arXiv:2302.07012v1 [math.ST] 14 Feb 2023

带正则化高斯分布的稀疏贝叶斯推断*

Jasper M. Everink†、Yiqiu Dong†、和 Martin S. Andersen†

摘要。 正则化是变分逆问题中的常用工具，用于对问题参数施加假设。其中一种假设是稀疏性，通常通过套索和类似全变差的正则化来促进。尽管许多此类正则化逆问题的解可以视为精心选择的后验分布的概率最大点，但从这些分布中抽取的样本通常并不稀疏。本文提出一个框架，用于隐式定义一种概率分布，将促进稀疏性的正则化与高斯分布的作用结合起来。与连续分布不同，这些隐式分布可以为稀疏向量赋予正概率。我们研究了采用各种正则化函数的这些正则化分布，包括全变分正则化和分段线性凸函数。我们将所发展的理论应用于贝叶斯线性逆问题的不确定性量化，并为贝叶斯层次模型推导出一个吉布斯采样器。为展示我们这种诱导稀疏性的框架与连续分布之间的差异，我们将该框架应用于小规模去模糊和计算机断层扫描示例。

关键词： 贝叶斯推断、正则化、逆问题、不确定性量化**AMS 学科分类：** 62F15, 65C05, 90C25

1. 引言。 一种常见的信号重建方法是 $\boldsymbol{x} \in \mathbb{R}^n$ 根据含噪测量值 $\boldsymbol{b} = A\boldsymbol{x} + \boldsymbol{e}$，其中 $A \in \mathbb{R}^{m\times n}$ 是线性前向算子，$\boldsymbol{e} \in \mathbb{R}^m$ 是噪声；该方法通过求解以下形式的正则化线性最小二乘问题来实现：

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + f(\boldsymbol{x}) \right\}, \tag{1.1}$$

其中，函数 f 的选取旨在使优化问题适定，并通过惩罚不需要的行为来改善重建效果。一种常见的选择是通过选取 l_1-基正则化函数来促进重建结果的稀疏性，其形式为 $f(\boldsymbol{x}) = \gamma\|D\boldsymbol{x}\|_1$，其中 $\gamma > 0$ 表示正则化强度，D 是一个线性算子，表示重建结果应当具有稀疏性的基。这类正则化函数已应用于多个领域，例如采用基追踪算法的压缩感知 [7] 以及采用全变分正则化的图像处理 [9]。

使用（1.1）重建信号的依据是，其解是适当后验分布的最大后验（MAP）估计，即：

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + f(\boldsymbol{x}) \right\} = \operatorname*{argmax}_{\boldsymbol{x}\in\mathbb{R}^n} \pi(\boldsymbol{x} \mid \boldsymbol{b}), \tag{1.2}$$

* **资助：** 本研究获得 Villum 基金会（资助编号：25893 和 VIL50096）及诺和诺德基金会（资助编号：NNF20OC0061894）的支撑。†丹麦技术大学应用数学与计算机科学系。Richard Petersens Plads，324 号楼，DK-2800 孔恩斯灵比，丹麦。（jmev@dtu.dk、yido@dtu.dk、mskan@dtu.dk）

其中，

$$\pi(\boldsymbol{x} \,|\, \boldsymbol{b}) \propto \pi(\boldsymbol{b} \,|\, \boldsymbol{x})\pi(\boldsymbol{x}) \propto \exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 - f(\boldsymbol{x})\right). \tag{1.3}$$

似然函数 $\pi(\boldsymbol{b} \,|\, \boldsymbol{x}) \propto \exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2\right)$ 是在假设误差$\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0},\ \lambda^{-1}I)$服从均值为零、协方差为 $\lambda^{-1}I$ 的正态分布时得到的，而先验 $\pi(\boldsymbol{x}) \propto \exp(-f(\boldsymbol{x}))$ 则描述了我们对重建所作的其他假设。

尽管对于合适的正则化函数 f，最大后验估计（1.2）可以保证具有稀疏性，但相应的后验分布（1.3）对这些稀疏解赋予的概率为零。这是因为稀疏性由低维子空间表示，而连续概率分布（如（1.3））总是对这些子空间赋予零概率。

一种构造对低维子空间赋予正概率的概率分布的方法是使用变维模型，见 [14]。在这些方法中，会构造多种模型 M_i，每种模型都有各自的概率分布 $\pi(\boldsymbol{x} \,|\, M_i)$，其支撑可能位于低维子空间上。随后，这些模型分布会与模型先验分布 $\pi(M_i)$结合。对于稀疏性，每种模型分布的支撑都位于一个稀疏子空间上，而模型先验分布则对应于对重建的稀疏性指定一个先验分布。

