12 Bayesian Regression
贝叶斯回归

用贝叶斯推断求解回归模型参数

审视数学，你会发现，它不仅是颠扑不破的真理，而且是至高无上的美丽——那种冷峻而朴素的美，不需要唤起人们任何的怜惜，没有绘画和音乐的浮华装饰，纯粹，只有伟大艺术才能展现出来的严格完美。

Mathematics, rightly viewed, possesses not only truth, but supreme beauty — a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show.

—— 伯特兰·罗素 (Bertrand Russell) | 英国哲学家、数学家 | 1872 ~ 1970

- `pymc3.Normal()` 定义正态先验分布
- `pymc3.HalfNormal()` 定义半正态先验分布
- `pymc3.plot_posterior()` 绘制后验分布
- `pymc3.sample()` 产生随机数
- `pymc3.traceplot()` 绘制后验分布随机数轨迹图

12.1 回顾贝叶斯推断

贝叶斯推断 (Bayesian inference) 把模型参数看作随机变量。在得到样本之前，根据主观经验和既有知识给出未知参数的概率分布叫做先验分布 (prior)。获得样本数据后，根据贝叶斯定理，基于给定的样本数据先计算似然分布 (likelihood)，然后模型参数的后验分布 (prior)。

上面这段文字对应如下这个公式：

$$\overbrace{f_{\Theta|X}(\theta|x)}^{\text{Posterior}} = \frac{\overbrace{f_{X|\Theta}(x|\theta)}^{\text{Likelihood}}\overbrace{f_{\Theta}(\theta)}^{\text{Prior}}}{\int_{\vartheta} f_{X|\Theta}(x|\vartheta) f_{\Theta}(\vartheta)\mathrm{d}\vartheta} \tag{1}$$

最后根据参数的后验分布进行统计推断。贝叶斯推断对应的优化问题为最大化后验概率 (Maximum A Posteriori, MAP)。本章介绍如何利用贝叶斯推断完成线性回归。

大家如果对 (1) 感到陌生的话，请回顾《统计至简》第 20、21 两章。

线性回归模型

为了配合贝叶斯推断，把多元线性回归模型写成：

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_D x_D^{(i)} \tag{2}$$

其中，i 为样本序号，D 为特征数。

当 $D = 1$ 时，一元线性回归模型为：

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} \tag{3}$$

似然

似然函数可以写成：

$$f_{\gamma|\Theta}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(i)} - \left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_D x_D^{(i)}\right)\right)^2}{2\sigma^2}\right) \tag{4}$$

这意味着残差 ε 服从 $N(0, \sigma^2)$。

贝叶斯定理

利用贝叶斯定理，我们可以得到后验分布：

$$f_{\Theta|\gamma}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) = \frac{f_{\gamma|\Theta}(\boldsymbol{y}\,|\,\boldsymbol{\theta})\cdot f_{\Theta}(\boldsymbol{\theta})}{f_{\gamma}(\boldsymbol{y})} \tag{5}$$

最大后验优化：

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} f_{\Theta|\gamma}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) \tag{6}$$

