \mid \hat{\theta}_n) - \frac{d}{2}\log n + O_p(1).$$

结合奇异学习理论的展开式，可得到近似误差在大-$n$时的行为。

命题 3 （秩为$r$ 的回归中拉普拉斯近似/贝叶斯信息准则的误差）。 在秩为-$r$ 的线性回归模型中，在上述条件下，

$$\log Z_n^{Lap} - \log Z_n = \Big(\frac{d}{2} - \lambda(r)\Big)\log n + O_p(1) = \frac{d-r}{2}\,\log n + O_p(1).$$

证明。 根据奇异学习理论的展开式，我们有

$$\log Z_n = \log p(D_n \mid \theta^\star) - \lambda(r)\log n + O_p(1),$$

而拉普拉斯近似/贝叶斯信息准则给出

$$\log Z_n^{\mathrm{Lap}} = \log p(D_n \mid \hat{\theta}_n) - \frac{d}{2}\log n + O_p(1).$$

ˆ 在设定正确的高斯情形下，最大似然估计量 $\theta_n$ 收敛到 KL ˆ 极小化者 $\theta^\star$，且差值 $\log p($ $D_n \mid \theta_n) - \log\ p(D_n \mid \theta^\star)$ 为 $O_p(1)$。将两个展开式相减，便得到

$$\log Z_n^{\mathrm{Lap}} - \log Z_n = \Big(\frac{d}{2} - \lambda(r)\Big)\log n + O_p(1).$$

代入 $\lambda(r) = r/2$ ，即可得到所述形式。 □

命题 3 意味着，在奇异模型（$r < d$）中，拉普拉斯近似/贝叶斯信息准则误差随 $\log n$线性增长，斜率为（$d-r$）$/2$；而在正则模型（$r = d$）中，斜率为零，仅存在 $O_p(1)$ 级别的波动。

ˆ 受此启发，我们定义一个实对数典范阈值的经验证据斜率估计量：给定不同样本量 $n \in \mathcal{N}$下 $\log$ $Z_n$的值 $\ell_n$，我们将 ˆ $\ell_n$对 $\log\ n$ 进行回归，并令

$$\hat{\lambda}_{\mathrm{emp}} := -\frac{1}{2}\,\widehat{\mathrm{slope}}\big(\hat{\ell}_n \text{ vs } \log n\big).$$

在我们的秩为 $r$ 的高斯分布族中，命题 2 保证，当 $|\mathcal{N}| \to \infty$ 且 $\min_{n\in\mathcal{N}} n \to\infty$ 时，ˆ $\lambda_{\mathrm{emp}} \to \lambda(r)$，误差为通常的回归误差。

# 5. 线性字典中的表示不变性

现在，我们将字典示例形式化：不同的参数化共享同一个子空间，但列数不同。

### 5.1. 子空间协方差与精确证据

回顾字典模型

$$y_i = Dz_i + \varepsilon_i, \qquad z_i \sim \mathcal{N}(0, \tau^2 I_d), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_p).$$

对 $z_i$ 积分可知， $y_i$ 的边际分布为高斯分布，其协方差为

$$\Sigma_y(D) = \tau^2 DD^\top + \sigma^2 I_p.$$

因此， $D_n = \{y_i\}_{i=1}^n$ 的联合边际似然为

$$\log p(D_n \mid D) = -\frac{n}{2} \log \det(2\pi \Sigma_y(D)) - \frac{1}{2} \sum_{i=1}^n y_i^\top \Sigma_y(D)^{-1} y_i.$$

假设 $D$ 和 $D'$ 具有相同的列空间 $V \subset \mathbb{R}^p$ ，且 $\dim V = r$。那么存在一个可逆矩阵 $R$ ，使得 $D' = DR$。因此， $D'D'^\top = DRR^\top D^\top$ 与 $DD^\top$具有相同的值域和秩，并且存在 $Q \succ 0$ ，属于该 $r$ 维子空间，满足 $D'D'^\top = DQD^\top$。这意味着， $\Sigma_y(D)$ 和 $\Sigma_y(D')$ 的谱在 $r$ 个信号方向上相同，仅相差不随 $n$缩放的 $O(1)$ 因子。