从这些变维模型中采样颇具挑战性。采样此类分布的一种常见方法是可逆跳跃马尔可夫链蒙特卡罗（RJMCMC）方法 [8]。这种方法需要大量调参，在高维问题中也容易失效。另一种方法是收缩阈值梅特罗波利斯调整朗之万算法（STMALA） [17] ，它基于 MALA，为类似 RJMCMC 的算法提供了一种高效的提议分布。尽管这种方法在规模较大的问题上效果更好，但仍需要进行困难的参数调优，而且很容易产生高度相关的样本。

在 [5], 中，我们证明，对于闭凸集 $\mathbf{C} \subset \mathbb{R}^n$，受约束的线性最小二乘问题

$$\operatorname*{argmin}_{\boldsymbol{x} \in \mathbf{C}} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x} - \hat{\boldsymbol{b}}\|_2^2 + \frac{\delta}{2}\|L\boldsymbol{x} - \hat{\boldsymbol{c}}\|_2^2 \right\}, \tag{1.4}$$

其中 $\mathrm{Null}(A) \cap \mathrm{Null}(L) = \{\boldsymbol{0}\}$，$\hat{}\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{b}, \lambda^{-1}I)$ 且 $\hat{}\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{c}, \delta^{-1}I)$，等同于从高斯后验分布中采样

$$\pi(\boldsymbol{x} \,|\, \boldsymbol{b}) \propto \pi(\boldsymbol{b} \,|\, \boldsymbol{x})\pi(\boldsymbol{x}) \propto \exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 - \frac{\delta}{2}\|L\boldsymbol{x} - \boldsymbol{c}\|_2^2\right), \tag{1.5}$$

并按照范数 $\mathbf{C}$将样本投影到约束集 $\|\cdot\|_{\lambda A^T A + \delta L^T L}$上。此外，如果约束集$\mathbf{C}$ 是多面体的，那么得到的投影高斯分布会以正概率分配给 $\mathbf{C}$的各个低维面。该投影高斯分布的结构与变维模型类似，并能产生独立样本；然而，对应的概率分布是隐式的，而且对于每个样本，都必须精确求解一个优化问题（1.4）。

这种方法与其他基于优化的采样方法类似，例如用于从高斯分布中采样的扰动优化 [12] 以及先随机化再优化

(RTO) [19]，用于非线性逆问题。然而，这些方法仅适用于从明确描述的连续概率分布中采样，而我们的方法所产生的概率分布通常没有明确的表达式。

本文将通过研究以下优化问题，对 [5] 中关于（1.4）的大部分结果进行推广：

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2+\frac{\delta}{2}\|L\boldsymbol{x}-\hat{\boldsymbol{c}}\|_2^2+f(\boldsymbol{x})\right\}, \tag{1.6}$$

其中， f 是促进稀疏性的正则化。此前已有研究考察了 $f(\boldsymbol{x})=\|D\boldsymbol{x}\|_1$的情形，此时 D 是对角矩阵，见 [6]。我们主要考虑凸分段线性函数 f这一类。该类包括基于 l_1-范数的正则化函数，以及 [5]中考虑的多面体约束。我们对通过正则化随机线性最小二乘问题（1.6）得到的分布进行了刻画。我们将所得性质应用于贝叶斯线性逆问题和层次模型，并通过数值实验进行检验。

本文结构如下。第2节利用近端算子将问题（1.6）转化为一个通用框架，并研究由此得到的正则化高斯分布的性质。第3节将应用这一理论研究贝叶斯线性逆问题，第4节将为贝叶斯层次模型推导一种算法。最后，第5节介绍基于这些模型的数值实验。

2. 正则化多元分布。我们讨论一种理论，用于将正则化的影响结合到概率分布中。在第2.1小节中，我们给出正则化高斯分布的定义，并介绍正则化分布的一些一般理论。在第2.2小节中，我们将重点讨论与高斯分布结合时会引入稀疏性的正则化函数。

2.1. 一般理论。若 $A\in\mathbb{R}^{m\times n}$ 的秩为 n，则具有随机数据 $\hat{\boldsymbol{b}}\sim\mathcal{N}(\boldsymbol{b},\ \lambda^{-1}I)$ 的正则化最小二乘问题可变换如下：

$$\begin{aligned}
&\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2+f(\boldsymbol{x})\right\}\\
&=\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\boldsymbol{x}-(A^TA)^{-1}(A^T\hat{\boldsymbol{b}})\|_{\lambda A^TA}^2+f(\boldsymbol{x})\right\}\\
&=\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}^\star\|_{\Sigma^{-1}}^2+f(\boldsymbol{x})\right\},
\end{aligned} \tag{2.1}$$

其中 $\boldsymbol{x}^\star=(A^TA)^{-1}(A^T\hat{\boldsymbol{b}})$ 服从密度为

$$\pi_{\boldsymbol{x}^\star}(\boldsymbol{x})\propto\exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2\right), \tag{2.2}$$

