经验贝叶斯与 James-Stein 估计量 *

大规模推断讨论班

Charles Stein 在 1955 年证明, 使用 100 多年的极大似然估计方法，对于超过二维的高斯模型是不容许的（即存在比极大似然更好的估计），这一论断震惊了当时整个统计界。虽然基于某些原因，极大似然估计法仍被广泛应用，但 Stein 估计已经从根本上指出了新的视角，即用经验贝叶斯的方法解决高维统计推断问题，包括估计、检验以及预测。经验贝叶斯是一些列方法的统称，我们以 Stein 估计为例来展示其应用。

Stein(1956) 发表后，学者并没有意识到其与经验贝叶斯的关联，但 Stein 的工作仍可看做是经验贝叶斯理论的滥觞。经验贝叶斯理论另外一个源头的开创者是 Herbert Robbins，他创造了“经验贝叶斯”这个学术名词, 试图展示频率论学派也可以在像贝叶斯学派那样进行有效的大规模并行研究（即大规模推断）[1]。然而，大规模并行研究在 20 世纪 50 年代很少被提及，而 Stein 估计却可以在小的数据集上得到漂亮的应用，所以 Robbins 的理论并没有像 Stein 估计那样大的影响。

这一切在 21 世纪发生了改变，最新的科学技术提出了对并行研究的需求[2]（例如，以基因芯片为例，我们可以获取 100 名实验者（50 名正常人与 50 名患病者）的基因数据，每个人的基因长度为 6000，我们意在寻找那些在实验者与正常人之间区别较大的基因，也就是要对每个基因做假设检验，检验它在正常人群体与患病者群体之间是否不同，现在的问题是，我们不是有一个基因要分析（检验），而是有 6000 个基因要分析（检验），这就是所谓的并行研究），Robbins 的思想处理这类问题得心应手，而其思想将在本书随后的章节贯穿始终。

Stein 的理论关注估计，而 Robbins 的理论侧重假设检验，在 2.6 节我们会看到上述两种理论是紧密联系融为一体的，是经验贝叶斯理论的一体两面。经验贝叶斯理论使得参数估计与假设检验，频率学派与贝叶斯学派方法的界限变得模糊。

1 贝叶斯公式与多元正态分布

这一节简要的回顾一下贝叶斯方法在多元正态估计方面的应用。贝叶斯公式体现了简单却意义深远的统计思维。虽然通常是在离散情形下表达贝叶斯公式，但是也可以清晰地从概率密度的角度来阐述它。设定模型如下：

$$\mu \sim g(.) \quad \text{和} \quad z \mid \mu \sim f_\mu(z)。 \tag{1.1}$$

$g(\mu)$ 是未知参数向量 μ 的先验分布，$f_\mu(z)$ 是在给定 μ 时的条件概率密度。

在观测值 z 条件下，可以利用贝叶斯公式，求解 μ 的条件概率分布（它的后验分布），即

$$g(\mu|z) = g(\mu)f_\mu(z)/f(z) \tag{1.2}$$

其中，$f(z)$ 是 z 的边际分布，即

*本文作者为大规模推断讨论班，成员：杨晓康、张洋、宋培培、张猛、刘博、朱祁恒和高磊.

$$f(z)=\int g(\mu)f_\mu(z)d\mu, \tag{1.3}$$

(1.3) 是计算（1.2）式最困难的部分，但通常并没有必要计算它。一般情况下，知道后验分布 $g(\mu|z)=g(\mu)f_\mu(z)$（即先验分布 $g(\mu)$ 和似然函数 $f_\mu(z)$ 的乘积）成比例（可参照下面的 (1.4) 式）就足够了。因为对于参数的任意两个可能取值 μ_1、μ_2，由（1.2）可得，

$$\frac{g(\mu_1|z)}{g(\mu_2|z)}=\frac{g(\mu_1)}{g(\mu_2)}\frac{f_{\mu_1}(z)}{f_{\mu_2}(z)} \tag{1.4}$$

也就是说，我们所关心的参数两个取值的后验比率，是先验比率和似然比率的乘积，与（1.3）式无关，因此没有必要花费精力计算（1.3）。

练习 1.1 设 μ 服从先验分布为均值为 0，方差为 A 的正态分布，给定 μ 下，z 的条件分布为均值为 μ 方差为 1 的正态分布，即

$$\mu\sim N(0,A)\quad 和\quad z\mid\mu\sim N(\mu,1) \tag{1.5}$$

请证明

$$\mu\mid z\sim N(Bz,B)\qquad 其中B=A/(A+1)。 \tag{1.6}$$

练习 1.1 解答：按（1.2）式，有，

$$g(\mu|z)=g(\mu)f_\mu(z)/f(z)\propto g(\mu)f_\mu(z)$$

其中，

$$g(\mu)=\frac{1}{\sqrt{2\pi}\sqrt{A}}e^{-\frac{\mu^2}{2A}},f_\mu(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-\mu)^2}{2}}$$

因此，μ 的后验密度可以写为:

$$g(\mu|z)\propto g(\mu)f_\mu(z)=Const\times e^{-\frac{\mu^2}{2A}-\frac{(z-\mu)^2}{2}}$$

补全平方项，得，

$$g(\mu|z)=Const\times e^{-\frac{(\mu-\frac{A}{A+1}z)^2}{2\frac{A}{A+1}}}$$

令 $B=A/(A+1)$, 则，

$$g(\mu|z)=Const\times e^{-\frac{(\mu-Bz)^2}{2B}}$$

上式是正态分布密度函数形式，所以，

$$\mu\mid z\sim N(Bz,B)\qquad 其中B=A/(A+1)。 \tag{1.6}$$

注：虽然在贝叶斯框架下，没有必要求 $f(z)$，但经验贝叶斯情形下，$f(z)$ 需要求出。下面给出在（1.5）式的假设下，求解 z 的边际分布 $f(z)$ 过程：

在计算 $f(z)$ 时，可尝试套用下面的积分公式以方便计算：

$$\int_{-\infty}^{+\infty}e^{-\frac{(x-c)^2}{b}}dx=(b\pi)^{0.5},\qquad 其中\ \mathrm{b}、\mathrm{c}\ 均为常数。 \tag{*}$$

为求解 $f(z)$，将（1.5）式的先验分布以及条件分布密度函数带入（1.3）式，并化简得，

$$f(z)=\frac{1}{2\pi\sqrt{A}}\int_{-\infty}^{+\infty}e^{-\frac{\mu^2}{2A}}e^{-\frac{(z-\mu)^2}{2}}d\mu,$$