命题 4 (实对数典范阈值与证据的表示不变性)。 考虑两个字典 $D \in \mathbb{R}^{p \times d}$ 和 $D' \in \mathbb{R}^{p \times d'}$ ，其列空间相同且维数为 $r$。那么它们的实对数典范阈值相同，且等于$\lambda(r)$，其对数证据满足

$$\log p(D_n \mid D) = \log p(D_n \mid D') + O_p(1)$$

与 $n \to \infty$ 相同。特别是，对于同一子空间的所有参数化，首要的 $-\lambda(r) \log n$ 项都相同。

证明概要。 这两个模型在 $y$ 上诱导出相同的高斯分布族，其信号协方差的秩为 $r$ ；参数化只改变了将该协方差写作 $DD^\top$ 还是 $D'D'^\top$ 的方式。因此，KL 极小化点附近的似然局部几何，在 $r$ 信号方向上经平滑重新参数化后等价，而冗余坐标则对应平坦方向。奇异学习理论不变性结果表明，实对数典范阈值仅取决于内在模型类，而不取决于特定坐标系，因此得到 $\lambda(D) = \lambda(D') = \lambda(r)$ [3, 4]。对数证据的 $O_p(1)$ 差异源于：在两种参数化下， $y$ 的经验协方差都收敛到相同的总体极限，因此依赖于 $n$ 的首要项一致。

□

相比之下，BIC式近似在惩罚项中使用环境维数。

注1（过完备字典中BIC的不具不变性）。如果我们对 $D$ 和 $D'$ 的边际似然进行如下近似

$$\log Z_n^{BIC}(D) \approx \log p(D_n \mid \hat{D}) - \frac{d}{2} \log n, \qquad \log Z_n^{BIC}(D') \approx \log p(D_n \mid \hat{D}') - \frac{d'}{2} \log n,$$

那么，即使 $D$ 和 $D'$ 表示同一个子空间，且数据拟合项相近，*BIC*分数仍会相差约$\frac{1}{2}(d - d') \log n$。因此，*BIC*并非表示不变：即使不同参数化描述的是同一组分布，它也会人为地偏好坐标较少的参数化。

## 6. 实验

现在，我们通过合成线性高斯模型对上述论断进行数值演示。所有实验均采用高斯先验和高斯噪声，因此精确边际似然具有闭式形式。我们报告：（i）拉普拉斯近似/贝叶斯信息准则误差$\log Z_n^{\mathrm{Lap}} - \log Z_n$ 相对于$\log n$的斜率；以及（ii）经RLCT校正的误差$\log Z_n^{\mathrm{RLCT}} - \log Z_n$的斜率，其中

$$\log Z_n^{\mathrm{RLCT}} := \log p(D_n \mid \hat{\theta}_n) - \lambda(r) \log n$$