和协方差 $\Sigma=(\lambda A^TA)^{-1}$。由于这一变换，我们可以研究一般情形下的优化问题（2.1），其中 $\boldsymbol{x}^\star$ 服从任意高斯分布。这引出了以下假设，以及本节研究重点对象的定义。

假设 2.1。设 $\boldsymbol{x}^\star$ 是满足 $\boldsymbol{x}^\star \sim \mathcal{N}(\boldsymbol{\mu},\ \Sigma)$ 均值为$\boldsymbol{\mu} \in \mathbb{R}^n$ 、协方差矩阵为 $\Sigma \in \mathbb{S}^n_{++}$的高斯随机向量，其中 $\mathbb{S}^n_{++}$ 表示阶数为 n的实对称正定矩阵集合。

定义 2.2。 在假设 *2.1*成立时，若 $f\colon\ \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是一个真下半连续凸函数，则经 $\mathcal{N}(\boldsymbol{\mu}$ 正则化的高斯分布 $\Sigma)$ 定义为 f如下形式

$$\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) := \underset{\boldsymbol{z}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{x}^\star\|^2_{\Sigma^{-1}} + f(\boldsymbol{z}) \right\}, \tag{2.3}$$

其中，近端算子$\operatorname{prox}_f^{\Sigma^{-1}}$ 是相对于范数 $\|\cdot\|_{\Sigma^{-1}}$的近端算子。

由于它与正则化线性最小二乘的关系，本文重点研究正则化高斯分布，但近端算子可应用于任意概率分布。若 π 是 $\mathbb{R}^n$上的任意概率测度，则根据近端算子 $\operatorname{prox}_f^{\Sigma^{-1}}$ 的连续性，推前测度 $\pi\circ\left[\operatorname{prox}_f^{\Sigma^{-1}}\right]^{-1}$ 是$\mathbb{R}^n$上定义良好的概率测度。此外，本文所呈现的一些结果并不依赖于 π 为高斯分布，因此很容易推广到其他分布。

对于大多数正常下半连续凸函数 f，我们无法计算 $\operatorname{prox}_f^{\Sigma^{-1}}$ 的显式表达式；不过，对于一些简单的例子，确实存在显式表达式。令 $f(\boldsymbol{z}) = \frac{1}{2}\|D\boldsymbol{z}-\boldsymbol{d}\|_2^2$，其中任意 $D \in \mathbb{R}^{k\times n}$ 和 $\boldsymbol{d} \in \mathbb{R}^k$ 均成立，即广义蒂霍诺夫正则化。对于这种正则化，正则化高斯分布为

$$\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) = (\Sigma^{-1} + D^TD)^{-1}(\Sigma^{-1}\boldsymbol{x}^\star + D^T\boldsymbol{d}),$$

它仍然是一个高斯分布。然而，一般而言，正则化高斯分布不再是高斯分布。有关正则化高斯分布显式描述的一个更复杂示例，请参见 [17, 引理 II.2]。令 $f(\boldsymbol{x}) = \chi_{\mathbf{C}}(\boldsymbol{x})$ 为闭凸集 $\mathbf{C}$ $\subset \mathbb{R}^n$ 的示性函数，即 $f(\boldsymbol{x})$ 在 $\mathbf{C}$ 上为零，在其他地方为无穷大。则 $\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) = \Pi^{\Sigma^{-1}}$ $_{\mathbf{C}}(\boldsymbol{x}^\star)$，其中 $\Pi^{\Sigma^{-1}}$ $_{\mathbf{C}}(\boldsymbol{x})$ 表示将 $\boldsymbol{x}$ 投影到 $\mathbf{C}$ 上的斜投影。在这种情况下，$\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star)$ 在 $\mathbf{C}$ 的内部具有与 $\boldsymbol{x}^\star$ 相同的密度，但 $\mathbf{C}$ 外部的所有质量都会被转移到 $\mathbf{C}$ 的边界上。我们在 [5] 中专门研究了这种投影情形。

In 一般而言，以下命题描述了正则化高斯分布的支撑 .