将指数（即 e 右上角上部分）合并同类项，并凑完全平方式得，

$$f(z)=\frac{1}{2\pi\sqrt{A}}e^{-\frac{z^2}{2(A+1)}}\int_{-\infty}^{+\infty}e^{-\frac{(\mu-\frac{Az}{A+1})^2}{\frac{2A}{A+1}}}d\mu,$$

利用 (*) 式的结论得，

$$\int_{-\infty}^{+\infty}e^{-\frac{(\mu-\frac{Az}{A+1})^2}{\frac{2A}{A+1}}}d\mu=\sqrt{\frac{2A\pi}{A+1}}$$

将上式结论代入 $f(z)$ 求解式并化简得，

$$f(z)=\frac{1}{\sqrt{2\pi}\sqrt{A+1}}e^{-\frac{z^2}{2(A+1)}},$$

由此可得，

$$z\sim N(0,A+1)$$

如果现在要同时处理许多（1.5）式，这就是进行大规模推断，即

$$\mu_i\sim N(0,A)\quad 和\quad z_i\mid\mu_i\sim N(\mu_i,1)\quad [i=1,2,.....N],\tag{1.7}$$

其中每对 (μ_i,z_i) 之间相互独立，令 $\boldsymbol{\mu}=(\mu_1,\mu_2,.....\mu_N)'$，那么我们就可以用简洁的标准符号把（1.7）式表示成 N 维正态分布了。即

$$\boldsymbol{\mu}\sim N_N(0,AI)\tag{1.8}$$

以及

$$\mathbf{z}|\boldsymbol{\mu}\sim\boldsymbol{N}_N(\boldsymbol{\mu},\boldsymbol{I})\tag{1.9}$$

其中 $\boldsymbol{I}$ 是 N 阶单位矩阵。仿照练习 1.1 的解答，可以得到 $\boldsymbol{\mu}$ 的后验分布

$$\boldsymbol{\mu}|\mathbf{z}\sim\boldsymbol{N}_N(\mathbf{Bz},\mathbf{B}\boldsymbol{I})\qquad[\mathbf{B}=A/A+1]\tag{1.10}$$

以上是贝叶斯公式在多元正态分布总体均值推断中的应用。在统计实践中，给定观测值 $\mathbf{z}$, 我们希望用估计量 $\hat{\boldsymbol{\mu}}=t(\mathbf{z})$ 来估计 $\boldsymbol{\mu}$，

$$\hat{\boldsymbol{\mu}}=(\hat{\mu}_1,\hat{\mu}_2,...,\hat{\mu}_N)'\tag{1.11}$$

我们用总平方误差损失来计算用 $\hat{\boldsymbol{\mu}}$ 来估计 $\boldsymbol{\mu}$ 的误差，如下所示:

$$\mathbf{L}(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})=\parallel\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\parallel^2=\sum_{i=1}^{N}(\hat{\mu_i}-\mu_i)^2\tag{1.12}$$

与此相关的风险函数就是在给定 $\boldsymbol{\mu}$ 的时候，$\mathbf{L}(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})$ 的期望值, 即:

$$\boldsymbol{R}(\boldsymbol{\mu})=\mathbf{E}_{\boldsymbol{\mu}}\{\mathbf{L}(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})\}=\mathbf{E}_{\boldsymbol{\mu}}\{\parallel t(\mathbf{z})-\boldsymbol{\mu}\parallel^2\}\tag{1.13}$$

其中 $\mathbf{E}_{\boldsymbol{\mu}}$ 就是关于 $\mathbf{z}\sim\boldsymbol{N}_N(\boldsymbol{\mu},\boldsymbol{I})$,$\boldsymbol{\mu}$ 固定时的期望。

很明显，我们用 $\mathbf{z}$ 来估计 $\boldsymbol{\mu}$，即:

$$\hat{\boldsymbol{\mu}}^{(MLE)} = \mathbf{z} \tag{1.14}$$

在模型（1.9）下,$\mathbf{z}$ 是 $\boldsymbol{\mu}$ 的极大似然估计量, 其推导过程很简单: 把 N 个正态分布密度函数连乘 $\longrightarrow$ 取对数 $\longrightarrow$ 对 $\boldsymbol{\mu}$ 求导并令导数等于 0。在给定 $\boldsymbol{\mu}$ 下，极大似然估计的风险为:

$$\mathbf{R}^{(MLE)}(\boldsymbol{\mu}) = N \tag{1.15}$$

(1.15) 推导: 由 (1.14) 和 $\hat{\boldsymbol{\mu}} = t(\mathbf{z})$，得 $t(\mathbf{z}) = \mathbf{z}$, 将其带入 (1.13) 式，(1.13) 式变为:

$$\mathbf{R}^{(MLE)}(\boldsymbol{\mu}) = \mathbf{E}_{\boldsymbol{\mu}}\{\| \mathbf{z} - \boldsymbol{\mu} \|^2\} = \mathbf{E}_{\boldsymbol{\mu}} \left\{ \sum_{i=1}^{N} (z_i - \mu_i)^2 \right\}$$

其中，z_i 服从 $N(\mu, 1)$, 所以 $z_i - \mu$ 服从 $N(0, 1)$，所以 $\sum_{i=1}^{N}(z_i - \mu_i)^2$ 服从 $\chi^2(N)$ 分布（多个标准正态随机变量的平方和服从卡方分布）。因为卡方分布的期望等于其自由度，所以，

$$E(\chi^2(N)) = N$$

由此 (1.15) 式得证。

在参数空间中每一个点我们都用 $\hat{\boldsymbol{\mu}}^{(MLE)}$ 来代替，这对于一般的估计目的看似是挺合理的。

先验信息（1.8）式告诉我们，参数 $\boldsymbol{\mu}$ 在原点 $\mathbf{0}$ 附近。根据 (1.10) 式，贝叶斯估计值为

$$\hat{\boldsymbol{\mu}}^{(Bayes)} = \mathbf{Bz} = \left(1 - \frac{1}{A+1}\right)\mathbf{z} \tag{1.16}$$

这是对应于平方误差损失函数的一种选择。

注：(1.10) 式是用贝叶斯方法，得到的参数后验分布，按理说，贝叶斯分析到这，就大功告成。可是，我们总是想知道这个参数的估计值是多少，因此，光有后验分布还不行。从后验分布到参数估计值，有好多选择，比如期望、众数、中位数等，选哪个值呢？在不同的评判标准下，会有不同的选择。如果我们评判工具是平方损失损失函数，那么，就如（1.16）式，就把参数后验分布的期望作为参数估计值。