使用命题 2 中的解析 RLCT $\lambda(r)$。在所有情况下，我们都对设计和噪声的多次随机抽样取平均。

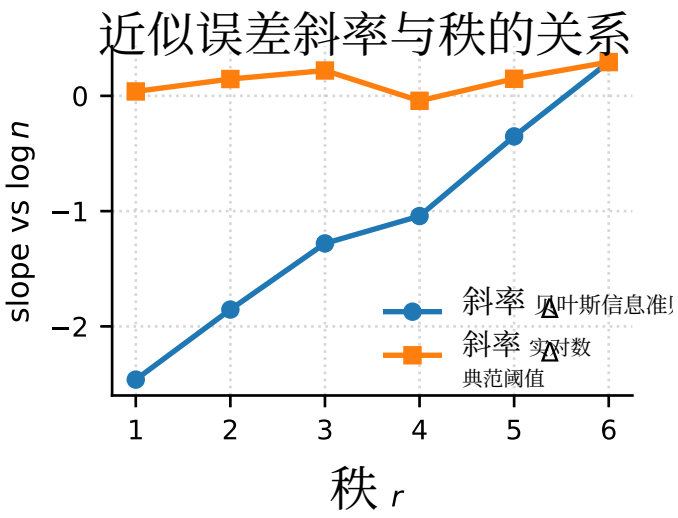


图 1：线性回归中的秩扫描。对于每个内在秩 $r$，我们估计贝叶斯信息准则误差 $\Delta_{\mathrm{BIC}}(n)$ 和经 RLCT 校正的误差 $\Delta_{\mathrm{RLCT}}(n)$ 随 $\log n$ 变化的斜率。当 $r$ 接近环境维数 $d$ 时，贝叶斯信息准则斜率趋近于零。对于较小的 $r$，贝叶斯信息准则斜率显著为负，而经 RLCT 校正的斜率在所有 $r$ 范围内都接近于零。

## 6.1. 实验设置

对于秩-$r$ 回归实验，我们生成输入 $x_i \sim \mathcal{N}(0,\ I_p)$，其中 $p$ 固定；选择一个秩为 $= r$ 的真实矩阵 $B^\star \in \mathbb{R}^{p\times d}$，满足 rank($B^\star$)，并抽取一个真实参数 $\theta^\star \sim \mathcal{N}(0,\ \tau^2 I_d)$。然后我们设定

$$y_i = x_i^\top B^\star \theta^\star + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

对于样本量 $n \in \{50、100、200、400、.....\}$。对于每个 $n$，我们计算：

- 使用命题1计算精确对数证据 $\log Z_n$ ；
- 拉普拉斯近似/贝叶斯信息准则近似 $\log Z_{\mathrm{Lap}n} \quad = \log p(D_n \mid \hat{\theta}_n) - \frac{d}{2}\log n$;
- RLCT感知近似 $\log\ Z_{\mathrm{RLCT}n} \qquad = \log\ p(D_n \mid \hat{\theta}_n)\ - \lambda(r)\log\ n$，其中 $\lambda(r)\ = r/2$。

我们对多个随机种子重复此过程，并对每个误差项 $\Delta_{\mathrm{BIC}}(n) := \log Z_n^{\mathrm{Lap}} - \log Z_n$ 和 $\Delta_{\mathrm{RLCT}}(n) := \log Z_n^{\mathrm{RLCT}} - \log Z_n$ 关于 $\log n$ 进行简单线性回归，以估计其斜率。

对于字典实验，我们使用第3.3节的线性子空间模型，固定一个 $p$ 维、秩为 $r$ 的子空间 $V$，并构造：(i) 一个“最小”字典 $D \in \mathbb{R}^{p\times r}$，其列构成 $V$ 的一组基；(ii) 一个“过完备”字典 $D' \in \mathbb{R}^{p\times d'}$ $(d' > r)$，其列张成相同的$V$。我们使用固定真实字典从模型生成 $y_i$，并比较 $D$ 和 $D'$的精确证据与BIC分数。

## 6.2. 奇异回归中的秩扫描

我们的第一个实验改变内在秩 $r$，同时保持环境维数 $d$ 不变。对于每个$r \in \{1,\ 2、\ \ldots、6\}$，我们构造一个秩为$r$ 的矩阵 $B^\star$，并对$n$的若干取值运行上述流程。图 1 展示了 $\Delta_{\mathrm{BIC}}(n)$和 $\Delta_{\mathrm{RLCT}}(n)$的估计斜率随 $r$变化的情况。