命题 2.3。 在假设 *2.1* 下，若 $f\colon\ \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是适当的下半连续凸函数，则有：

$$\operatorname{supp}(\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star)) = \operatorname{cl}(\operatorname{dom}(f)),$$

其中$\operatorname{cl}(\mathbf{C})$ 表示$\mathbf{C}$ 的闭包，$\operatorname{supp}(\boldsymbol{z}^\star)$ 表示随机向量$\boldsymbol{z}^\star$的支撑，即任何包含该点的开集都具有正概率，随机向量 $\boldsymbol{z}^\star$亦是如此。

证明。 利用引理 (近端算子$\operatorname{prox}_f^{\Sigma^{-1}}) = \operatorname{dom}(f)$ 中的值域 B.1。 ■

注意，对于 $f(\boldsymbol{x}) = \chi_{\mathbf{C}}(\boldsymbol{x})$ 投影 $\Pi^{\Sigma^{-1}}$ $_{\mathbf{C}}$ 会将约束集 $\mathbf{C}$ 之外的所有样本映射到边界 $\operatorname{bd}(\mathbf{C})$，而该集合是勒贝格零测集。

此外，可以证明，[5, 引理 2.4]表明 $\Pi^{\Sigma^{-1}}{}_{\mathbf{C}}(\boldsymbol{x}^\star)$ 在满足 $\mathbf{C} \neq \mathbb{R}^n$时，会对 $\mathrm{bd}(\mathbf{C})$ 赋予正概率，因此 $\Pi^{\Sigma^{-1}}{}_{\mathbf{C}}(\boldsymbol{x}^\star)$ 相对于勒贝格测度没有密度，因而不是连续随机变量。下述命题给出了 f 的光滑性条件，使正则化高斯近端算子$\mathrm{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star)$ 成为连续随机变量。

命题 2.4。 在假设 *2.1*下，设 $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$，且 $\mathrm{dom}(f)$ 为开集。若 f是定义在 $\mathrm{dom}(f)$ 上的处处可微凸函数，且具有局部 *Lipschitz* 连续梯度，则 近端算子$\mathrm{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star)$ 是定义在 $\mathrm{dom}(f)$上的连续分布。

证明。 注意，与 $\boldsymbol{z}$ =近端算子$\mathrm{prox}_f^{\Sigma^{-1}}(\boldsymbol{x})$ 相关的最优性条件为

$$\boldsymbol{x} \in \boldsymbol{z} + \Sigma \partial f(\boldsymbol{z}), \tag{2.4}$$

并因 此，近端算子的逆像可表示为如下集值映射

$$\left[\mathrm{prox}_f^{\Sigma^{-1}}\right]^{-1}(\boldsymbol{z}) = \boldsymbol{z} + \Sigma \partial f(\boldsymbol{z}). \tag{2.5}$$

因此，根据假设，$\left[\text{近端算子}_f^{\Sigma^{-1}}\right]^{-1}(\boldsymbol{z}) = \boldsymbol{z} + \Sigma \nabla f(\boldsymbol{z})$ 是一个局部 Lipschitz 连续的单值映射，因此结论可由推论 B.3 得出。 ■

命题 2.4表明，如果正则化函数 f 足够光滑，那么正则化高斯分布具有连续分布。该命题未考虑正则化函数可微但不具有局部 Lipschitz 连续梯度的情形[1]。对于许多不可微函数，我们可以证明其逆命题。以下定理给出了一个关于函数的条件，使得与其对应的正则化分布不连续。

定理 2.5。 在假设 *2.1*下，设 $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是一个适当的下半连续凸函数。此外，设 $\mathbf{U} \subset \mathbb{R}^k$ 是紧集，且 $s\colon \mathbf{U} \to s(\mathbf{U}) \subseteq \mathbb{R}^n$是一个同胚。此外，设 $\mathbf{V} \subset \mathbb{R}^{n-k}$ 是紧集，且 $d\colon \mathbf{U} \times \mathbf{V} \to d(\mathbf{U}, \mathbf{V}) \subseteq \mathbb{R}^n$在 $\boldsymbol{v}$ 上连续且可逆，且对于任意 $\boldsymbol{u} \in \mathbf{U}$，

$$d(\boldsymbol{u}, \mathbf{V}) \subseteq \partial f(s(\boldsymbol{u})).$$

那么，

$$\mathbb{P}\left(\mathrm{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) \in s(\mathbf{U})\right) > 0.$$

证明。 注意，事件 $\{\text{近端算子}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) \in s(\mathbf{U})\}$ 等价于事件 $\{\boldsymbol{x}^\star \in (\mathrm{Id} + \Sigma\partial f)(s(\mathbf{U}))\}$。由于 $\mathrm{supp}(\boldsymbol{x}^\star) = \mathbb{R}^n$，只需证明集合 $(\mathrm{Id} + \Sigma\partial f)(s(\mathbf{U}))$包含一个开集。

根据假设，集合 $(\mathrm{Id} + \Sigma\partial f)(s(\mathbf{U}))$ 包含映射的像

$$h : \mathbf{U} \times \mathbf{V} \to \mathbb{R}^n, \quad \text{defined by} \quad h(\boldsymbol{u}, \boldsymbol{v}) = s(\boldsymbol{u}) + \Sigma d(\boldsymbol{u}, \boldsymbol{v}).$$