假定 A=1，那么得到的贝叶斯估计值 $\hat{\boldsymbol{\mu}}^{(Bayes)} = (1 - \frac{1}{2})\mathbf{z}$，就相当于把极大似然估计 $\hat{\boldsymbol{\mu}}^{(MLE)}$ 向 $\mathbf{0}$ 收缩，而且，收缩幅度竟然达到二者差距的一半。

(1.16) 式贝叶斯估计值的风险为

$$\mathbf{R}^{Bayes}(\boldsymbol{\mu}) = (1 - \mathbf{B})^2 \| \boldsymbol{\mu} \|^2 + N\mathbf{B}^2 \tag{1.17}$$

(1.17) 式推导: 根据风险函数的定义得

$$\mathbf{R}^{Bayes}(\boldsymbol{\mu}) = \mathbf{E}_{\boldsymbol{\mu}}[(\mathbf{Bz} - \boldsymbol{\mu})'(\mathbf{Bz} - \boldsymbol{\mu})]$$

展开得，

$$\mathbf{R}^{Bayes}(\boldsymbol{\mu}) = \mathbf{E}_{\boldsymbol{\mu}}[\mathbf{B}^2\mathbf{z}^{'}\mathbf{z} - 2\mathbf{B}\mathbf{z}^{'}\boldsymbol{\mu} + \boldsymbol{\mu}^{'}\boldsymbol{\mu}]$$

运用二阶矩、期望、方差的关系可得，

$$\mathbf{R}^{Bayes}(\boldsymbol{\mu}) = \mathbf{B}^2(\parallel \boldsymbol{\mu} \parallel^2 + N) - 2\mathbf{B} \parallel \boldsymbol{\mu} \parallel^2 + \parallel \boldsymbol{\mu} \parallel^2$$

$$\mathbf{R}^{Bayes}(\boldsymbol{\mu}) = (\mathbf{B}^2 - 2\mathbf{B} + 1) \parallel \boldsymbol{\mu} \parallel^2 + N\mathbf{B}^2$$

进一步可得，

$$\mathbf{R}^{Bayes}(\boldsymbol{\mu}) = (1 - \mathbf{B})^2 \parallel \boldsymbol{\mu} \parallel^2 + N\mathbf{B}^2 \tag{1.17}$$

估计的整体风险是对以上风险再求期望（把 $\boldsymbol{\mu}$ 看做随机变量），即:

$$\mathbf{R}_A^{(Bayes)} = \mathbf{E}_A\{\mathbf{R}^{(Bayes)}(\boldsymbol{\mu})\} = N\frac{A}{A+1} \tag{1.18}$$

(1.18) 式推导: 由于 $\mathbf{B} = \frac{A}{A+1}$ 且 $\mathbf{E}(\parallel \boldsymbol{\mu} \parallel^2) = NA$, 把其带入

$$\mathbf{E}_A\{(1 - \mathbf{B})^2 \parallel \boldsymbol{\mu} \parallel^2 + N\mathbf{B}^2\}$$

就可得到 (1.18) 式的结果。

练习 1.2 证明（1.17）和（1.18）（略）

根据（1.15）式，相应地极大似然估计 $\hat{\mu}^{(MLE)}$ 的整体风险是：

$$R_A^{(MLE)} = N$$

如果先前的（1.8）式成立，则 $\hat{\mu}^{(Bayes)}$ 的风险比 $\hat{\mu}^{(MLE)}$ 小，有：

$$R_A^{(MLE)} - R_A^{(Bayes)} = \frac{N}{(A+1)} \tag{1.19}$$

当 A=1 时，$\hat{\mu}^{(Bayes)}$ 的风险只有 $\hat{\mu}^{(MLE)}$ 一半大小。

2 经验贝叶斯估计

假定（1.8）式正确，由于不知道 A 的值，因此不能使用 $\hat{\mu}^{(Bayes)}$。这时候经验贝叶斯就派上用场了。（1.8）（1.9）式暗示了 Z 的边缘分布（对 $Z \sim N_N(0, AI)$ 积分，积掉 μ，此时 $\mu \sim N_N(0, AI)$）得：

$$Z \sim N_N(0, (A+1)I) \tag{1.20}$$

令 $S = \|Z\|^2$，则 S 服从自由度为 N 的卡方分布（尺度为 $A+1$），即:

$$S \sim (A+1)\chi_N^2 \tag{1.21}$$

因此

$$E\{\frac{N-2}{S}\} = \frac{1}{A+1} \tag{1.22}$$

练习 1.3 证明（1.22）。

练习 1.3 解答： 由于 $S \sim (A+1)\chi^2_N$，则

$$\frac{S}{(A+1)} \sim \chi^2_N$$

那么，

$$\frac{(A+1)}{S} \sim Inverse—\chi^2_N$$

根据逆卡方分布的性质可知，

$$E\{\frac{(A+1)}{S}\} = \frac{1}{(N-2)}$$

化简得，

$$E\{\frac{N-2}{S}\} = \frac{1}{A+1}$$

James-Stein 估计量[3] [4] 被定义为：

$$\hat{\boldsymbol{\mu}}^{(JS)} = \{1 - \frac{(N-2)}{S}\}\mathbf{z} \tag{1.23}$$

这里用无偏估计量 $\frac{(N-2)}{S}$ 代替（1.16）式里的未知项 $\frac{1}{(A+1)}$ 。我们通过经验数据估计得到先验分布中的某些参数，这也是经验贝叶斯中的“经验”一词的缘由所在[5] [6] [7] [8]。

不难得出 James-Stein 估计量的全部贝叶斯风险是：

$$R^{(JS)} = \frac{NA}{(A+1)} + \frac{2}{(A+1)} \tag{1.24}$$

这比（1.18）式得出的真正的贝叶斯风险稍微大一些：

$$\frac{R_A^{(JS)}}{R_A^{(Bayes)}} = 1 + \frac{2}{(NA)} \tag{1.25}$$

举例来说，当 N=10，A=1 时，$R_A^{(JS)}$ 只比（1.18）全部贝叶斯风险大 20%。

James-Stein 估计量带给统计学的震撼并非来自于 (1.24) 和 (1.25)。更大的震撼来自于 James 和 Stein 在 1961 年证明的一个定理。下面，就介绍一下这个定理。