表 1 总结了一次具有 $d = 6$和 $r \in \{1、.....、6\}$ 的代表性运行结果。BIC 斜率在 $r = 1$ 处的强负值与 $r = d$ 处的近似零值之间插值，而经 RLCT 校正的斜率始终接近零，这与命题 3 一致。

| 秩 $r$ | 斜率 $\Delta_{\mathrm{BIC}}$ | 斜率 $\Delta_{\mathrm{RLCT}}$ |
|---|---|---|
| 1 | -2.50 | 0.01 |
| 2 | -1.97 | 0.03 |
| 3 | -1.49 | 0.01 |
| 4 | -0.98 | 0.03 |
| 5 | -0.44 | 0.06 |
| 6 | 0.07 | 0.07 |

表 1：秩为 $r$ 的回归中，$\Delta_{\mathrm{BIC}}(n)$ 和 $\Delta_{\mathrm{RLCT}}(n)$ 相对于 $\log n$ 的估计斜率（一个代表性配置，$d = 6$）。在强奇异情形（$r \ll d$）下，BIC误差会衰减（斜率为负），这反映了 $(d - r)/2\log n$ 造成的过度惩罚；而经RLCT校正的误差在所有 $r$ 下的斜率都接近于零。

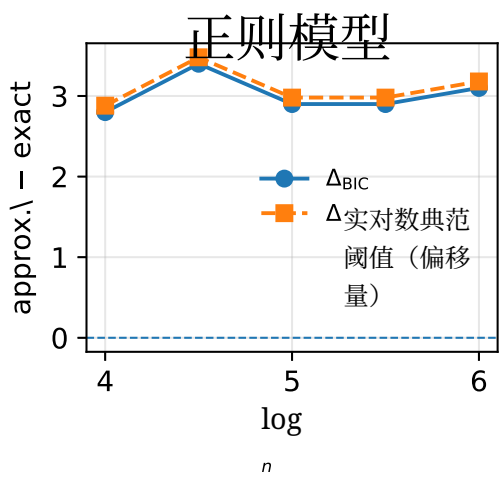


图 2：正则模型：$\Delta_{\mathrm{BIC}}(n)$ 和 $\Delta_{\mathrm{RLCT}}(n)$ 相对于 $\log n$ 的变化。当 $\lambda = d/2$ 时，正如预测的那样，两条斜率都接近于零。

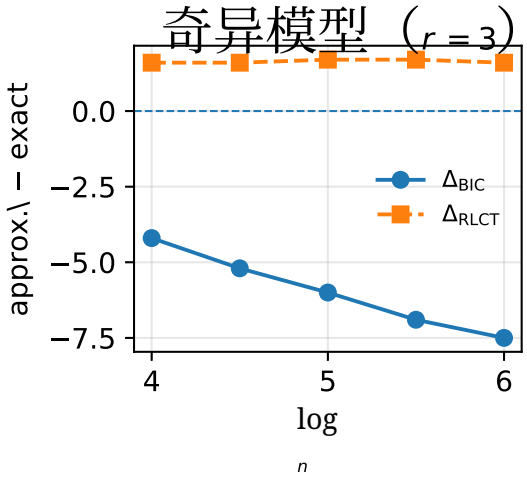


图 3：奇异模型：$\Delta_{\mathrm{BIC}}(n)$ 和 $\Delta_{\mathrm{RLCT}}(n)$ 相对于 $\log n$ 的变化。BIC误差的斜率大幅为负，而经RLCT校正的误差则几乎保持平坦。

## 6.3. 正则与奇异情形下的证据校正

为明确呈现正则与奇异对比，我们还使用相同的数据生成过程，对比了一个完全正则模型（$r = d$）和一个奇异模型（$r < d$）。对于每个 $n$，我们计算近似对数证据与精确对数证据之间的差值，并再次对其与 $\log n$ 进行回归。

图 2 和图 3 展示了这一现象：在正则情形下，贝叶斯信息准则和考虑实对数典范阈值的近似都与精确证据密切吻合，且 $\Delta_{\mathrm{BIC}}(n)$ 的斜率在统计上与零无异。在奇异情形下，贝叶斯信息准则误差呈现明显的负斜率线性趋势，而经RLCT校正的误差则保持平坦。