[1]命题2.4 可以推广到可微凸函数 f 使得 $\boldsymbol{z} \mapsto \boldsymbol{z} + \Sigma\nabla f(\boldsymbol{z})$ 保持勒贝格可测性。

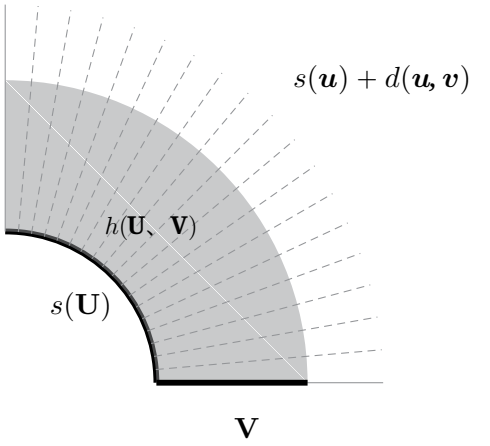

图 2.1：四分之一圆盘曲线边界的定理 2.5可视化。

若 $\boldsymbol{u}, \boldsymbol{u}' \in \mathbf{U}$ 和 $\boldsymbol{v}, \boldsymbol{v}' \in \mathbf{V}$ 满足 $h(\boldsymbol{u}, \boldsymbol{v}) = h(\boldsymbol{u}', \boldsymbol{v}')$。则

$$s(\boldsymbol{u}) = \operatorname{prox}_f^{\Sigma}(h(\boldsymbol{u}, \boldsymbol{v})) = \operatorname{prox}_f^{\Sigma}(h(\boldsymbol{u}', \boldsymbol{v}')) = s(\boldsymbol{u}'),$$

因此，由于 s 是同胚，$\boldsymbol{u} = \boldsymbol{u}'$。此外，由于 $s(\boldsymbol{u}) + \Sigma d(\boldsymbol{u}, \boldsymbol{v}') = s(\boldsymbol{u}) + \Sigma d(\boldsymbol{u}, \boldsymbol{v}')$ 且 d 在 $\boldsymbol{v}$中可逆，我们得到 $\boldsymbol{v} = \boldsymbol{v}'$。因此，$h$是 n维集合 $\mathbf{U} \times \mathbf{V}$ 与 $(\mathrm{Id} + \Sigma \partial f)(s(\mathbf{U}))$的一个子集之间的同胚。因此，我们可以得出结论，$(\mathrm{Id} + \Sigma \partial f)(s(\mathbf{U}))$ 包含一个开集。 ■

定理 2.5 涵盖了大多数值得关注的情形，包括大多数稀疏性正则化函数，正如第 2.2 小节将讨论的那样。作为另一个例子，设 C 为四分之一圆盘，即单位圆盘与非负正交象限的交集，并考虑 $f(\boldsymbol{x}) = \chi_{\mathbf{C}}(\boldsymbol{x})$，也就是说，将线性最小二乘问题限制在四分之一圆盘 C 上。f 的次微分是法锥 $\partial f(\boldsymbol{x}) = N_{\mathbf{C}}(\boldsymbol{x})$。为简便起见，考虑 $\Sigma = I$，并令 $s(\mathrm{U})$ 为边界的一维曲线部分，如图 2.1 所示。f 在 s 上的次微分可参数化为 $d(u、v) = vs(u)$，其中 $v \geqslant 0$。来自 $s(\boldsymbol{u}) + \mathrm{N}_{\mathbf{C}}(s(\boldsymbol{u}))$的所有质量，包括将 v限制到 $[0、1]$后得到的子集 $h(\mathrm{U}, \mathrm{V})$，都会投影到 $s(\mathrm{U})$上，从而使 $\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star)$ 为 $s(\mathrm{U})$赋予正概率。

2.2. 稀疏性。 尽管定理 2.5 适用于任意曲面，但通常关注的是线性子空间。特别是，如果我们感兴趣的是位于由线性方程组 $D\boldsymbol{x} = \boldsymbol{d}$定义的子空间中的解，就会添加正则化函数 $\|D\boldsymbol{x} - \boldsymbol{d}\|$。该范数在零点处不可微，因此正则化函数在指定子空间中不可微。下面的定理表明，对于这类正则化函数，正则化高斯分布在这些子空间上具有正概率。

Theo第 2.6 节。设 $p \geqslant 1$，$D_i \in \mathbb{R}^{c_i \times n}$和 $\boldsymbol{d}_i \in \mathbb{R}^{c_i}$对于 $i = 1, \ldots, k$。然后定义一个适当的 r

连续凸函数 $f\colon\ \mathbb{R}^n \to \mathbb{R}$，定义为

$$f(\boldsymbol{x}) := \sum_{i=1}^{k} \|D_i\boldsymbol{x} - \boldsymbol{d}_i\|_p. \tag{2.6}$$