定理：当 $N \geq 3$ 时,James-Stein 估计量的均方误差总是小于极大似然估计量的均方误差，即，

$$E_\mu\left(\| \hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu} \|^2\right) < E_\mu\left(\| \hat{\boldsymbol{\mu}}^{(MLE)} - \boldsymbol{\mu} \|^2\right) \tag{1.26}$$

传统上认为，正态分布的样本均值不管对几维都是容许的，但是 James 和 Stein 发现三维以上的样本均值是不容许的，James-Stein 估计对于多维均值的估计而言是一个更好的统计量。

Charls Stein 其人其事 以下内容是我根据吴建福先生在北大许宝騄讲座上的一段话总结的，吴建福先生在讲话中对 Stein 的学术成就以及人品风骨赞赏有加。

事 1 Stein 把他的惊人发现发表在了 1956 年的”Inadmissibility of the usual estimator for the mean of a multivariate distribution” 这篇文章里，后来 Charls Stein 和 Willard James 在 1961 年发表的“Estimation with Quadratic Loss”这篇文章里给出了前者的一个详细证明。Willard James

Figure 1: Charls Stein(1920-)

是 Stein 的一名非全职弟子，与老师合写的这篇文章是他发表的唯一作品, 写就这篇文章之后他就去当了一名社区老师。尽管如此，在 1961 年的这篇文章中，Stein 把 James 排在第一作者的位置，而且 Stein 在给这个特殊的估计方法取名时也加上了 James 的名字，并且也排在了首位，由此可见 Stein 胸怀之宽广。

事 2 五六十年代的美国，麦卡锡主义盛行，在这个背景下，州立大学要对美国国家效忠。在 Berkeley，Stein 是一个左派，但他拒绝效忠，结果他被迫离开 Berkeley（当时只有他一个人离开），去了 Stanford。我们知道，那个年代，Stanford 的统计学不如 Berkeley，Stein 相当于从一个统计学术重镇去了一个统计氛围不是很浓厚的地方。这是 Stein 非常让人佩服的一点，也就是我们中国人常讲的“不为五斗米折腰”。

事 3 在 Stanford 期间，Stein 因为抗议（具体事情不详），居然被抓到监狱里去呆了一段时间。堂堂一名美国科学院院士，被抓到监狱里去，这在中国恐怕是不敢想象的。

事 4 Condoleezza Rice(时任美国国务卿) 去 Stanford 大学，想要去各学院转一转。Stanford 许多教授以及院士就写抗议信，抗议 Rice 干涉大学事务。抗议信写好后，谁去带头交给校长呢？Stein 交出去的，他带头交给了校长。当时 Stein 年事已高，身体也不好，连说话都不是很利索，但仍然顶住压力，敢于向“权贵”挑战。

James-Stein 估计量 $\hat{\mu}^{JS}$ 使每一个观测值 z_i 均向 0 收缩。当然，没有必要必须朝 0 收缩，更为一般的版本如下所示：

$$\mu_i \underset{ind}{} N(M\ A)\ ,\ z_i|\mu_i \underset{ind}{} N(\mu_i, \sigma_0^2)\ \ ,\ \ i = 1, 2, \ldots, N \tag{1.32}$$

其中 M 和 A 分别是先验分布的均值和方差。有一点需要注意，先假设这些参数是知道的，按照贝叶斯分析思路，迅速得到 z_i 的边际分布和 μ 的后验分布，

$$z_i \underset{ind}{} N(M, A + \sigma_0^2)\ \ ,\ \ \mu_i|z_i \underset{ind}{} N(M + B(z_i - M), B\sigma_0^2)\ \ ,\ \ i = 1, 2, \ldots, N \tag{1.33}$$

其中

$$B = \frac{A}{A + \sigma_0^2} \tag{1.34}$$

从而可以得到贝叶斯估计 $\hat{\mu}_i^{(Bayes)} = M + B(z_i - M)$。

不过，上述推导都建立在先验分布参数 M 和 A 已知的情况下，而实际情形中，我们并不能很容易地设定这些参数值。在 M 和 A 未知的情况下，我们通过 z_i 的信息来估计 M 和 A，从而得到

经验贝叶斯估计 $\hat{\mu}_i^{(JS)}$:

$$\hat{\mu}_i^{(JS)} = \overline{z} + \left(1 - \frac{(N-3)\sigma_0^2}{S}\right)(z_i - \overline{z}) \tag{1.35}$$

其中 $\overline{z} = \sum z_i/N$, $S = \sum(z_i - \overline{z})^2$ 。

注：如何从贝叶斯估计表达式过渡到经验贝叶斯估计表达式，即如何对 M 和 B 进行经验贝叶斯估计？这里做一下解释。

由 1.33 式 z_i 的边际分布可以知道 $E(z_i) = M$, 所以可以用 z_i 的均值来估计 M，从而有

$$\hat{M} = \overline{z},$$

由标准正态分布性质可得

$$\frac{\sum(z_i - \overline{z})^2}{A + \sigma_0^2} = \frac{S}{A + \sigma_0^2} \sim \chi^2(N-1),$$

从而有

$$\frac{A + \sigma_0^2}{S} \sim inverse - \chi^2(N-1),$$

由逆卡方分布性质可知

$$E(\frac{A + \sigma_0^2}{S}) = \frac{1}{N-3},$$

进一步可得

$$E(\frac{N-3}{S}) = \frac{1}{A + \sigma_0^2},$$

上式左右两侧同乘以 σ_0^2

$$E(\frac{(N-3)\sigma_0^2}{S}) = \frac{\sigma_0^2}{A + \sigma_0^2},$$

根据式 (1.34), 上式右侧可以变形为

$$E(\frac{(N-3)\sigma_0^2}{S}) = \frac{\sigma_0^2}{A + \sigma_0^2} = 1 - B,$$

最终得到 B 的经验贝叶斯估计

$$\hat{B} = 1 - \frac{(N-3)\sigma_0^2}{S}.$$

值得注意的是，在 z_i 并不是向着 0 收缩，而是向着值 M 收缩时，如果定理 1.26 式仍然成立，则要求 $N \geq 4$ 。

如果 1.26 式中 $\hat{\mu}^{(JS)}$ 和 $\hat{\mu}^{(MLE)}$ 这两种估计方法的风险差别比较小的话，$\hat{\mu}^{(JS)}$ 将仅仅是一个有趣的理论小新闻。实际上, 使用 $\hat{\mu}^{(JS)}$ 的好处多多。来看棒球（图 2) 比赛中，安打率计算的一个例子。