例如，一组具有代表性的拟合结果为：

$$\text{regular:} \quad \mathrm{slope}(\Delta_{\mathrm{BIC}}) \approx 0.20, \quad \mathrm{slope}(\Delta_{\mathrm{RLCT}}) \approx 0.20,$$

$$\text{singular:} \quad \mathrm{slope}(\Delta_{\mathrm{BIC}}) \approx -0.99, \quad \mathrm{slope}(\Delta_{\mathrm{RLCT}}) \approx 0.01,$$

这与命题 3 的 $\frac{d-r}{2}\log n$ 预测一致。

## 6.4. 字典参数化与特征值谱

最后，我们转向字典模型，比较同一子空间的“最小”表示和“过完备”表示。图 4 显示了最小字典的 $D^\top D$ 和 $D'^\top D'$ 的特征值

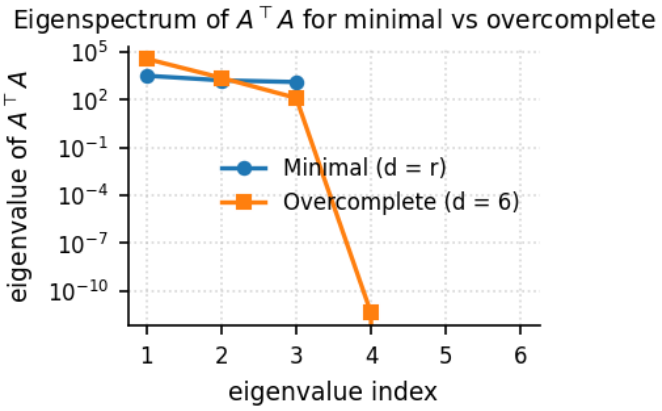


图 4：对于相同的底层子空间， $D^\top D$ （最小字典， $d = r$）和 $D'^\top D'$ （过完备字典，$d' > r$）的特征值谱。二者均具有 $r$ 个较大的特征值以及一组近零特征值，从而说明了过完备表示中的冗余坐标。

| 数量 | 值 |
|---|---|
| 最小字典（$d = r = 3$）：$\log Z_{\text{exact}}$ | $-297.53$ |
| 过完备字典（$d' = 6$）：$\log Z_{\text{exact}}$ | $-297.76$ |
| 最小 BIC 近似 | $-293.80$ |
| 过完备 BIC 近似 | $-301.75$ |
| 最简 RLCT 感知近似 | $-293.80$ |
| 过完备 RLCT 感知近似 | $-293.80$ |

表 2：表示同一子空间的最小字典和过完备字典的精确对数证据与近似对数证据。精确证据仅相差 $O(1)$，这与命题 4 一致。RLCT感知近似在惩罚项中使用内在秩 $r$ ，因此不受列数影响。BIC 近似按 $d/2\log n$进行惩罚，因此对过完备参数化给出的评分显著更差。

$D \in \mathbb{R}^{p\times r}$ 以及一个列空间相同的过完备字典 $D' \in \mathbb{R}^{p\times d'}$ 。两者均显示$r$ 个非零特征值，后面跟着一组（近似）为零的特征值，对应于冗余方向。

表 2 报告了在固定样本量 $n$ 下，这一类模型对中的一组模型的精确对数证据，以及 BIC/RLCT 类型的近似值。
这些实验支持我们的两项主要论点：（i）在秩亏线性模型中，拉普拉斯近似/贝叶斯信息准则的误差以 $(d/2 - \lambda)\log n$ 的斜率增长，而基于实对数典范阈值的校正可恢复正确的斜率；（ii）有效复杂度取决于内在子空间，而非用于表示它的坐标数量，因此基于 RLCT 的惩罚项具有表示不变性，而 BIC 则不具有。

