arXiv:2302.02950v1 [stat.ME] 6 Feb 2023

基于评分规则的多维客观先验分布。

Isadora Antoniano-Villalobos[1],Cristiano Villa[2],和 Stephen G. Walker[3]

[1] *Ca' Foscari University of Venice, Italy*
[2] *Duke-Kunshan University, Suzhou, China*
[3] *University of Texas at Austin, USA*

摘要

对于多维参数空间而言，构造客观先验充其量是一项极具挑战性的工作。一种常见做法是假设独立性，并将联合先验设为通过“标准”客观方法（如 Jeffreys 先验或参考先验）获得的边缘分布的乘积。然而，先验独立性的假设并不总是合理的，而且能否将其严格视为客观的，仍有待讨论。本文将此前针对一维情形提出的、基于评分规则的客观方法加以扩展，提出一种适用于多维参数空间的新型客观先验，该先验能够产生依赖结构。所提出的先验具有一个很有吸引力的性质：它是适定的，并且不依赖于所选择的模型，而只取决于所考虑的参数空间。

1 引言

众所周知，在多维参数空间中推导客观先验颇具挑战。一种可能的解决方案是假设先验独立，并将联合先验设为各边际分布的乘积。若要为参数向量 $\boldsymbol{\theta} = (\theta_1, \ldots\ldots, \theta_d)$ 指定客观先验，独立先验的思路是令

$$\pi(\boldsymbol{\theta}) = \pi_1(\theta_1) \times \cdots \times \pi_d(\theta_d),$$

其中 $\pi_j(\theta_j)$ 对于 $j = 1$、……、p，可采用任何“标准”的客观方法求得，例如 Jeffreys先验（Jeffreys，1961年）或参考先验（Bernardo，1979年）。然而，随着模型规模不断扩大，基于模型为所有参数设定相互独立的先验，或使用 Jeffreys先验或参考先验的做法，都存在太多问题；参见 Consonni等人（2018年）。尽管在某些情况下，上述为多维参数

空间设定客观先验的做法或许可取，但必须考虑到，先验独立性的假设 并非总是令人满意的默认选择。由于缺乏更好的选择（例如特定的论证依据），很难令人信服地将某种策略认定为客观。

本文着重于提出一种适用于多维参数空间的先验，该先验基于客观性考量，并扩展了Walkerand Villa（2021）的工作。

对现有用于处理多维参数空间的客观方法进行回顾相当直接。Jeffreys本人（Jeffreys，1961年）承认，采用其方法（即*Jeffreys*规则先验）为多维参数空间获取先验分布存在局限，因此提出采用先验独立性（即独立*Jeffreys*先验）。另见Kass and Wasserman (1999)中的讨论。至于参考先验这一类先验分布，尽管一维情形已有形式化定义（Berger等，2009），但针对任意参数空间的扩展尚未给出。对于某些特定情形，也存在特设解决方案，例如如何处理干扰参数，见（Liseo，1993）；但当参数不能被视为干扰参数时，这些方案并未明确说明应如何处理。此外，所得先验分布往往会因赋予感兴趣参数的“重要性顺序”不同而有所差异。

除了上述显而易见的局限之外，在处理客观先验时还必须考虑一个方面，即它们通常是不适定的。众所周知，只要相应的后验分布是适定的，这本身 $per\ se$ 并不构成问题，但后验适定性仍然需要得到验证。对于多维参数空间而言，这项任务的繁琐性和困难显而易见。关于这一问题的讨论，请参阅 Kass and Wasserman (1999)。

本文旨在针对多维参数空间提出一种新颖的客观先验，其具有一个颇具吸引力的性质，即先验本身是适定的。自然地，这种先验无需验证后验适定性；更重要的是，在不适合使用不适定先验分布的所有情形下都可以采用它，例如处理贝叶斯因子时。该思想建立在 Walker and Villa (2021) 提出的程序之上，该程序利用信息、散度与评分规则之间的联系来获得一种客观先验分布。事实上，该方法假设评分函数为常数，这意味着相对于其他参数值，我们不偏好任何特定的参数值。它还具有使信息度量最小化的性质。该先验的一个关键方面是，它不依赖于所选择的模型，而只依赖于参数空间。

本文的组织结构如下。第 2 节介绍 $d = 2$ 情形下先验分布的推导，并说明如何推导一般的多维情形。第 3 节通过模拟数据和真实数据，阐述所提出的先验分布的实现；我们讨论的示例包括单样本结果以及基于重复抽样的频率学派分析。最后，第 4 节以简短的讨论和一些说明作结。

2 先验推导

我们首先回顾 Walker 和 Villa (2021) 提出的思想。这一点很有用，因为扩展到多维情形需要更复杂的计算，但其基本原理大体保持不变。考虑散度、信息和评分之间的一般关系，即对于密度 p 和 q

$$D(p,q) = I(p) + \int p\, S(q), \tag{1}$$

其中 D 表示两个密度函数之间的散度，$I(p) = -\int p\, S(p)$ 是相应的信息，而 $S(q)$ 是评分函数。对于一维参数，Walker and Villa (2021) 采用了布雷格曼散度（Bregman，1967），$D(p,q) = \int B_\phi(p,q)$，其中

$$B_\phi(p,q) = \phi(p) - \phi(q) - \frac{\partial \phi}{\partial q}(p-q) - \frac{\partial \phi}{\partial q'}(p'-q'),$$

其中 $\phi(q,q')$ 是某个凸函数。对于局部评分函数，要得到形式（1），必须令

$$\phi(q,q') = q\frac{\partial \phi}{\partial q} + q'\frac{\partial \phi}{\partial q'},$$

选择 $\phi(u,v) = u\,\alpha(v/u)$ 即可保证这一点，其中 α 是某个凸函数。接着令评分函数为 0，得到客观先验 q，它是以下方程的解

$$\frac{\partial \phi}{\partial q} = \frac{d}{dx}\frac{\partial \phi}{\partial q'}.$$