如图 1 所示，随着样本不断引入，MAP 优化结果不断接近真实参数。

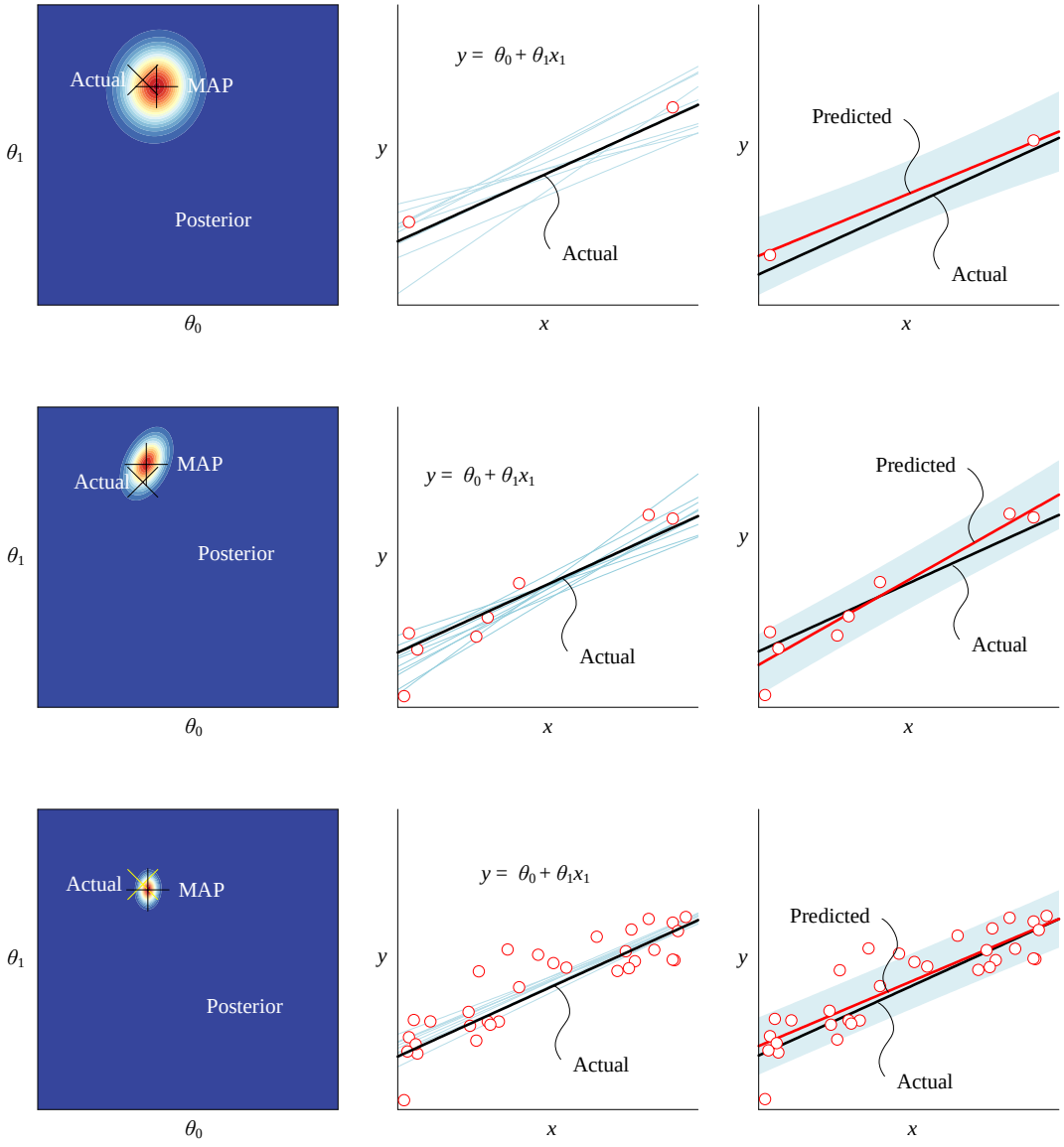

图 1. 贝叶斯回归后验概率

由于后验 ∝ 似然 × 先验，最大后验优化等价于：

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\gamma}|\Theta}\left(\boldsymbol{y}\,|\,\boldsymbol{\theta}\right)\cdot f_{\Theta}\left(\boldsymbol{\theta}\right) \tag{7}$$

为了避免算数下溢，取对数后，优化问题可以写成：

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ln\left(f_{\boldsymbol{\gamma}|\Theta}\left(\boldsymbol{y}\,|\,\boldsymbol{\theta}\right)\cdot f_{\Theta}\left(\boldsymbol{\theta}\right)\right) \tag{8}$$

从书之前介绍过，算术下溢 (arithmetic underflow) 也称为浮点数下溢 (floating point underflow)，是指计算机浮点数计算的结果小于可以表示的最小数。

(8) 进一步整理为：

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ln f_{\boldsymbol{\gamma}|\Theta}\left(\boldsymbol{y}\,|\,\boldsymbol{\theta}\right) + \ln f_{\Theta}\left(\boldsymbol{\theta}\right) \tag{9}$$

12.2 贝叶斯回归：无信息先验

《统计至简》第 20 章介绍过无信息先验 (uninformative prior)。也就是说，没有先验信息，或者先验分布不清楚，我们可以用常数或均匀分布作为先验分布，比如 $f(\boldsymbol{\theta}) = 1$。最大后验优化就可以写成：