# 7. 结论与未来工作

我们采用简单的线性高斯模型，明确展示了拉普拉斯近似/贝叶斯信息准则在奇异情形下的失效模式：当内在秩 $r$ 小于环境参数维数 $d$时，拉普拉斯近似/贝叶斯信息准则的误差会以斜率 $(d - r)\log n/2$ 增长，而基于实对数典范阈值的校正则能恢复正确的证据斜率。在一个线性字典模型中，我们进一步证明，采用具有相同张成空间的过完备重新参数化时，实对数典范阈值和证据保持不变，而贝叶斯信息准则则不会。

尽管我们的分析仅限于教科书式高斯族，但我们强调的现象——费希尔几何中的退化、 $d/2$ 与实对数典范阈值不匹配，以及表示不

变性贝叶斯信息准则——在过参数化模型中普遍存在。接下来可以立即着手将基于斜率的实对数典范阈值估计方法拓展到线性模型之外，并探究能否在深度神经网络等更复杂的架构中恢复（或有意强制实现）类似的表示不变性。

## 参考文献


[1] Gideon Schwarz。模型维数的估计。《统计年刊》，6(2):461–464，1978年。[2] Robert E. Kass 和 Adrian E. Raftery。贝叶斯因子。《美国统计学会会刊》，90(430):773–795，1995年。[3] Sumio Watanabe。《代数几何与统计学习理论》。剑桥大学出版社，2009年。[4] Sumio Watanabe。《贝叶斯统计学的数学理论》。Chapman and Hall/CRC，2018年。[5] Luke Tierney 和 Joseph B. Kadane。后验矩和边缘密度的精确近似。《美国统计学会会刊》，81(393):82–86，1986年。[6] Michael E. Tipping 和 Christopher M. Bishop。概率主成分分析。《皇家统计学会会刊：*B*辑（统计方法学）》，61(3):611–622，1999年。[7] Zoubin Ghahramani 和 Geoffrey E. Hinton。因子分析器的期望最大化算法。载于 Michael I. Jordan（编），《图模型学习》，第177–199页。Kluwer Academic Publishers，1998年。最初以技术报告 CRG-TR-96-1 的形式发表。[8] Sumio Watanabe。不可识别学习机器的代数分析。《神经网络》，13 (4–5):445–455，2000年。在奇异学习理论文献中常被引用为早期的实对数典范阈值结果。[9] Masayuki Aoyagi 和 Sumio Watanabe。奇异统计模型中贝叶斯估计量和吉布斯估计量的渐近学习曲线。《机器学习研究期刊》，11:1–47，2010年。[10] Sepp Hochreiter 和 Jürgen Schmidhuber。平坦极小值。《神经计算》，9(1):1–42，1997年。[11] Nitish Shirish Keskar、Dheevatsa Mudigere、Jorge Nocedal、Mikhail Smelyanskiy 和 Ping Tak Peter Tang。关于深度学习的大批量训练：泛化差距与尖锐极小值。载于国际学习表征会议（国际学习表征会议），2017年。[12] Tengyuan Liang、Alexander Rakhlin 和 Karthik Sridharan。神经网络的费舍尔–饶度量、几何与复杂度。载于第*36*届国际机器学习会议论文集（国际机器学习会议），2019年。探讨费舍尔信息和有效维数。[13] Gintare Karolina Dziugaite 和 Daniel M. Roy。计算非空泛的泛化界：针对参数远多于训练数据的深度（随机）神经网络。载于第*33*届人工智能中的不确定性会议论文集（人工智能中的不确定性），2017年。[14] Tomoyuki Matsuda 和 Sumio Watanabe。奇异统计模型的信息准则。《机器学习研究期刊》，20(75):1–52，2019年。[15] Jorma Rissanen。《统计建模中的信息与复杂性》。施普林格，2007年。关于最小描述长度和随机复杂度的奠基性参考文献。[16] Bruno A. Olshausen 和 David J. Field。使用超完备基组的稀疏编码：一种初级视觉皮层理论。《视觉研究》，37(23):3311–3325，1997年。