当凸函数 α 取为“简单”形式 $\alpha(u) = u^{-(k+1)}$ 时，结果为

$$q(x) = \frac{k\, a^k}{(a+x)^{k+1}},$$

其中 $k \geq 1$ 和 $a > 0$ 为某些参数（详情见附录 A）。这是 Lomax 分布族中的一个成员，其密度函数记为 $\mathrm{LM}(a,k)$。此处我们暂不进一步指定 a 和 k，只指出，就不变性和信息最小化而言，$a = 1$ 和 $k = 1$ 分别是有充分依据的选择。为进一步阐明这些观点，若现在进行变换 $y = 1/x$，则

$$q(y) = \frac{k\, a^k\, y^{k-1}}{(1+ay)^{k+1}},$$

这仅在 $a = k = 1$ 时与 $q(x)$ 相匹配。此外，令 $a = 1$，这对后续讨论不失一般性，然后考虑负熵

$$I(k) = \log k + (k+1)\int \log(1+x)\,\frac{k}{(1+x)^{k+1}}\,dx.$$

积分为 $1/k$，因此 $I(k) = \log k + 1 + 1/k$，其最小值在 $k = 1$ 处取得。

2.1 维度2

为简化记号，我们用 x 和 y 表示先验 q的两个参数。在介绍评分规则的思路之前，我们先介绍多元 Lomax 分布的解及其一些引人注目的性质。

如果先验上不假设 X 与 Y 独立，那么条件先验分布也应采用 Lomax 分布的形式，因为它是一维的，并应遵循上一节讨论的思想；因此， $\pi(y \mid x)$ 将服从参数为 $a(x)$ 和 $k(x)$ 的 Lomax 分布。若边际先验 $\pi(x)$ 满足：对于 X，其分布为 LM(a, k)，并且取 $a(x) = a + x$和 $k(x) = k + 1$，则

$$\pi(x, y) = \frac{(k+1)\,(a+x)^{k+1}}{(a+x+y)^{k+2}} \frac{k\,a^k}{(a+x)^{k+1}} = \frac{k(k+1)a^k}{(a+x+y)^{k+2}}, \tag{2}$$

这是一个二元洛马克斯密度函数。此外，两个边际分布均为 LM(a, k)。

现在我们考虑布雷格曼散度的二元形式

$$D(\boldsymbol{p}, \boldsymbol{q}) = \int_{\Omega} B_{\phi}(\boldsymbol{p}, \boldsymbol{q})\, dx\, dy,$$

其中积分是在空间 $(x, y) \in \Omega \subseteq \mathbb{R}^2$ 上进行的，并且不失一般性，我们暂且假设 $\Omega = (0, \infty)^2$。定义

$$\boldsymbol{p}(x, y) = (p(x, y), p_x(x, y), p_y(x, y)),$$

对于 $\boldsymbol{q}$ 也有类似结论；对于二元函数 $f(x, y)$，记 $f_x = \partial f / \partial x$，对于 f_y 也有类似记法。具体而言，我们有

$$B_{\phi}(\boldsymbol{p}, \boldsymbol{q}) = \phi(\boldsymbol{p}) - \phi(\boldsymbol{q}) - \frac{\partial \phi}{\partial q}(p - q) - \frac{\partial \phi}{\partial q_x}(p_x - q_x) - \frac{\partial \phi}{\partial q_y}(p_y - q_y). \tag{3}$$

重新整理后，在假设密度在端点处消失的条件下使用分部积分，式 (3) 的积分变为

$$\int \left\{ \phi(\boldsymbol{p}) + p \left[-\frac{\partial \phi}{\partial q} + \frac{d}{dx}\frac{\partial \phi}{\partial q_x} + \frac{d}{dy}\frac{\partial \phi}{\partial q_y} \right] - \left[\phi(\boldsymbol{q}) - q\frac{\partial \phi}{\partial q} - q_x \frac{\partial \phi}{\partial q_x} - q_y \frac{\partial \phi}{\partial q_y} \right] \right\}.$$

为得到方程 (1) 的分解，我们需要

$$\phi(\boldsymbol{q}) = q\frac{\partial \phi}{\partial q} + q_x \frac{\partial \phi}{\partial qx} + q_y \frac{\partial \phi}{\partial q_y}, \tag{4}$$

从而得到评分函数

$$S(\boldsymbol{q}) = -\frac{\partial \phi}{\partial q} + \frac{d}{dx}\frac{\partial \phi}{\partial q_x} + \frac{d}{dy}\frac{\partial \phi}{\partial q_y}.$$

遵循 Walker 和 Villa (2021) 概述的推理，我们提出

$$\phi(q, q_x, q_y) = q\,\alpha\left(\frac{q_x}{q}, \frac{q_y}{q}\right), \tag{5}$$

对于某个凸函数 $\alpha(u, v)$。在缺乏信息的情况下，合理的假设是，对于伯格曼散度和评分规则的定义而言，x 和 y 具有同等的重要性。对称条件 $\alpha(u, v) = \alpha(v, u)$ 保证了这一点。我们还可以基于增加维度这一概念，为多维情形的加性版本提供动机，因此我们取

$$\alpha(u, v) = \frac{1}{u^{k+2}} + \frac{1}{v^{k+2}}.$$

现在，令 $S(q) = 0$ （我们提醒一下，这是我们的“客观”论元），可得到解 $q(x, y) = \pi(x, y)$，如式（2）所示。

2.2 一般多维情形

扩展到 d 维情形（其中 $d > 2$）相对直接，使用以下凸函数:

$$\alpha(u_1, \ldots, u_d) = \sum_{j=1}^{d} \frac{1}{u_j^{k+d}}.$$

如果我们采用上文所述的相同策略，则得到的解为

$$q(x_1, \ldots, x_d) = \frac{k(k+1)\ldots(k+d-1)\,a^k}{(a + x_1 + \cdots + x_d)^{k+d}},$$

这就是多元Lomax分布的密度，记为 LM(d, k, a)。所有边际密度和条件密度函数也都属于 Lomax 分布族。

特别地，对于 $S \subset \{1, \ldots, d\}$，$\mathbf{x}_S$ 的边缘密度为 LM$(|S|,\ k,\ a)$，其基于以下积分

$$\int_0^\infty (a + y)^{-k}\, dy = a^{-k+1}/(k-1)$$

对于 $k > 1$，给定 $\mathbf{x}_T$ 的 $\mathbf{x}_S$ 的条件分布为 LM$(|S|,\ k + d - |S|,\ a + \sum_{i \in T} x_i)$。

这是一个重要的相容性条件：所有边缘先验分布和条件先验分布都属于同一个族。关于多元 Lo-max 分布的更多内容，请参阅 Hutchinson（1979年）、Lindley 和 Singpurwalla（1986）以及 Nayak（1987）。

3 示例

本节将介绍多元Lomax分布作为客观先验的一些应用。具体而言，我们会进行一些单样本分析，并

分析该先验在重复样本上的频率学派表现。对于后者，我们将评估先验密度在均方误差（MSE）方面的表现，或采用信息量更高的相对均方根误差版本，并考察95%后验可信区间的覆盖率。