安打率是什么？安打率是基本的棒球统计数据之一，指的是选手在打击上的表现，通常计算方式是选手总安打数除以总打数，例如，某球员出赛 10 场总共有 35 个打数，击出 11 支安打，那他的安打率就是 0.314，通常打击率只算到小数点后 3 位，台湾通常会用「成」这个字来叙述打击率，0.314 就叫 3 成 14 安打率。

Figure 2: Roberto Clemente Walker(1934-1972)

棒球运动安打率是衡量选手打击能力的工具之一，单以安打率来说，若能达到 3 成，也就是 0.300，就是一个好击球员，也是中心击球员必须达成的目标，2 成 5 则算是下限，如果不到 2 成 5 就算击球不好了。

Table 1所示为 18 位棒球运动员在 1970 年早期赛季的安打率，记为 z_i (Table1第二列)，由中心极限定理可知 z_i 近似服从正态分布。球员真实安打率取的是剩余赛季比赛的平均安打率 μ_i，剩余赛季的击球共有 370 多次。我们可以用早期赛季的结果来估计 μ_i，估计方法有极大似然估计（Table1第三列）$\hat{\mu}_i^{MLE} = z_i$ 或者 James-Stein 估计（Table1第五列）$\hat{\mu}_i^{JS} = \overline{z} + \left(1 - \frac{(N-3)\sigma_0^2}{S}\right)(z_i - \overline{z})$。James-Stein 估计式中，$\overline{z}$ 为总体安打率，计算可得 $\overline{z} = 0.265$，在计算 $\hat{\mu}_i^{JS}$ 的过程中，还需要知道 σ_0^2, 我们用样本方差估计 σ_0^2, 即 $\sigma_0^2 = \overline{z}(1-\overline{z})/45$。

为了比较极大似然估计和 James-Stein 估计的效果，我们进一步计算极大似然估计和 James-Stein 估计的误差平方和，进而得到两种估计的误差平方和的比率，如下所示：

$$\sum_1^{18}\left(\hat{\mu}_i^{(JS)} - \mu_i\right)^2 \Big/ \sum_1^{18}\left(\hat{\mu}_i^{(MLE)} - \mu_i\right)^2 = 0.28$$

这个比率表明相对于极大似然估计，James-Stein 估计的误差平方和更小，大概是极大似然估计的 30%, 可见，在这个例子中 James-Stein 估计的优势是明显的。

对于 Stein 现象最开始的反应是认为它是一个悖论：在表最上方的 Clemente 与在表底部的 Munson 的成绩是相互独立的，为什么 Clemente 取得好成绩会提高我们对 Munson 的成绩的预测值？在 James-Stein 估计中，Clemente 取得好成绩会提高总体平均水平 $\overline{z}$,$\overline{z}$ 的增加会提高 $\hat{\mu}^{(JS)}$ 增大。这表明隐藏在队员之间的间接信息，对每一个队员安打率的估计也有影响。经典的贝叶斯理论通过先验分布建立了不同运动员之间的相互联系，但对于经验贝叶斯而言更加神奇，如果我们把经验贝叶斯分析比作一部机器，先验仅仅被看做一个这部机器的启动器。

3 估计独立成分

上一节中提到，James-Stein 估计量的性质要优于 MLE 估计量，那么在统计应用中，为什么我们仍在大量的使用 MLE 估计量呢？我们通过一个模拟的例子来探讨 James-Stein 估计量的局限

Table 1: 18 位球员的击球率数据统计

Name	hits/AB	$\hat{\mu}_i^{MLE}$	$\hat{\mu}_i$	$\hat{\mu}_i^{JS}$
Clemente	18/45	0.400	0.346	0.294
F Robinson	17/45	0.378	0.298	0.298
F Howard	16/45	0.356	0.276	0.285
Johnstone	15/45	0.333	0.222	0.280
Berry	14/45	0.311	0.273	0.275
Spencer	14/45	0.311	0.270	0.275
Kessinger	13/45	0.298	0.263	0.270
L Alvarado	12/45	0.267	0.210	0.266
Santo	11/45	0.244	0.269	0.261
Swoboda	11/45	0.244	0.230	0.261
Unser	10/45	0.222	0.264	0.256
Williams	10/45	0.222	0.256	0.256
Scott	10/45	0.222	0.303	0.256
Petrocelli	10/45	0.222	0.264	0.256
E Rodriguez	10/45	0.222	0.226	0.256
Campaneris	9/45	0.200	0.286	0.252
Munson	8/45	0.178	0.316	0.247
Alvis	7/45	0.156	0.200	0.242
Grand Average		0.265	0.265	0.265

性。如表 2所示，$N = 10$, 第一列为 $\mu_1, \mu_2, \ldots, \mu_{10}$ 的真值，其中 $\mu_{10} = 4$, 与其他值有较大的差异。我们在这些真值参数的基础上，随机产生正态分布的样本，然后再依据样本，分别用 MLE 和 James-Stein 估计量两种方法得到参数 μ_i 的估计值 $\hat{\mu}_i^{MLE}$ 和 $\hat{\mu}_i^{JS}$，并且我们还分别计算估计的误差 $(\hat{\mu}_i^{MLE} - \mu_i)^2$ 和 $(\hat{\mu}_i^{JS} - \mu_i)^2$。多次模拟后，对 $(\hat{\mu}_i^{MLE} - \mu_i)^2$ 求平均可得到 MLE 估计的均方误差 $MSE_i^{(MLE)}$, 对 $(\hat{\mu}_i^{JS} - \mu_i)^2$ 求平均可得到 James-Stein 估计量的的均方误差 $MSE_i^{(MLE)}$，通过对比这两个均方误差的大小，我们就可以分析两种估计方法的好坏。

以上模拟过程的 R 代码如下：

```
set.seed(123)
u <- c(-.81, -.39, -.39, -.08, -.69, -.71, 1.28, 1.32, 1.89, 4.00)
u.JS <- NULL
u.MLE <- NULL
for (i in 1:1000) {
  z <- rnorm(10, u, 1)
  u.mle <- z
  u.MLE <- rbind(u.MLE, z)
  z.average <- mean(z)
  S <- sum((z - z.average)^2)
```