对于任意 $\mathbf{I} \subseteq \{1,\dots,k\}$，令 $\mathbf{F_I}$ 为 $\{\boldsymbol{x} \in \mathbb{R}^n \mid D_i\boldsymbol{x} - \boldsymbol{d}_i = 0, \forall i \in \mathbf{I}$ 的非空紧子集，且 $D_i\boldsymbol{x} - \boldsymbol{d}_i \neq 0, \forall i \notin \mathbf{I}\}$。那么，

$$\mathbb{P}\left(\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) \in \mathbf{F_I}\right) > 0.$$

证明。 首先假设 $p > 1$。对于 $\boldsymbol{x} \in \mathbf{F_I}$，$f$ 的次微分可分解如下。映射 $\boldsymbol{x} \mapsto \|D_i\boldsymbol{x} - \boldsymbol{d}_i\|_p$ 对任意 $i \notin \mathbf{I}$ 均可微。对于 $i \in \mathbf{I}$，$\boldsymbol{x} \mapsto \|D_i\boldsymbol{x} - \boldsymbol{d}_i\|_p$不可微，其次梯度为 $D_i^T\partial(\|\cdot\|_p)(\mathbf{0})$，其中 $\partial(\|\cdot\|_p)(\mathbf{0})$ 是一个 c_i 维单位 l_p 范数球，以 $\mathbf{0}$为中心。因此，我们可以将 f 的次梯度拆分为在 $\mathbf{F_I}$ 上为常数的集合和平移连续项，即：

$$\partial f(\boldsymbol{x}) = \mathbf{E} + \nabla g(\boldsymbol{x}), \tag{2.7}$$

with

$$\mathbf{E} = \sum_{i\in\mathbf{I}} D_i^T\partial(\|\cdot\|_p)(\mathbf{0}) \quad \text{and}$$
$$g(\boldsymbol{x}) = \sum_{i\notin\mathbf{I}} \|D_i\boldsymbol{x} - \boldsymbol{d}_i\|_p,$$

其中前一个和是闵可夫斯基和。注意$\mathbf{E}$ 与$\mathbf{F_I}$ 且 $\dim(\mathbf{E}) + \dim(\mathbf{F_I}) = n$。因此，结论由定理2.5得出。

对于情形 $p = 1$，注意任意标量 $a \in \mathbb{R}$ 都满足 $\|a\|_2 = |a|$。因此，对于任意 $\boldsymbol{z} \in \mathbb{R}^d$，我们有 $\|\boldsymbol{z}\|_1 = \sum_{j=1}^d \|\boldsymbol{z}_j\|_2$，其中 $\boldsymbol{z}_j$ 表示 $\boldsymbol{z}$ 的各个分量。

形如 (2.6) 的函数示例包括各向（非）同性全变差等正则化函数。然而，如果分解 (2.7) 中的可微部分 g 为零，则可以得到更多结论。以下假设将正则化函数限制为这样一类函数：不同的、可能是低维的子空间具有恒定的次微分。

假设 2.7。 设 $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是任意一个适当函数，其上图为多面体集合；等价地，f 是一个凸分段线性函数 *[15, 定理 2。49]*。注意，任何此类函数都可以写成

$$f(\boldsymbol{x}) = \chi_{\mathbf{C}}(\boldsymbol{x}) + \max_{i\in\mathbf{I}}\{\boldsymbol{a}_i^T\boldsymbol{x} + b_i\},$$

其中，$\mathbf{C} \subseteq \mathbb{R}^n$ 是一个多面体集合，$\mathbf{I}$ 是一个有限指标集，而各对 $(\boldsymbol{a}_i、\ b_i)$ 定义仿射函数。注意，$\mathbf{C}$ 是 f 的有效定义域。

满足假设2.7 的函数示例包括多面体约束，如非负约束（$f(\boldsymbol{z}) = \chi_{\mathbb{R}^n_+}(\boldsymbol{z})$），以及 l_1-范数型正则化，如各向异性全变分（$f(\boldsymbol{z}) = \gamma\|D\boldsymbol{z}\|_1$）和某些 Besov 范数 [19]（$f(\boldsymbol{z}) = \gamma\|WB\boldsymbol{z}\|_1$）。请注意，各向同性全变分以及更一般的组稀疏（$f(\boldsymbol{z}) = \sum_i \|D_i\boldsymbol{z}\|_2$）不属于此类。