我们首先将所提出的先验用于威布尔分布的参数估计，以此展示参数维度为二的情形。随后，我们重点讨论 Dagum 分布的三个参数（Dagum，1975），给出一个关键示例，说明如何在维度大于二的参数空间中实际应用客观多元先验分布。最后，我们使用多元Lomax先验估计线性回归模型的参数，该模型具有正态误差项和两个协变量。该先验同时赋予截距、两个系数和回归方差；据我们所知，这是唯一一种能够以直接方式实现这一点的客观先验分布。

3.1 威布尔分布

文献中已有一些研究关注使用客观先验对威布尔分布的参数进行贝叶斯估计，尤其聚焦于 Jeffreys先验和参考先验。Sun（1997）对可能的客观先验进行了讨论和比较。作者指出参考先验更优，但将哪个参数（若有）选作干扰参数至关重要，因为不同的选择会产生不同的先验分布。

我们考虑采用以下参数化形式的威布尔分布

$$f(x|\theta,\beta) = \beta x^{\beta-1}\,\theta^{-\beta}\exp\left\{-(x/\theta)^{\beta}\right\}, \qquad x>0,$$

其中，θ 为尺度参数，β 为形状参数。我们用于与所提出的方法进行比较的客观先验，是在两个参数都受到关注的情形下的参考先验，即 $\pi(\theta,\ \beta) \propto (\theta\beta)^{-1}$。遵循 Sun (1997) 中模拟研究的指导原则，我们从参数为 $\theta=1$ 和 $\beta=0.5$, 1，10 的威布尔分布中，分别抽取了样本量为 $n=30$ 和 $n=100$ 的250个独立样本。这是因为尺度参数的变化不会影响先验的行为。

模拟研究的结果汇总于表1（ n =30）和表2（ $n=100$）。关于相对均方根误差（均方误差，MSE），其定义为 $\sqrt{\text{MSE}(\theta-\widehat{\theta})}/\theta$，其中 $\widehat{\theta}$ 为后验均值，我们注意到，在所有情形下，Lomax先验都优于参考先验。此外，正如预期的那样，对于两种先验分布，样本量越大，均方误差越小。就95%后验可信区间的覆盖率而言，我们没有发现任何明显偏离合理预期的情况，因为这些数值都在通常的容许范围内。

最后，对本研究进行总结时，我们分析一些取自 Ellah（2012）的真实数据。这些数据由绝缘流体在 34 千伏电压下、两电极之间发生击穿的 19 个时间（分钟）组成（表3）。

均方误差: β	$\beta = 0.5$	$\beta = 1$	$\beta = 10$	均方误差: θ	$\beta = 0.5$	$\beta = 1$	$\beta = 10$
参考值	3.92	3.90	3.91		3.13	3.13	3.12
Lomax	3.33	3.27	3.21		2.59	2.61	2.69
覆盖率: β				覆盖率: θ			
	$\beta = 0.5$	$\beta = 1$	$\beta = 10$		$\beta = 0.5$	$\beta = 1$	$\beta = 10$
参考值	0.90	0.91	0.91		0.91	0.92	0.91
Lomax	0.91	0.91	0.90		0.95	0.96	0.96

表 1：威布尔分布参数的均方误差及 95%后验可信区间的覆盖率，其中 $\theta = 1$ 和样本量为 $n = 30$；每种参数组合基于 250 个独立样本。

均方误差: β	$\beta = 0.5$	$\beta = 1$	$\beta = 10$	均方误差: θ	$\beta = 0.5$	$\beta = 1$	$\beta = 10$
参考值	1.85	1.86	1.94		1.36	1.37	1.37
Lomax	1.77	1.75	1.76		1.29	1.29	1.30
覆盖率: β				覆盖率: θ			
	$\beta = 0.5$	$\beta = 1$	$\beta = 10$		$\beta = 0.5$	$\beta = 1$	$\beta = 10$
参考值	0.94	0.95	0.94		0.94	0.93	0.94
Lomax	0.93	0.93	0.92		0.95	0.94	0.94

表2：基于每种参数组合的250个独立样本，威布尔分布参数的均方误差和95%后验可信区间的覆盖率，其中 $\theta = 1$ ，样本量为 $n = 100$。

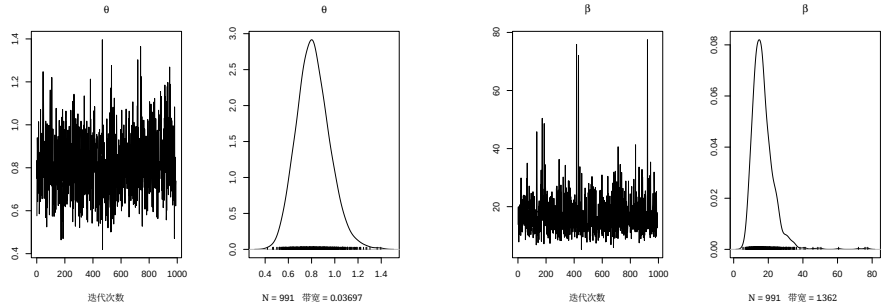
0.96	4.15	0.19	0.78	8.01	31.75	7.35	6.50	8.27	33.91
32.52	3.16	4.85	2.78	4.67	1.31	12.06	36.71	72.89	

表3：在34千伏电压下，电极间绝缘流体发生击穿的19次时间观测值（单位：分钟）。

$\hat{}$这些参数的最大似然估计值为 $\theta = 0.77$ 和 $\backsim\beta = 12.22$。查看参考先验和 Lomax 先验下后验分布的后验统计量（表4），可以发现，采用两种先验分布得到的区间均包含相应参数的最大似然估计值。对于θ，两种先验分布的表现相当接近。然而，对于形状参数β，Lomax 先验的表现优于参考先验；比较两者的后验方差和 95% 可信区间即可明显看出这一点。为完成这一示例，我们在图1中展示后验链和估计密度。

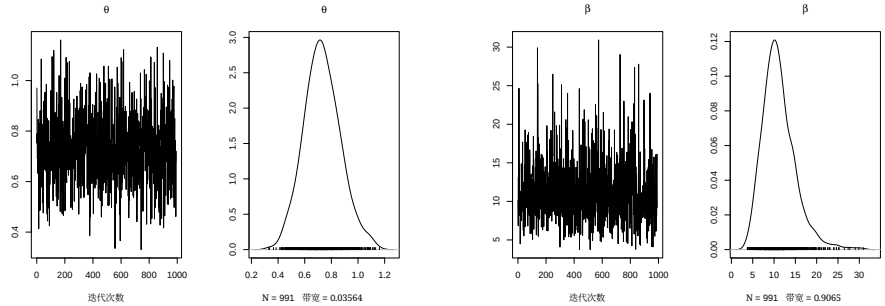
	参考值			Lomax		
	Mean	方差	95%置信区间	Mean	方差	95%置信区间
θ	0.8	0.02	(0.55,1.10)	0.73	0.02	(0.48,1.02)
β	16.84	44.93	(8.51,31.89)	11.11	15.08	(5.07,20.36)