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ln f_{\boldsymbol{\gamma}|\Theta}\left(\boldsymbol{y}\,|\,\boldsymbol{\theta}\right) \tag{10}$$

这和 MLE 的目标函数一致：

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \ln f\left(\boldsymbol{y};\boldsymbol{\theta}\right) \tag{11}$$

将 (4) 代入 $\ln f\left(\boldsymbol{y}\,|\,\boldsymbol{\theta}\right)$ 得到：

$$\begin{aligned}\ln f_{\boldsymbol{\gamma}|\Theta}\left(\boldsymbol{y}\,|\,\boldsymbol{\theta}\right) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_D x_D^{(i)}\right)\right)^2 + \underbrace{n\ln\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}} \\ &= -\frac{\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2}{2\sigma^2} + \underbrace{n\ln\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}}\end{aligned} \tag{12}$$

忽略常数，最大化后验 MAP 优化问题等价于如下最小化问题：

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 \tag{13}$$

12.3 使用 pymc3 完成贝叶斯回归

本节利用 pymc3 完成模型为 $y = \theta_0 + \theta_1 x_1$ 贝叶斯回归。如图 2 所示，黑色线为真实模型，参数为截距 $\theta_0 = 1$、斜率 $\theta_1 = 2$。图 2 中蓝色散点为样本点。

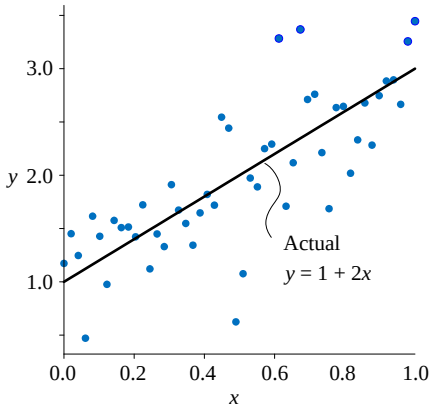

图 2. 真实模型和样本点

图 3 所示为三个参数的后验分布随机数轨迹图。随机数轨迹由 pymc3 中马尔科夫链蒙特卡洛 (Markov Chain Monte Carlo, MCMC) 生成。图中只绘制达到平稳状态后的轨迹。每个参数模拟两条轨迹。

前文提过残差 ε 服从 $N(0, \sigma^2)$，所以残差也是一个模型参数。代码中，残差的先验分布为半正态分布 (half normal distribution)，如图 4 所示。有关半正态分布，大家可以参考：

https://www.pymc.io/projects/docs/en/latest/api/distributions/generated/pymc.HalfNormal.html