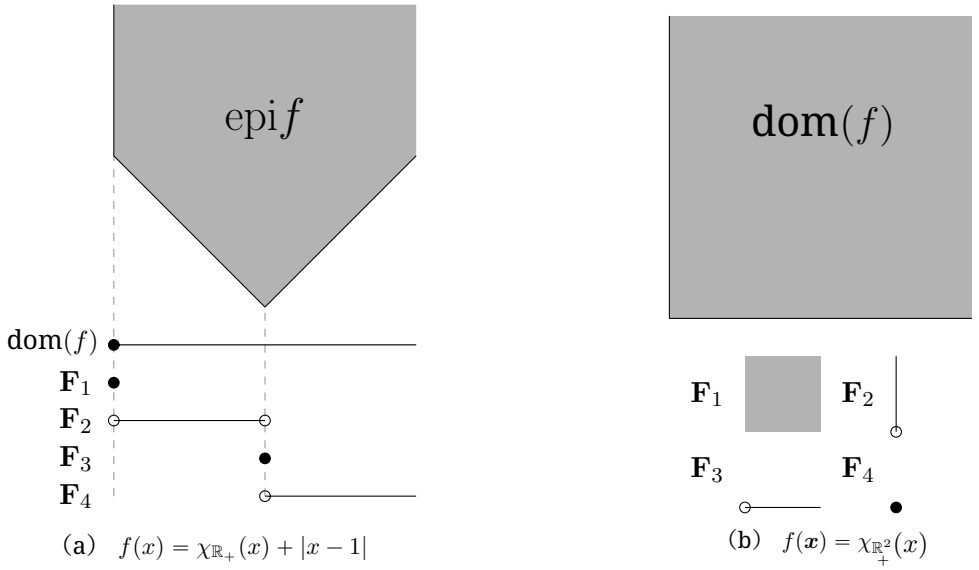

图 2.2：一维和二维中的多面体划分示例。

以下定义给出了正则化函数定义域的一个划分，使得每个元素都对应于一个可能感兴趣的低维子空间的一部分。

定义 *2.8*。在假设 *2.7* 下，我们定义 $\boldsymbol{dom}(f)$ 的多面体划分 $\{F_j\}_{j\in\mathbf{J}}$ 如下。记 $\hat{F}_j \subset \mathbb{R}^{n+1}$ 为多面体上图 $\boldsymbol{epi}(f)$ 的面，使得 $(\boldsymbol{x}, t) \in \hat{F}_j\hat{}$ 蕴含 $f(\boldsymbol{x}) = t$。记 F_i 为将投影 $(\boldsymbol{x}, t) \mapsto \boldsymbol{x}$ 应用于 $\boldsymbol{relint}(F_i)$所得的集合，其中 *relint* 表示相对内部。

等价地，$\{\mathbf{F}_j\}_{j\in\mathbf{J}}$ 是通过次微分的等价关系对 $\mathrm{dom}(f)$ 进行的划分，即对于所有 $j \in \mathbf{J}$，当且仅当 $\partial f(\boldsymbol{x}) = \partial f(\boldsymbol{y})$ 时，有 $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{F}_j$。

图 2.2 展示了这一多面体划分。请注意，在某些情况下，该划分包含许多低维集合。我们将证明，正则化高斯分布为多面体划分中的每个集合都赋予正概率。此外，正则化高斯分布可以表示为定义在每个集合上的、可能为低维的密度的混合，并且这些密度除平移外与原始高斯分布成比例。下面的引理描述了一个正交性，该正交性是由于平移分量 ∇g在（2.7）中为零而成立的。

引理 *2.9.* 考虑定义 *2.8* 中的多面体划分 $\{F_j\}_{j\in\mathbf{J}}$。对于 $\boldsymbol{x} \in F_i$，次微分 $\partial f(\boldsymbol{x})$ 不依赖于 $\boldsymbol{x}$，因此记为 $\partial f_{\mathbf{F}_j}$。此外，将 F_j 参数化为 $\boldsymbol{x}_j + F_j\boldsymbol{u}$，其中固定 $\boldsymbol{x}_j \in F_j$、$F_j \in \mathbb{R}^{n\times\dim(\mathbf{F}_j)}$，且局部坐标为 $\boldsymbol{u} \in U \subseteq \mathbb{R}^{\dim(\mathbf{F}_j)}$。那么，存在 $\boldsymbol{d} \in \{\mathbf{0}、\boldsymbol{a}_1、\ldots\ldots、\boldsymbol{a}_{|\mathbf{I}|}\}$，使得 $(\partial f_{\mathbf{F}_j} - \boldsymbol{d})^T F_j\boldsymbol{u} = \{0\}$。

ˆ证明。 对于每个 $\mathbf{F}_j$，都存在一个最大的非空活跃仿射函数集合 $\mathbf{I} \subseteq\mathbf{I}$，使得 $f|_{\mathbf{F}_j}(\boldsymbol{x}) = \boldsymbol{a}_i^T\boldsymbol{x} + b_i$对所有 $i \in \hat{\mathbf{I}}$。对于第一部分，注意对于 $\boldsymbol{x} \in \mathbf{F}_j$，$\partial f(\boldsymbol{x}) =$