表4：威布尔分布真实数据示例的后验汇总。最大似然估计为 $\hat{\theta} = 0.77$ 和 $\hat{\beta} = 12.22$。



（a）参考先验： θ

（b）参考先验： β



（c）Lomax先验： θ

（d）Lomax先验： β

图1： 使用参考先验（上）和Lomax先验（下）时，威布尔分布尺度参数（左）和形状参数（右）的后验链与估计密度。

3.2 Dagum分布

Dagum分布最早由 Dagum（1975）提出，现广泛应用于经济学，例如用于对收入和财富数据进行建模。如果随机变量 X 具有以下概率密度函数，则称其服从 Dagum分布：

$$f(x|a,b,p) = \frac{a\,p}{x}\left(\frac{\{x/b\}^{ap}}{\{(x/b)^a+1\}^{p+1}}\right), \qquad x>0,$$

其中， $p>0$ 和 $a>0$ 是形状参数， $b>0$ 是尺度参数。

关于采用客观先验估计 Dagum分布参数的贝叶斯文献很少，Nawash等（2017）是个例外。该文作者提出了 Jeffreys先验的一种扩展；然而，严格来说，所提出的先验并非客观先验，因为它依赖于一个必须设定的超参数，而该超参数会影响先验本身的表现。因此，为了将多元Lomax分布与另一种方案进行适当比较，我们将考虑一个由三个独立弥散伽马密度的乘积构成的三参数联合先验。也就是说， $\pi(p,\ a,\ b)\ =\mathrm{Ga}(a_1,\ b_1)\cdot\mathrm{Ga}(a_2,\ b_2)\cdot\mathrm{Ga}(a_3,\ b_3)$，其中 a_1、 a_2、 a_3、 b_1、 b_2和 b_3 都相对较小（例如等于 0.1 或 0.01）。

为了将多元Lomax先验与上述弥散先验进行比较，我们分析两种情景。首先，考虑参数为 ($p=1,\ a=2.1,\ b=1$) 的 Dagum分布，然后考虑参数为 ($p=2,\ a=2.1,\ b=$2) 的 Dagum分布。我们选择了 $a>2$ ，以确保该分布的均值和方差均存在。对于每种情景，我们分别以 $n=30$ 和 $n=100$ 为样本量重复抽取 250 个独立样本。参数的后验分布没有封闭形式，因此我们采用 Metropolis-within-Gibbs算法从中抽样。

在表 5 和表 6 中，我们报告了上述情景在两种先验分布下的相对均方误差（平方根）以及 95% 后验可信区间的覆盖率。我们注意到，在这两种情况下，多元Lomax

MSE	模糊Lomax分布		模糊Lomax分布	
参数	n = 30		n = 100	
p	13.28	6.59	7.87	4.89
a	2.80	2.05	1.45	1.24
b	4.26	3.18	2.92	2.44
COV	模糊Lomax分布		模糊Lomax分布	
参数	n = 30		n = 100	
p	0.97	1.00	0.95	0.97
a	0.96	0.99	0.96	0.96
b	0.98	1.00	0.94	0.96

表5：基于250次重复抽样（样本量分别为 $n=30$ 和 $n=100$，其中 $p=1$、 $a=2.1$ 和 $b=1$），Dagum分布参数的相对均方误差（上）和95%后验可信区间的覆盖率（下）。表中比较了使用多元Lomax分布和三个独立弥散伽马密度乘积得到的数值。

在均方误差方面优于联合无信息Gamma密度。对于参数 p 的估计，这一优势似乎更为显著，表明 Lomax先验具有更高的准确性。就覆盖率而言，各先验分布的表现趋于相似；正如预期，样本量较小时区间相对较宽。

最后，为结束对 Dagum 密度的示例说明，我们分析英国 Quar-

MSE	模糊Lomax分布		模糊Lomax分布	
参数	n = 30		n = 100	
p	8.05	3.55	9.21	4.52
a	2.38	2.03	1.24	1.11
b	3.96	2.80	3.09	2.38
COV	模糊Lomax分布		模糊Lomax分布	
参数	n = 30		n = 100	
p	0.99	0.99	0.97	0.98
a	0.96	0.98	0.97	0.98
b	1.00	1.00	0.95	0.98

表 6：基于 250 次重复抽样（样本量分别为 $n = 30$ 和 $n = 100$），Dagum 分布参数 $p = 2$、$a = 2.1$ 和 $b = 2$ 的相对均方误差（上）和 95% 后验可信区间的覆盖率（下）。该表比较了使用多元Lomax分布和三个独立弥散伽马密度的乘积所得的数值。

期天然气消费量数据，涵盖1960–1986年（Durbin and Koopman，2001），这是一个公开数据集。我们将三维多元Lomax分布与三个弥散 Gamma 密度的乘积进行比较。图 2 展示了将多元Lomax分布用作Dagum分布三个参数的联合先验时所得的后验轨迹和估计密度；这里，Dagum分布被假设为数据的统计模型。图 3 展示了将三个独立弥散 Gamma 密度的乘积用作联合先验时的相同图形。显然，在两种先验分布下，三个边际后验密度似乎都集中在相同的值附近，但Lomax先验下后验分布所体现的变异性，也即不确定性，似乎小于竞争性的弥散 Gamma 密度。这一点也得到了表 7 所报告的后验统计量的证实。

参数	Lomax		Gamma	
p	25.5	(6.61, 104.19)	27.2	(3.40, 209.14)
a	1.77	(1.50, 2.04)	1.76	(1.52, 2.10)
b	30.66	(12.86, 76.68)	29.1	(7.31, 109.23)

表 7：真实数据示例中 Dagum 分布参数的后验中位数和后验95%可信区间。左侧为使用多元Lomax先验得到的值，右侧为使用由三个独立弥散伽马密度的乘积构成的先验得到的值。