Table 2: 模拟实验

	μ_i	$MSE_i^{(MLE)}$	$MSE_i^{(JS)}$
1	-0.81	0.95	0.61
2	-0.39	1.04	0.62
3	-0.39	1.03	0.62
4	-0.08	0.99	0.58
5	0.69	1.06	0.67
6	0.71	0.98	0.63
7	1.28	0.95	0.71
8	1.32	1.04	0.77
9	1.89	1.00	0.88
10	4.00	1.08	2.04!!
Total Sqerr		10.12	8.13

```
  u.js <- z.average + (1- (10-3) * 1 / S)*(z-z.average)
  u.JS <- rbind(u.JS, u.js)
}
apply((u.MLE-matrix(rep(u, 1000), nrow = 1000, byrow = T))^2, 2, mean)
apply((u.JS-matrix(rep(u, 1000), nrow = 1000, byrow = T))^2, 2, mean)
```

如表所示，从估计的均方误差角度看，在前九种情形下，$\hat{\mu}_i^{JS}$ 的估计效果要优于 $\hat{\mu}_i^{(MLE)}$，而在第十种情形下，$\hat{\mu}_{10}^{JS}$ 估计的均方误差比较大，估计效果不如 $\hat{\mu}_{10}^{(MLE)}$。不过，对估计的整体效果而言，James-Stein 估计的均方误差是 8.13，比极大似然估计的均方误差要小。

James-Stein 估计量关注的是整体估计的好坏，其目的是使整体误差损失函数 $\sum(\hat{\mu}_i - \mu_i)^2$ 最小, 在这样的目标下，其整体的估计效果比较好，但是对于那些本质上并不能划为一个类别的情形，比如上例中的 μ_{10}, 其估计效果就会打折扣了。棒球球迷知道，Clemente（图 2) 是一个非常优秀有着特殊天赋的棒球手，在对其击球率进行估计时，我们就不应该把他的水平向他的那些天分不是很高的队友的平均水平收缩，其中的原因就是不能笼统地把 Clemente 和他的队员混合在一起，这里 Clemente 可以看做是一个“变异”或“异常值”。

罗伯托・克莱门特・沃克（西班牙语：Roberto Clemente Walker，1934 年 8 月 18 日－ 1972 年 12 月 31 日）为波多黎各的棒球选手之一，曾经效力于美国职棒大联盟匹兹堡海盗队。1952 年成为布鲁克林道奇队自由契约球员，1954 年因规则五选秀转到匹兹堡海盗。17 年的职棒生涯，平均打击率 0.317，240 支本垒打，1305 分打点。1971 年他还率领海盗夺下个人第二枚世界大赛冠军戒指（七场世界冠军赛平均打击率 0.414，还附送对手二发本垒打），没想到在隔年 (1972 年)，却在搭机前往尼加拉瓜赈灾途中不幸坠机身亡。死后旋即在 1973 年以 92.7% 的得票率入选美国棒球名人堂。

在现在的统计应用实践中，人们似乎更加关注个体的推断，不愿意接受个体推断受整体水平影

响的想法，这也就解释了为什么极大似然估计在统计推断中这么流行。上世纪的前半叶，Fisher 这些统计元勋们开创的极大似然估计对于他们面临的数据而言是合适的，然而，在二十一世纪，我们面临的数据再也不是 Fisher,Neyman 所能想象的了，极大似然估计的这种推断方法对于大规模推断并不是最佳选择，以经验贝叶斯为核心的大规模统计推断方法更适用于现在的数据推断要求。

以上我们提到了 JS 估计量的一个不足，即有可能忽视个体因素的影响，那么如何能让这个估计量既能捕获整体平均水平为个体推断带来的好处，又能保护一些异常的独立观测所含有的信息呢？继续讨论上边提到的棒球例子，显然，对 Clemente 的估计不能简单地使用 JS 估计量的那种不分青红皂白的强制收缩，这里我们提出一个修正的 JS 估计量，又称为有限的转移估计：

$$\hat{\mu}_i^{(D)} = \begin{cases} max(\hat{\mu}_i^{(JS)}, \hat{\mu}_i^{(MLE)} - D\sigma_0) & , z_i > \bar{z} \\ min(\hat{\mu}_i^{(JS)}, \hat{\mu}_i^{(MLE)} + D\sigma_0) & , z_i \leq \bar{z} \end{cases}$$

当 z_i 远大于 $\bar{z}$ 时，不在将极大似然估计收缩到 $\hat{\mu}_i^{(JS)}$, 只是进行一个较小的收缩 $D\sigma_0$，当 z_i 远小于 $\bar{z}$ 时, 情况类似。下面我们采用这种有限转移估计量对击球率进行估计，并做出 $(z_i, \hat{\mu}_i^{(D)})$ 的散点图。

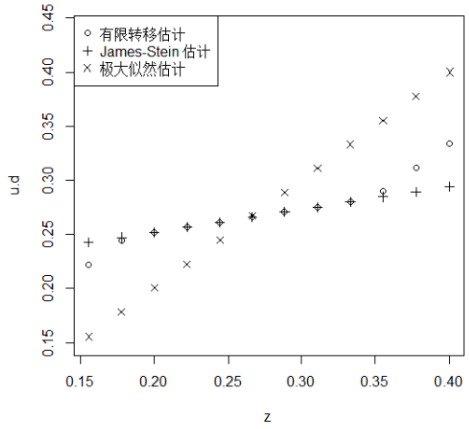


Figure 3: 三种估计方法比较

由图可见，有限转移估计可以看做是极大似然估计与 JS 估计的综合，尤其是在两端进行估计的时候。我们取 $D = 1$，这意味着，$\hat{\mu}_i^{(D)}$ 与极大似然估计的距离最大差距也就是 1 个标准差 $sigma_0 = 0.294$，因此，对棒球运动员 Clemente 的击球率估计值 $\hat{\mu}_1^{(D)} = 0.334$, 与 $\hat{\mu}_1^{(JS)} = 0.294$ 相比，更加靠近极大似然估计值。虽然，这可能会牺牲上文提到的 James-Stein 估计量带来的好处（总估计均方误差较小），但是牺牲的不是很多，$\hat{\mu}_i^{(D)}$ 较 $\hat{\mu}_i^{(JS)}$ 估计整体均方误差值提高了大约 10%。

图 3绘制 R 代码：

```
plot(z, u.d, ylim = c(0.15, 0.44))
points(z, u.JS, pch = 3)
points(z, u.mle, pch = 4)
legend("topleft",
```