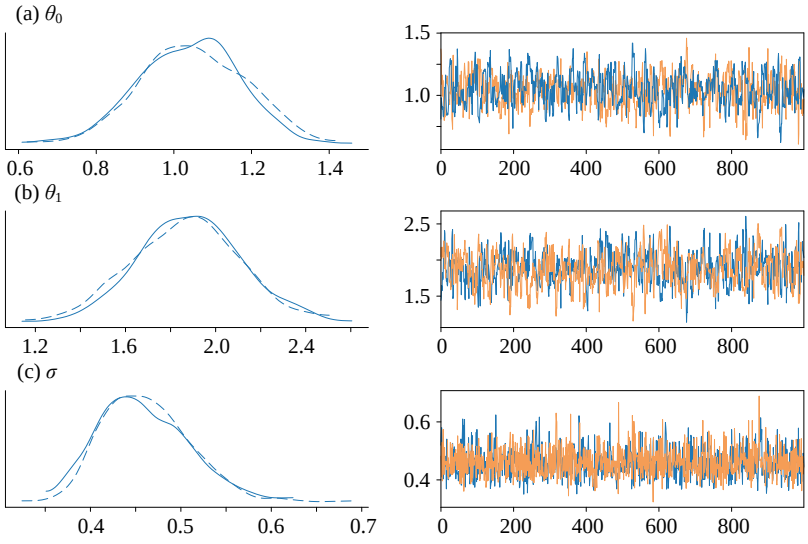

图 3. 后验分布随机数轨迹图

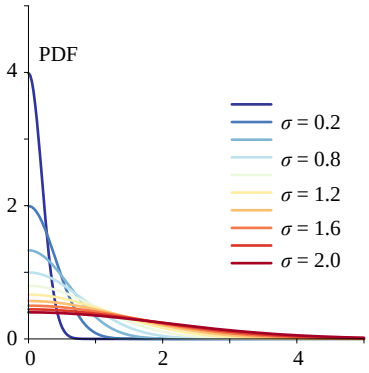

图 4. 半正态分布概率密度曲线

图 5 所示为后验分布随机数的直方图。直方图合并两条 MCMC 轨迹。图中均值可以视作 MAP 的优化解。HDI 代表最大密度区间 (highest density interval)，即后验分布的可信区间。可信区间越窄，后验分布的确信度越高。图 6 所示为参数 θ_0 和 θ_1 后验分布随机生数构成的分布。

图 7 所示为贝叶斯线性回归的结果，图中红色线为预测模型。图中的浅蓝色线为 50 条后验分布的采样函数，它们对应图 6 中的 50 个散点。红色线相当于这些浅蓝色线“取平均”。

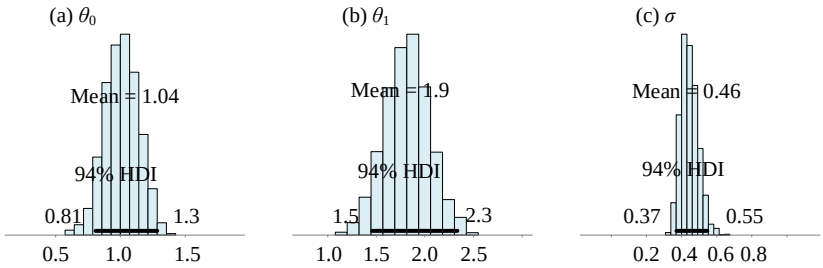

图 5. 后验分布直方图

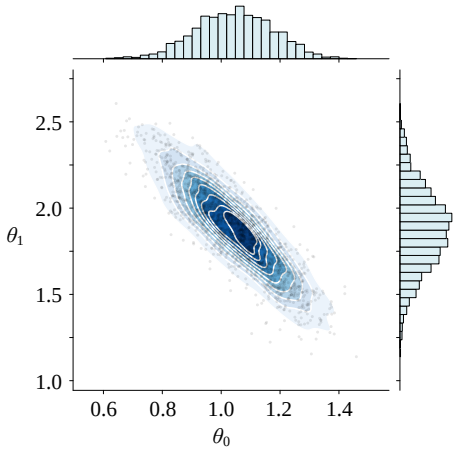

图 6. 参数 θ_0 和 θ_1 后验分布随机生数构成的分布

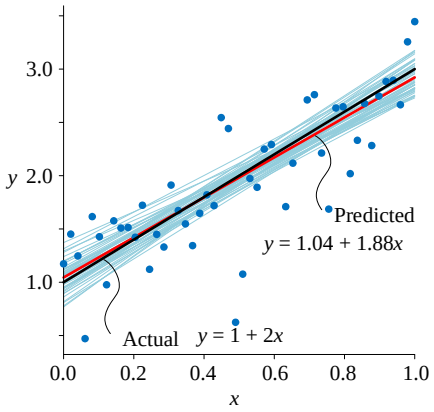

图 7. 贝叶斯线性回归结果

Bk6_Ch12_01.ipynb 绘制本节图像。

12.4 贝叶斯视角理解 Ridge 正则化

上一章的岭回归可以从贝叶斯推断角度理解。

本章中假设线性回归参数服从正态分布：

$$f_{\Theta_j}\left(\theta_j\right)=\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\theta_j^2}{2\tau^2}\right) \tag{14}$$