3.3 线性回归

所提出的先验有一项有趣的应用，即估计线性回归模型的参数。通常，这类模型的客观先验分布会将

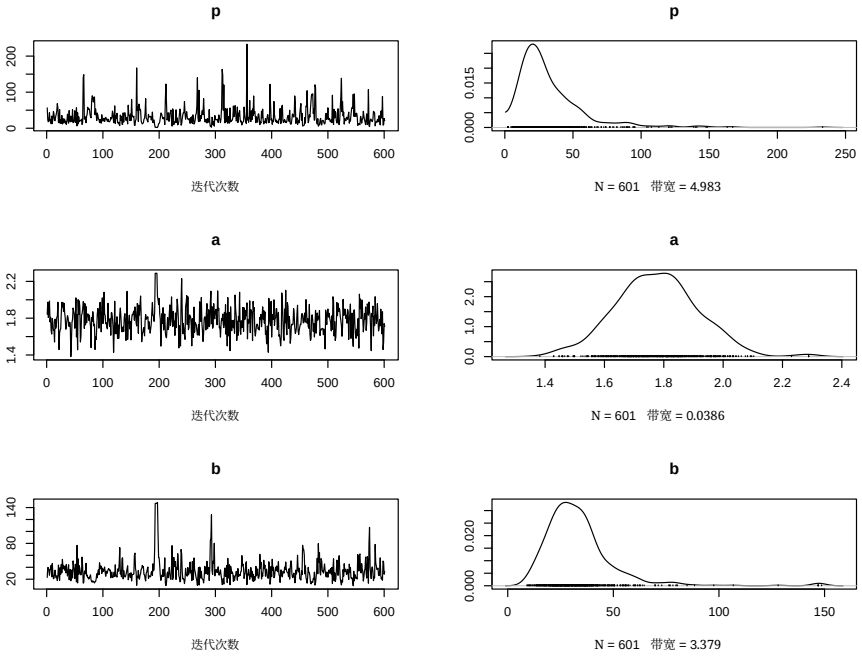


图 2：采用多元Lomax先验时 Dagum 参数的后验链和估计密度。

系数与截距和回归方差分开。换言之，对于该模型

$$y_i = \beta_0 + \sum_{i=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad \text{for} \quad i = 1, \ldots, n,$$

当 $\varepsilon_i \sim N(0,\ \sigma^2)$ 时，客观先验的类型将为

$$\pi(\beta_0, \beta_1, \ldots, \beta_p, \sigma) = \pi(\beta_0, \sigma)\pi(\beta_1, \ldots, \beta_p).$$

通常，在客观贝叶斯设定中，人们会设定 $\pi(\beta_0,\ \sigma) \propto \sigma^{-1}$，而系数的先验可以采用不同形式，例如 Zellner 和 Siow（1980年）、Fern´andez 等人（2001年）、Liang 等人（2008年）和 Bayarri 等人（2012年）提出的形式。这里，我们通过一个单样本示例，讨论所提出的先验的实现，并在重复抽样方案下进行频率学派分析。

3.3.1 线性回归：单样本

考虑从包含两个协变量的线性回归模型中抽取的一个大小为 $n = 100$ 的样本，其中协变量为 x_1 和 x_2 （二者均定义在 $\mathbb{R}$上），截距为 $\beta_0 = 20$，系数为 $\boldsymbol{\beta} = (10, -1)$，方差为 $\sigma^2 = 2$。这个（4 维）Lomax先验具有

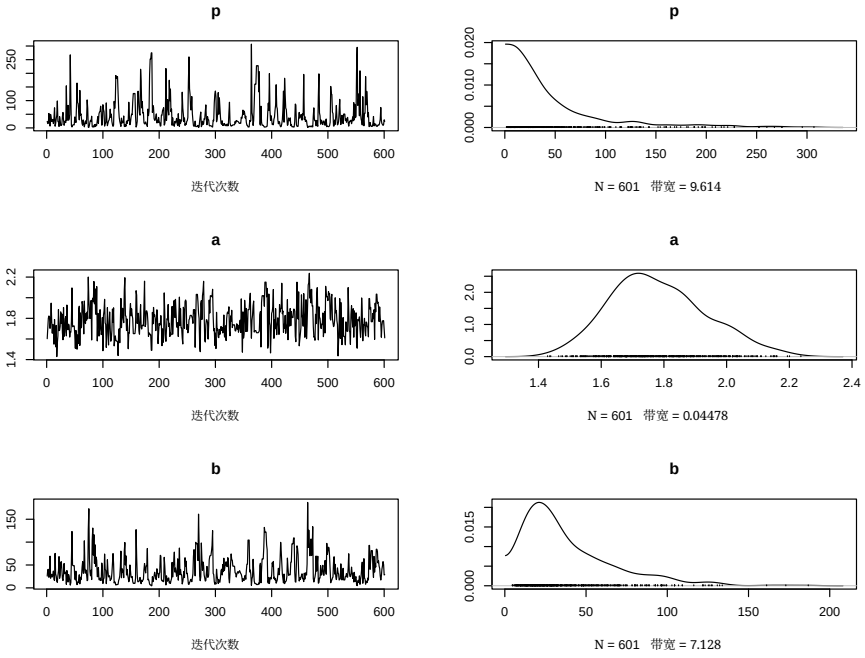


图 3：采用相互独立的弥散伽马先验时，Dagum 分布参数的后验链和估计密度。

以下形式，

$$\pi(\beta_0, \beta_1, \beta_2, \sigma^2) = 3 \left(1 + \sum_{j=0}^{2} |\beta_j| + \sigma^2\right)^{-5}, \tag{6}$$

其中 $\beta_j \in \mathbb{R}$，对于 $j = 0$，1，2，以及 $\sigma^2 > 0$。为了研究后验分布，我们采用了马尔可夫链蒙特卡罗（MCMC）算法，运行 100,000 次迭代，舍弃前 10,000 个样本作为预热期，并设置 50 个样本的抽稀间隔。

图 4 展示了四个边际后验样本的直方图，表 8 则报告了相应的后验汇总统计量。我们

参数	中位数	95%置信区间
β_0	20.15	(19.84, 20.47)
β_1	10.05	(9.96, 10.14)
β_2	-1.02	(-1.10, -0.94)
σ^2	2.41	(1.83, 3.26)