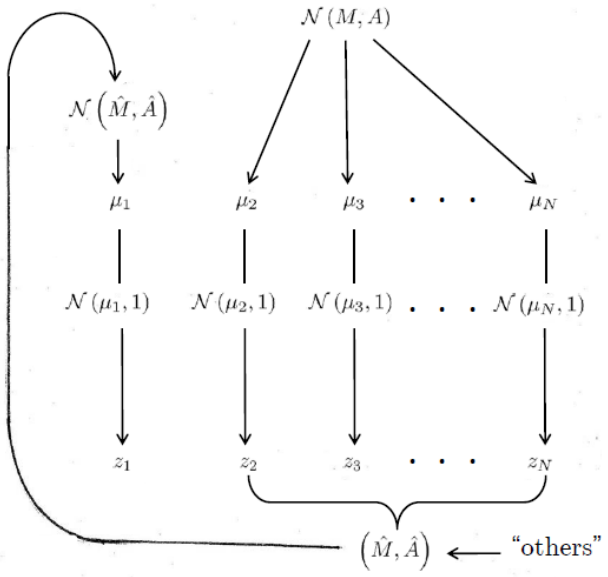


Figure 4: James-Stein 估计的流程：个体 1 的估计借用了其他个体的信息

```
legend = c有限转移估计("", "James-Stein 估计", 极大似然估计""),
pch = c(1, 3, 4))
```

4 利用其它个体的信息

贝叶斯和经验贝叶斯方法包含利用其它个体信息的思想，例如，每个棒球运动员安打率的估计要利用其它 17 个队员的信息。这里通常产生了这样一个问题：“其它个体包含哪些？”本书第十章将把这个问题回归到假设检验上, 在那里我们将面对上千个其它个体（远非 17 个）, 这大量增加了其它信息。

图 4[9] 展示了 James-Stein 估计的流程，个体 1 的估计利用了其它 (N-1) 个个体的信息。具体如下，我们假设首先观测到其它（N-1）个个体，利用这些个体信息可以得到先验分布参数的估计 $(\hat{M}, \hat{A})$，并用它们来替代（1.32）式中未知的先验分布参数（取 σ_0 为 1）。我们将得的先验分布 $\mathrm{N}(\hat{M}, \hat{A})$ 和 $z_1 \sim N(\mu_1, 1)$ 结合，利用贝叶斯公式可以估计得到 μ_1 的估计（实际上是 μ_1 的后验分布）。相同图示的其他版本应用了随后更为复杂的经验贝叶斯。

利用其他个体信息的方法并非贝叶斯分析所独有的，常用的回归方法也包含了利用其它个体信息的思想。在一项医学研究中，我们通过某种复杂的医学方法对 N=157 名健康的志愿者进行肾功能检验, 肾功能的好坏用一个得分来表示，高分代表肾功能良好，低分代表肾功能下降。我们利用分数和年龄数据绘制散点图 5，如图所示，随着年龄的增长肾功能会下降。实线为利用最小二乘法得到的肾功能分数对年龄的回归拟合直线。

现在假设有一位 55 岁肾脏捐赠者加入，我们不再使用实验设备，而是利用其它志愿者的检测数据对这位肾脏捐赠者的肾功能进行估计。如图 5所示，我们有两种方法对新加入者的肾功能得分进行估计：第一种，157 名志愿者中唯一一个 55 岁的个体（图 5中星状点）的肾功能得分，分值为

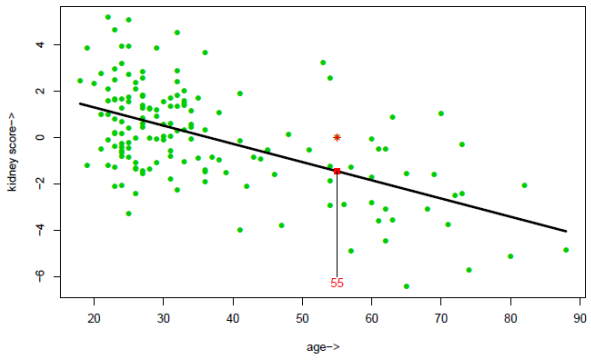


Figure 5: 肾功能与年龄

-0.01；第二种，利用最小二乘法所得到的 55 岁所对应的肾功能估计得分，分值为 -1.46。无论是频率学派还是贝叶斯学派，大多数统计学家都会采用最小二乘法。

著名统计学家 John Tukey 曾有过“借力”即“borrow strength”的提法，这句话点到了回归思想的精髓。回归的思想就是“利用其它个体的信息”，但是这一思想的运用要建立在比图 1.1 更严格的框架下。在肾功能检测的例子中，年龄是一个将志愿者与潜在的捐献者联系起来的很具有说服力的协变量。由年龄所建立起来的联系，在棒球的那个例子也扮演了重要角色。

Figure 6: John Wilder Tukey(1915-2000)

一般情况下，上述两种方法可以综合起来。我们可以把模型（1.32）拓展为：

$$\mu_i \sim N(M_0 + M_1.age_i, A) \quad 和 \quad z_i \sim N(\mu_i, \sigma_0^2)。 \tag{1.38}$$

(1.35) 式中 JS 估计量将变为下面的形式：

$$\hat{\mu}_i^{JS} = \hat{\mu}_i^{reg} + \left(1 - \frac{(N-4)\sigma_0^2}{S}\right)(z_i - \hat{\mu}_i^{reg}) \tag{1.39}$$

其中，$S = \sum(z_i - \hat{\mu}_i^{JS})^2$, $\hat{\mu}_i^{reg}$ 是 $(M_0 + M_1.age_i)$ 的线性回归估计值。在这种情况下，$\hat{\mu}_i^{JS}$ 是向线性回归估计值（即 $(M_0 + M_1.age_i)$）收缩，而不是趋向于 $\bar{z}$。

(1.39) 的求解过程如下： (1.38) 式中，$z_i \sim N(\mu_i, \sigma_0^2)$ 实质上是已知 μ_i 下 z_i 的条件分布，但

求解（1.39）式我们需要知道 z_i 的条件分布，我们可以套用 (1.3) 式求得

$$z_i \sim N(M_0 + M_1.age_i, A + \sigma_0^2),$$

所以

$$\frac{S}{A+\sigma_0^2} = \sum \left(\frac{z_i - \hat{\mu}_i^{reg}}{A+\sigma_0} \right)^2 \sim \chi^2(N-2),$$