图 8 所示为先验分布随 τ 变化。τ 越大代表越不确信，τ 越小代表确信程度越高。

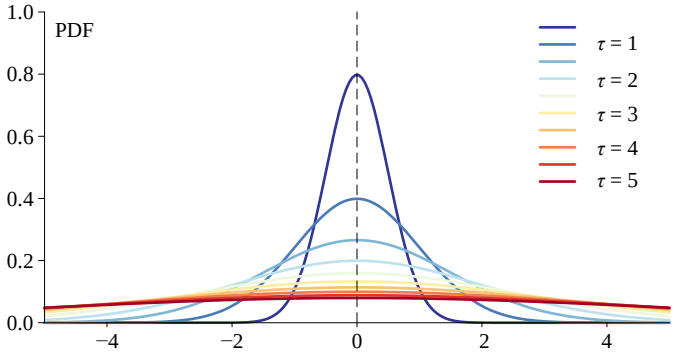

图 8. 先验分布随 τ 变化

(8) 所示的优化问题等价于：

$$\arg\max_{\theta}\left[\ln\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y^{(i)}-\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}+\cdots+\theta_Dx_D^{(i)}\right)\right)^2}{2\sigma^2}\right)+\ln\prod_{j=1}^{D}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\theta_j^2}{2\tau^2}\right)\right] \tag{15}$$

上式目标函数可以分为两部分整理。大家已经清楚，第一部分为：

$$-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}+\underbrace{n\ln\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}} \tag{16}$$

第二部分为：

$$-\frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2}+\underbrace{D\ln\frac{1}{\sqrt{2\pi\tau^2}}}_{\text{Constant}} \tag{17}$$

忽略常数后，优化问题为：

$$\arg\max_{\theta}\left[-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}-\frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2}\right] \tag{18}$$

将上式最大化问题调整为最小化问题：

$$\arg\min_{\theta}\frac{1}{2\sigma^2}\left(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2+\frac{\sigma^2}{\tau^2}\|\boldsymbol{\theta}\|_2^2\right) \tag{19}$$

令

$$\lambda=\frac{\sigma^2}{\tau^2} \tag{20}$$

(19) 等价于：

$$\arg\min_{\theta}\underbrace{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2}_{\text{OLS}}+\underbrace{\lambda\|\boldsymbol{\theta}\|_2^2}_{\text{L2 regularizer}} \tag{21}$$

这和上一章的结论完全一致。

《统计至简》第 20 章介绍过，先验的影响力很大，MAP 的结果向先验均值“收缩”。这种效果常被称作贝叶斯收缩 (Bayes shrinkage)。根据 (20)，σ 保持不变条件下，τ 越小代表确信度越高，λ 越大，通过 MAP 得到的优化解向 $\boldsymbol{0}$ (先验均值) 收缩。图 9 上可以看到，优化解随着约束项参数 λ 不断增大运动轨迹，“收缩”的这种现象显而易见。

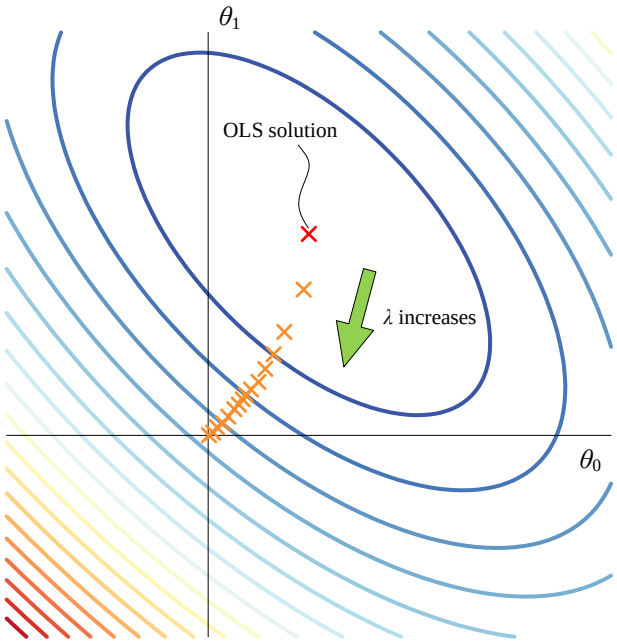

图 9. 不断增大 λ，岭回归优化解变化路径

12.5 贝叶斯视角理解套索正则化

如果先验分布为拉普拉斯分布：

$$f_{\Theta_j}\left(\theta_j\right)=\frac{1}{2b}\exp\left(-\frac{\left|\theta_j\right|}{b}\right) \tag{22}$$