表 8：单样本回归示例的后验汇总统计量。

可以看到，真实参数值均包含在后验 95%

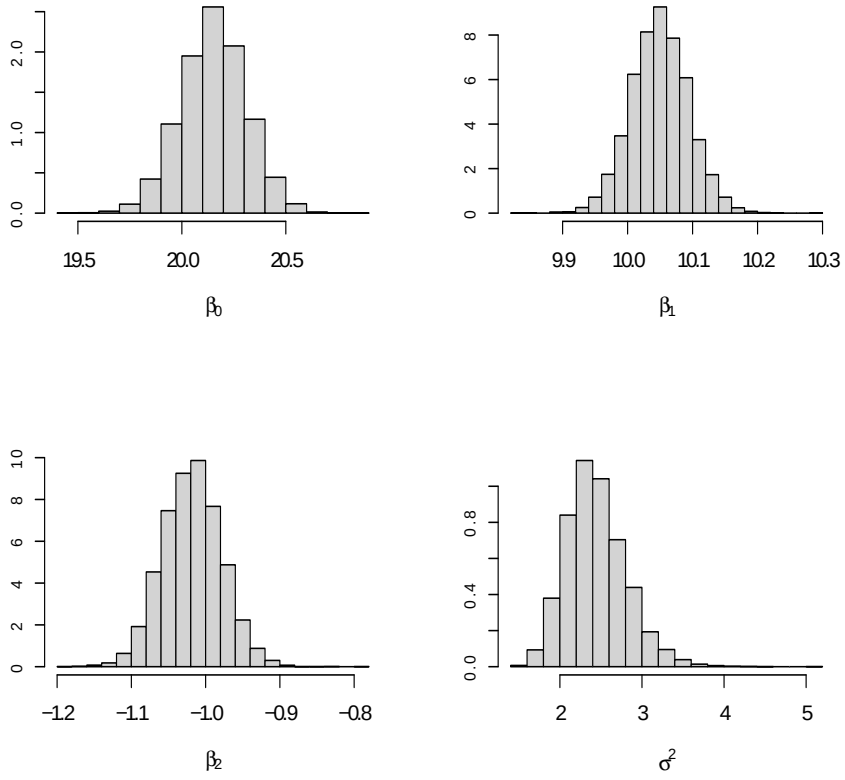


图4：线性回归模型参数的后验样本直方图，其中截距为 $\beta_0 = 20$，系数为 $\beta_1 = 10$ 和 $\beta_2 = -1$，方差为 σ^2=2

.

可信区间内，表明所提出的先验具有良好的推断表现。

3.3.2 线性回归：频率学派分析

为了展示所提出的先验的性能，并将其与其他方法进行比较，我们从一个线性回归模型中抽取了 250 个样本，每个样本的大小为 $n = 100$ 。该模型的截距为 $\beta_0 = 20$，系数为 $\beta_1 = 5$ 和 $\beta_2 = -3$，方差为 $\sigma^2 = 2$。式 (6) 给出了所提出的先验，并将其与一个

独立弥散先验进行比较，即

$$\begin{aligned}\pi(\beta_0, \beta_1, \beta_2, \sigma^2) &= \pi(\beta_0, \beta_1, \beta_2)\pi(\sigma^2) \\ &\propto \sigma^{-1}\,\mathrm{MVN}(\mathbf{0}, c\mathrm{I}_3),\end{aligned}$$

其中 I_3 是一个维度为 3 的对角矩阵，且 $c > 0$ 的取值足够大，可表示弥散先验信息。此外，为了将所提出的先验与被视为客观的先验进行比较，我们使用了 Zellner 的以下版本 g 先验（Zellner，1986）

$$\begin{aligned}\pi(\beta_0, \beta_1, \beta_2, \sigma^2) &= \pi(\sigma)\pi(\beta_0, \beta_1, \beta_2|\sigma) \\ &\propto \sigma^{-1}\,\mathrm{MVN}(\mathbf{0}, \sigma^2 g\mathbf{X}),\end{aligned}$$

其中 $g > 0$ 已校准为 500，$\mathbf{X}$ 是设计矩阵。表 9 报告了三种先验分布下最大后验（MAP）估计的相对平方误差（平方根）。可以看出，这些数值几乎相同。表 10 报告了以下参数的覆盖率

参数	Lomax	弥散先验	Zellner 的 g
β_0	0.998	0.998	0.998
β_1	1.000	1.000	1.000
β_2	1.001	1.002	1.001
σ^2	1.098	1.095	1.094

表 9：线性回归模型的最大后验估计平方误差，对比所提出的先验与独立弥散先验及 Zellner 的 g 先验。

各先验下的 95% 后验可信区间。同样，各先验分布的表现几乎相同。

参数	Lomax	弥散先验	Zellner 的 g
β_0	0.95	0.94	0.95
β_1	0.97	0.96	0.97
β_2	0.97	0.98	0.98
σ^2	0.93	0.92	0.92

表 10：线性回归模型中 95% 后验可信区间的覆盖率，比较所提出的先验、独立弥散先验与 Zellner 的 g 先验。

4 讨论

本文提出了一种设定多维客观先验分布的通用方法。该先验是适当的，因为它服从多元 Lomax 分布，并且具有一些颇具吸引力的一致性性质

这些性质表现为：所有联合分布、边缘分布和条件分布也都是 Lomax 分布。

通过对模拟数据和真实数据进行分析，我们展示了所提出的先验在不同情景下的表现，包括假设数据服从威布尔分布或 Dagum 分布的情形，以及考虑线性回归模型的情形。在所有情况下，多元先验的表现都不逊于竞争性方案；对于威布尔分布和 Dagum 分布情形，它实际上优于替代先验分布。

关于线性回归模型，一个有趣的最后说明是，系数上的多元Lomax分布所产生的结果可以与 LASSO回归进行比较。事实上，可以注意到，多元Lomax分布是关于指数分布的独立拉普拉斯密度混合。也就是说，

$$p(x_1, ..., x_d) = \int \prod_{j=1}^{d} \text{Lap}(x_j|s)e^{-s}\, ds, \tag{7}$$

其中，Lap$(\cdot|s)$ 是位置参数为零、尺度参数为 s 的拉普拉斯密度。如果我们将上述公式（7）应用于线性回归模型，那么就需要最小化以下函数

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + (d+1)\log(1+|\beta|),$$

该函数可以与 LASSO 的最小化函数进行比较，后者的最后一项将是 $|\beta|$，而不是 log$(1+|\beta|)$）。其思想是，将惩罚项置于对数尺度上可以得到更稳健的估计。更多内容请参见附录B。

参考值

M. J. Bayarri, J. O. Berger, A. Forte 和 G. Garc´ıa-Donato。贝叶斯模型选择的准则及其在变量选择中的应用。统计年鉴，40(3):1550–1577，2012。ISSN 00905364, 21688966。J. O. Berger, J. M. Bernardo 和 D. Sun。参考先验的形式化定义。统计年鉴，37:905–938，2009。J. M. Bernardo。贝叶斯推断的参考后验分布。皇家统计学会杂志，*B*辑（方法论），41(2):113–147，1979年。L. Bregman。求凸集公共点的松弛法及其在求解凸规划问题中的应用。苏联计算数学与数学物理，7:200–217，1967。