利用卡方分布的性质可得

$$\frac{A+\sigma_0^2}{S} \sim inverse - \chi^2(N-2),$$

由逆卡方分布的期望公式得

$$E\left(\frac{A+\sigma_0^2}{S}\right) = \frac{1}{N-4},$$

进一步变换可得

$$E\left(1 - \frac{(N-4)\sigma_0^2}{S}\right) = 1 - \frac{\sigma_0^2}{A+\sigma_0^2} = \frac{A}{A+\sigma_0^2} = B.$$

即，$\left(1 - \frac{(N-4)\sigma_0^2}{S}\right)$ 是 B 的无偏估计。

练习 1.7 在肾功能检测的例子中，$S = 503$。假设 $\sigma_0^2 = 1$, 那么图 1.2 中星点处的 James-Stein 统计量是多少？（即对于一个 55 岁的健康志愿者）

练习 1.7 解答：由题干知：

$$S = 503,$$

$$\sigma_0^2 = 1。$$

由 1.4 节三、四段文字对例子的介绍可知：

$$\hat{\mu}_{53}^{reg} = -1.46,$$

$$N = 157,$$

$$z_{53} = -0.01。$$

将以上数据代入 1.39 式可得：

$$\hat{\mu}_{53}^{JS} = -0.45。$$

5 经验贝叶斯置信区间

让我们回到 1.1 节，假设有 N+1 个相互独立的正态观测值 z_i, 对应于 1.1 节我们可以得到

$$\mu_i \sim N(0, A) \quad 与 \quad z_i|\mu_i \sim N(\mu_i, 1), i = 0, 1, 2, \ldots, N+1 \tag{1.40}$$

.

现在，我们想给参数 μ_0 设定一个置信区间。在前面几小节中我们用的是点估计，例如极大似然估计、James-Stein 估计、贝叶斯估计等，虽然点估计能给出参数的具体数值，但是并不可靠，置信区间估计就弥补了这个缺点。我们可以利用图 1.1 中的经验贝叶斯信息，为参数 μ_0 设定一个经验贝叶斯置信区间，这种方法是古典置信区间方法论的进一步扩展。

如果 A 是已知的，根据 (1.10) 式我们可以得到 μ_0 的贝叶斯后验分布

$$\mu_0|z_0 \sim N(Bz_0, B) \qquad [B = A/A + 1] \tag{1.41}$$

相应地可以得到置信度为 95% 的后验区间

$$\mu_0 \in Bz_0 \pm 1.96\sqrt{B} \tag{1.42}$$

如果 A 是未知的，由于 B 是关于 A 的表达式，因此 B 也是未知的，这时我们可以用无偏估计量 $\hat{B}$ 来代替 B，运用 1.2 节的结论可得

$$\hat{B} = 1 - \frac{N-2}{S} \qquad [S = \| \mathbf{z} \|^2] \tag{1.43}$$

将其代入 (1.41) 式可以得到

$$\mu_0|z_0, \mathbf{z} \sim N(\hat{B}z_0, \hat{B}) \tag{1.44}$$

从而可得到与（1.42）式类似的置信区间 $\hat{B}z_0 \pm 1.96\sqrt{\hat{B}}$. 但是如果我们这么做的话，就会忽略 B 的估计量 $\hat{B}$ 的变异性, 这里给出了相对于 (1.44) 式更为准确的形式:

$$\mu_0|z_0, \mathbf{z} \sim \boldsymbol{N}\left(\hat{B}z_0, \hat{B} + \frac{2}{N-2}[z_0(1-\hat{B})]^2\right) \tag{1.45}$$

与此对应的后验区间为

$$\mu_0 \in \hat{B}z_0 \pm 1.96\left\{\hat{B} + \hat{B} + \frac{2}{N-2}[z_0(1-\hat{B})]^2\right\}^{\frac{1}{2}} \tag{1.46}$$

练习 1.8(a) 已知 (1.46) 式的区间长度与基于 (1.44) 式的区间长度，求其相对值即式 (1.47)。

$$\left\{1 + \hat{B} + \frac{2}{N-2}[z_0(1-\hat{B})]^2\right\}^{\frac{1}{2}} \tag{1.47}$$

(1.46) 式的区间长度:

$$2 * 1.96\sqrt{\hat{B}}$$

(b) 在 (1.47) 式中，若 N=17，B=0.21，当 $z_0 z$ 在 0 至 3 之间取值时，画出相应的图形。

练习 1.8 解答

(a) 基于 (1.44) 式的区间长度:

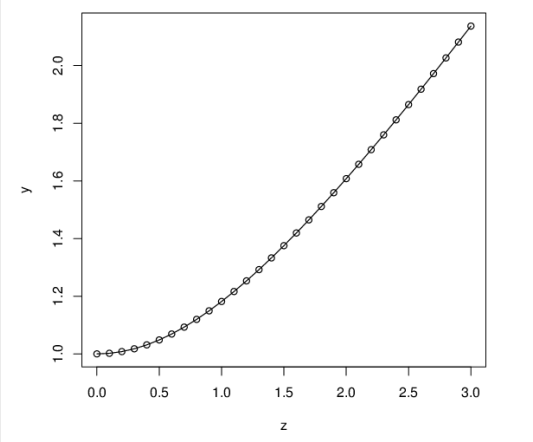
$$2 * 1.96\left\{\hat{B} + \hat{B} + \frac{2}{N-2}[z_0(1-\hat{B})]^2\right\}^{\frac{1}{2}}$$

两个式子相除便可得到 (1.47) 式。

(b) 可以用 R 软件作图：程序:

```
z=seq(from=0,to=3,by=0.1)
y=(1+2*0.79*0.79*z*z/(15*0.21))^0.5
plot(z,y)
lines(z,y)
```

图形:



参考文献

[1] Herbert Robbins. An empirical bayes approach to statistics. In *Herbert Robbins Selected Papers*, pages 41–47. Springer, 1985.

[2] Bradley Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465), 2004.

[3] Charles Stein et al. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

[4] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[5] Bradley Efron and Carl Morris. Limiting the risk of bayes and empirical bayes estimators¡ªpart ii: The empirical bayes case. *Journal of the American Statistical Association*, 67(337):130–139, 1972.

[6] Bradley Efron and Carl Morris. Stein's estimation rule and its competitors¡ªan empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[7] Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[8] Bradley Efron and Carl N Morris. *Stein's paradox in statistics*. WH Freeman, 1977.

[9] Bradley Efron. Empirical bayes methods for combining likelihoods. *Journal of the American Statistical Association*, 91(434):538–550, 1996.