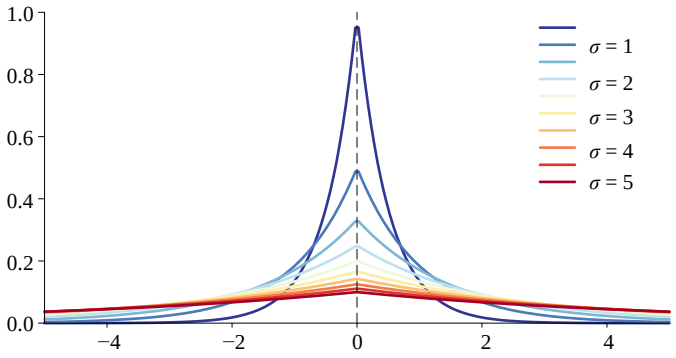

图 10. 先验分布随 b 变化

(8) 所示的优化问题等价于：

$$\arg\max_{\theta}\left[\ln\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y^{(i)}-\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}+\cdots+\theta_Dx_D^{(i)}\right)\right)^2}{2\sigma^2}\right)+\ln\prod_{j=1}^{D}\frac{1}{2b}\exp\left(-\frac{|\theta_j|}{b}\right)\right] \tag{23}$$

也是分两部分来看上式。第一部分和上一节完全相同：

$$-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}+\underbrace{n\ln\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}} \tag{24}$$

第二部分为：

$$-\frac{1}{b}\sum_{j=1}^{D}|\theta_j|+\underbrace{D\ln\frac{1}{2b}}_{\text{Constant}}=-\frac{1}{b}\|\boldsymbol{\theta}\|_1+\underbrace{D\ln\frac{1}{2b}}_{\text{Constant}} \tag{25}$$

忽略常数后，优化问题为：

$$\arg\max_{\theta}-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}-\frac{1}{b}\|\boldsymbol{\theta}\|_1 \tag{26}$$

最大化问题调整为最小化问题得到：

$$\arg\min_{\theta}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2+\frac{2\sigma^2}{b}\|\boldsymbol{\theta}\|_1 \tag{27}$$

令

$$\lambda=\frac{2\sigma^2}{b} \tag{28}$$

(27) 等价于

$$\arg\min_{\theta}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2+\lambda\|\boldsymbol{\theta}\|_1 \tag{29}$$

这和上一章套索回归的优化问题的目标函数本质上一致。

图 11 所示为不断增大 λ，套索回归参数变化轨迹；可以发现参数变化轨迹有两段，第一段从 OLS 结果为起始点，几乎沿着斜线靠近 y 轴 ($\theta_0 = 0$)，直至到达 y 轴。到达 y 轴时，回归系数 θ_0 为 0。第二段，沿着 y 轴朝着原点运动。

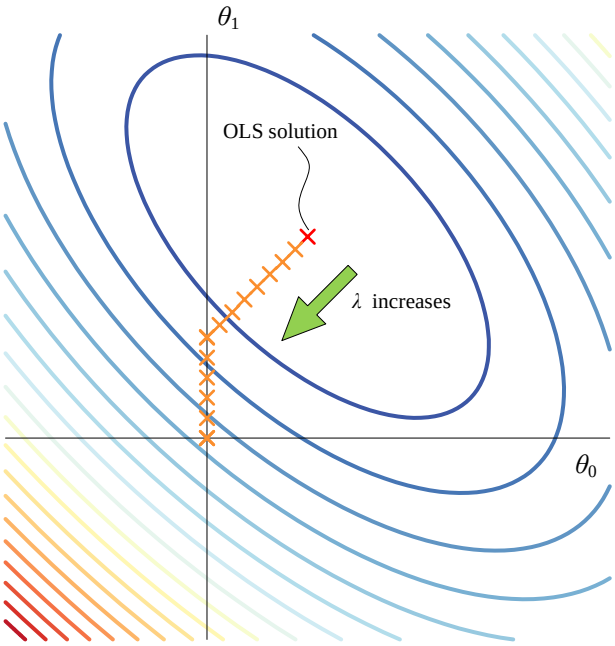

图 11. 不断增大 λ，套索回归优化解变化轨迹

想深入学习贝叶斯推断和贝叶斯回归的读者可以参考开源图书 *Bayesian Modeling and Computation in Python*：

https://bayesiancomputationbook.com/welcome.html