G. Consonni, D. Fouskakis, B. Liseo 和 I. Ntzoufras。客观贝叶斯分析的先验分布。贝叶斯分析，第627–679页，2018年。C. Dagum。收入分布模型及有限阶矩存在的条件。国际统计学会通报，46:199–205，1975年。J. Durbin 和 S. Koopman。状态空间方法的时间序列分析。牛津大学出版社，牛津，英格兰，2001年。A. H. Ellah。基于广义次序统计量的逆威布尔模型的贝叶斯与非贝叶斯估计。智能信息管理，4:23–31，2012年。C. Fern´andez, E. Ley 和 M. F. Steel。贝叶斯模型平均的基准先验。计量经济学杂志，100(2):381–427，2001年。T. P. Hutchinson。二元帕累托分布的四种应用。生物计量学杂志，21(6):553–563，1979年。doi: https://doi.org/10.1002/bimj. 4710210605。H. Jeffreys。概率论。牛津大学出版社，牛津，英格兰，1961年。R. E. Kass 和 L. Wasserman。通过形式规则选择先验分布。美国统计学会杂志，91:1343–1370，1999年。F. Liang, R. Paulo, G. Molina, M. A. Clyde 和 J. O. Berger。贝叶斯变量选择的g先验混合。美国统计学会杂志，103(481):410–423，2008年。D. V. Lindley 和 N. D. Singpurwalla。共享共同环境的系统组件寿命的多元分布。应用概率杂志，23(2):418–431，1986。doi: 10.2307/3214184。B. Liseo。利用参考先验消除干扰参数。*Biometrika*，80:295–304，1993年。S. Nawash, S. Ahmad 和 A. Ahmed。Dagum分布的贝叶斯分析。可靠性与统计研究杂志，10:123–136，2017年。T. K. Nayak。多元Lomax分布：性质及其在可靠性理论中的用途。应用概率杂志，24(1):170–177，1987。ISSN 00219002。E. Ovcharov。适当评分规则与布雷格曼散度。*Bernoulli*，24:53–79，2018年。D. Sun。关于威布尔分布非信息先验的简短说明。统计规划与推断杂志，61:319–338，1997年。

S. G. Walker 和 C. Villa。基于评分规则的客观先验。*Entropy*，23（7），2021年。

A. Zellner。关于使用 g 先验分布评估先验分布和贝叶斯回归分析。 载于 ***Goel, P.; Zellner, A.***（编）。《贝叶斯推断与决策技术：纪念 ***Bruno de Finetti*** 的论文集》。贝叶斯计量经济学与统计学研究。第 *6* 卷。，第 233–243 页，1986。

A. Zellner 和 A. Siow。所选回归假设的后验赔率比。*Trabajos de Estad´ıstica y de Investigaci´on Operativa*，31：585–603，1980年。

附录

A 数学细节

定义 $\mathbf{x} = (x_1, \ldots, x_d)$ 和 $\mathbf{q} = (q, q_1, \ldots, q_d)$，其中，对于 $j = 1, \ldots, d$

$$q_j = \frac{\partial q}{\partial x_j},$$

对于$\mathbf{p}$也同样如此。我们将概率测度P 和 Q 之间的散度定义为布雷格曼散度（参见Bregman（1967）和Ovcharov（2018年）：

$$D(P, Q) = \int_{\Omega} B_{\phi}(\boldsymbol{p}, \boldsymbol{q})\, d\mathbf{x}, \tag{8}$$

其中， p 和 q 分别是与 P 和 Q相关联的密度，并且

$$B_{\phi}(\boldsymbol{p}, \boldsymbol{q}) = \phi(\boldsymbol{p}) - \phi(\boldsymbol{q}) - \frac{\partial \phi}{\partial q}(p - q) - \sum_{i=1}^{d} \frac{\partial \phi}{\partial q_i}(p_i - q_i), \tag{9}$$

对于定义在 $\mathbb{R}^{d+1}$上的某个连续可微凸函数 ϕ 。

引理1 相应的*Bregman*评分函数为

$$S(\boldsymbol{q}) = -\frac{\partial \phi}{\partial q} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \frac{\partial \phi}{\partial q_i},$$

在极值满足以下条件的假设下

$$p_i(\mathbf{x})\, \frac{d}{dx_i}\, \frac{\partial \phi}{\partial q_i}$$

对每个 $i.$ 和 ϕ 都消失，且满足

$$\phi(\mathbf{q}) = q \frac{\partial \phi}{\partial q} + \sum_{i=1}^{d} q_i \frac{\partial \phi}{\partial q_i}. \tag{10}$$

证明。 引理陈述中的条件允许对每个 i 的 $p_i\,\partial\phi/\partial q_i$ 进行分部积分。因此，我们可以将 (8) 重写为

$$\int \left\{ \phi(\boldsymbol{p}) + p\left[-\frac{\partial \phi}{\partial q} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\frac{\partial \phi}{\partial q_i} \right] - \left[\phi(\boldsymbol{q}) - q\frac{\partial \phi}{\partial q} - \sum_{i=1}^{d} q_i \frac{\partial \phi}{\partial q_i} \right] \right\} d\mathbf{x}.$$

这表明评分函数如前所述，其中 ϕ 满足引理陈述中给出的约束条件

■

为满足 ϕ 的约束条件，我们定义 $\alpha : \mathbb{R}_+^d \to \mathbb{R}$ 为

$$\alpha(\mathbf{u}) = \sum_{i=1}^{d} u_i^{-(k+d)}$$

以及任意定义在 $(0,\infty)$ 上的密度 q 对应的 $u_i = q_i/q$。注意，一阶和二阶偏导数为

$$\alpha_i(\mathbf{u}) = -(k+d)u_i^{-(k+d+1)} \quad \text{and } \alpha_{ii}(\mathbf{u}) = (k+d)(k+d+1)u_i^{-(k+d+2)},$$

因此海森矩阵的行列式为

$$|H\alpha(\mathbf{u})| = (k+d)^d(k+d+1)^d \prod_{i=1}^{d} u_i^{-(k+d+2)}$$

因此，当 $k \in \mathbb{N}$ 为奇数时，对于 $u \in (0,\infty)^d$， $\alpha(\mathbf{u})$ 是凸函数。

Let

$$\phi(\mathbf{q}) = q\,\alpha\left(\frac{q_1}{q}, \ldots, \frac{q_d}{q}\right). \tag{11}$$

引理 2 方程 (11) 中的函数 ϕ 满足 *(10)*。

证明。 我们针对 $d = 2$ 的情形给出证明；对于更高维度的情形，数学推导同样成立。以下为 $q_1 = q_x$ 和 $q_2 = q_y$。现在

$$\begin{aligned}\frac{\partial \phi}{\partial q} &= q\left[\alpha_u\left(\frac{q_x}{q}, \frac{q_y}{q}\right)\left(\frac{q_x}{q^2}\right) + \alpha_v\left(\frac{q_x}{q}, \frac{q_y}{q}\right)\left(-\frac{q_y}{q^2}\right)\right] + \alpha\left(\frac{q_x}{q}, \frac{q_y}{q}\right) \\ &= \alpha\left(\frac{q_x}{q}, \frac{q_y}{q}\right) - \frac{1}{q}\left[q_x\alpha_u\left(\frac{q_x}{q}, \frac{q_y}{q}\right) + q_y\alpha_v\left(\frac{q_x}{q}, \frac{q_y}{q}\right)\right],\end{aligned}$$

and

$$\frac{\partial \phi}{\partial q_x} = q\,\alpha_u\left(\frac{q_x}{q}, \frac{q_y}{q}\right)\frac{1}{q} = \alpha_u\left(\frac{q_x}{q}, \frac{q_y}{q}\right),$$

and

$$\frac{\partial \phi}{\partial q_y} = \alpha_v\left(\frac{q_x}{q}, \frac{q_y}{q}\right),$$

这表明条件（10）得到满足。 ■

推论 3 与凸函数 ϕ （由(11)给出）相对应的布雷格曼散度(9) 下，参数$\mathbf{x} \in \mathbb{R}_+^d$ 的基于得分的先验为多元*Lomax*分布 $LM(d, a, k)$，其中 $\alpha(\mathbf{u}) = \sum_{i=1}^{d} u_i^{-(k+d)}$。

B 关于 LASSO 的替代方案的更多内容

这里我们讨论使用惩罚项 $\log(1+|\beta|)$ 进行变量选择的正则化思想，因此目标是最小化

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + 2\lambda\log(1+|\beta|),$$

对于某个 $\lambda > 0$。如果改为考虑一维情形（此时问题可简化为最小化），就更容易看出使用 $\log(1+|\beta|)$ 而非更常见的 $|\beta|$（它是 LASSO 的特征形式）的后果。

$$\beta^2 - 2\beta z + 2\lambda\log(1+|\beta|).$$

严格来说，这是假设 $\sum_{i=1}^{n} x_i^2 = 1$ 和 $z = \sum_{i=1}^{n} x_i\, y_i$。

这个问题的解很容易求得：

$$\widehat{\beta} = \begin{cases} 0 & z < \lambda \\ (z-1)/2 + \sqrt{z - \lambda + (z-1)^2/4} & z > \lambda, \end{cases}$$

如果 $z > 0$ 且

$$\widehat{\beta} = \begin{cases} 0 & z > -\lambda \\ (1+z)/2 - \sqrt{-\lambda - z + (1+z)^2/4} & z < -\lambda, \end{cases}$$

如果 $z < 0$。

在许多方面，这个解与 LASSO问题相似，即最小化

$$\beta^2 - 2\beta z + 2\lambda|\beta|,$$

$\widehat{}$ $\widehat{}$其解在 $|z| < \lambda$ 时为 $\beta = 0$，在 $z > \lambda$ 时为 $\beta = z - \lambda$，而在 $z < -\lambda$ 时为 $\sim\beta = z + \lambda$。因此，例如，当 $z \to \infty$ 时，LASSO 估计量永远不会回到标准无偏估计量 z。然而，使用 $\log(1+|\beta|)$ 惩罚时则会。我们看到，当 $z \to \infty$ 时，因此

$$(z-1)/2 + \sqrt{z - \lambda + (z-1)^2/4} - z \to 0.$$

$\widehat{}$需要注意的是，LASSO 和对数惩罚都能在 $|z| < \lambda$ 时达到 $\beta = 0$。因此，我们应该关注的是 $|z| > \lambda$ 时的情况，而这里的简单事实是，对于 $|z| > \lambda$，对数惩罚估计量接近 z，而后者本身更接近真实参数。

为了进行这样的实验，我们取了 $n = 100$ 并进行了抽样

$$y_i = \beta + \varepsilon_i,$$

其中，(ε_i) 是相互独立的标准正态随机变量。计算$\sim z = \bar{y}$，我们得到基于 LASSO 的估计量 β_L 和基于对数惩罚的估计量 $\sim\beta_N$；重复进行此类实验 1000 次后，我们得到每个估计量的均方

误差；即 $M_L = E(\widehat{\beta}_L - \beta)^2$ 和 $M_N = E(\widehat{\beta}_N - \beta)^2$。每种情形下的 M_N 都小于 M_L。例如，当 $\lambda = 1/2$且 $\beta = 2$时，我们得到 $M_L = 0.260$ 和 $M_N = 0.043$。保持 $\lambda = 1/2$ 不变，并将 $\beta = 1/2$ 设为.....，得到 $M_L = 0.216$ 和 $M_N = 0.197$。这些改进$\widehat{\ }$ $\widehat{\ }$是基于以下事实：对于 $z > \lambda$，我们有 $\beta_L < \beta_N < z$；对于 $z < -\lambda$，则有$\backsim$ $\backsim z < \beta_N < \beta_L$。见图5。

当使用 $\beta = 0$ 时，我们得到 $M_L = M_N = 0$，这表明在每种情况下都有 $|z| < \lambda$。如果保留 $\beta = 0$，现在取 $\lambda = 0.1$，则得到 $M_L = 0.0016$ 和 $M_N = 0.0019$。后一结果很容易解释，因为当 $|z| > \lambda$时，LASSO估计量更接近 0。然而，当 $\beta \neq 0$ 时，对数惩罚估计量具有更小的均方误差。

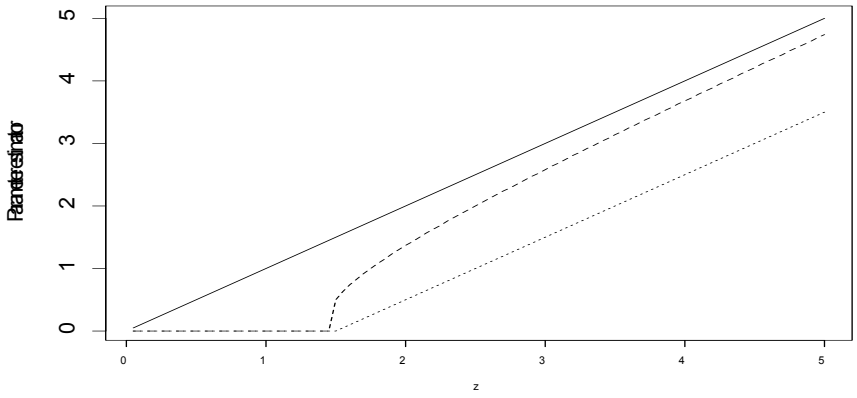


图 5：普通最小二乘估计量（粗线）、LASSO估计量（点线）和对数惩罚估计量（虚线）。