$\hat{}N_{\mathrm{dom}(\mathbf{f})}(\boldsymbol{x}) + \mathrm{conv}\{\boldsymbol{a}_i \,|\, i \in \mathbf{I}\}$，其中 $\mathrm{conv}\{\mathbf{S}\}$ 表示$\mathbf{S}$的凸包。因此，$\partial f(\boldsymbol{x})$ 与 $\boldsymbol{x} \in \mathbf{F}j$无关。

$\hat{}$ 对于第二部分，注意对于任意 j，$k \in \mathbf{I}$ 和 $\boldsymbol{u} \in \mathbf{U}$，

$$\boldsymbol{a}_j^T(\boldsymbol{x}_j + F_j\boldsymbol{u}) + b_j = \boldsymbol{a}_k^T(\boldsymbol{x}_j + F_j\boldsymbol{u}) + b_k \quad \text{and}$$
$$\boldsymbol{a}_j^T\boldsymbol{x}_j + b_j = \boldsymbol{a}_k^T\boldsymbol{x}_j + b_k.$$

由这些方程可得

$$0 = (\boldsymbol{a}_j^T\boldsymbol{x}_j + b_j) - (\boldsymbol{a}_k^T\boldsymbol{x}_j + b_k) + (\boldsymbol{a}_j - \boldsymbol{a}_k)^T F_j\boldsymbol{u} = (\boldsymbol{a}_j - \boldsymbol{a}_k)^T F_j\boldsymbol{u}.$$

因此，对于任意 $\boldsymbol{v} \in \mathrm{conv}\{\boldsymbol{a}_j \,|\, j \in \hat{}\mathbf{I}\}$，都有 $(\boldsymbol{v} - \boldsymbol{a}_k)^T F_j\boldsymbol{u} = 0$。此外，对于任意 $\boldsymbol{n} \in N_{\mathrm{dom}(\mathbf{f})}(\boldsymbol{x}_j + F_j\boldsymbol{u})$，都有 $\boldsymbol{n}^T F_j\boldsymbol{u} = 0$。由此可得，对于任意 $\boldsymbol{v} \in \partial f(\boldsymbol{x}) = N_{\mathrm{dom}(f)}(\boldsymbol{x}) + \mathrm{conv}\{\boldsymbol{a}_i \,|\, i \in \hat{}\mathbf{I}\}$，都有 $(\boldsymbol{v} - \boldsymbol{a}_k)^T F_j\boldsymbol{u} = 0$。

定理 2.10。 在假设 *2.1* 和 *2.7*下。考虑定义 $\{\mathbf{F}_j\}_{j\in\mathbf{J}}$中的多面体划分，该划分见定义*2.8*。在 近端算子$\mathrm{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star)$ 以 $\mathbf{F}j$ 为条件时，其概率可由定义在 $\pi_{\mathbf{F}_j}(\boldsymbol{x})$ 上的密度 $\mathbf{F}j$ 表示，该密度与协方差为 Σ的高斯分布成正比。此外，我们有

$$\pi_{\mathbf{F}_i}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|_{\Sigma^{-1}}^2 - f(\boldsymbol{x})\right).$$

证明。 式（2.3）的最优性条件为 $\boldsymbol{x}^\star \in \boldsymbol{z}^\star + \Sigma\partial f(\boldsymbol{z}^\star)$。对于 $\boldsymbol{z}^\star \in \mathbf{F}j$，这可简化为 $\boldsymbol{x}^\star \in \boldsymbol{z}^\star + \Sigma\partial f_{\mathbf{F}j}$。令 $\mathbf{E} \subseteq \mathbf{F}j$ 为可测集，则可写为

$$\begin{aligned}\mathbb{P}\left(\mathrm{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) \in \mathbf{E}\right) &= \int_{\mathbf{E}+\Sigma\partial f_{\mathbf{F}_j}} \pi_{\boldsymbol{x}^\star}(\boldsymbol{w})\mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbf{E}}\int_{\Sigma\partial f_{\mathbf{F}_j}} \pi_{\boldsymbol{x}^\star}(\boldsymbol{z}+\boldsymbol{v})\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{v}\end{aligned}$$

其中 $\pi_{\boldsymbol{x}^\star}$是 $\boldsymbol{x}^\star \sim \mathcal{N}(\boldsymbol{\mu},\ \Sigma)$的密度。因此，我们可以用密度$\pi_{\mathbf{F}_j}$ 来表示概率，如下所示

$$\mathbb{P}\left(\mathrm{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star) \in \mathbf{E}\right) = \int_{\mathbf{E}} \pi_{\mathbf{F}_j}(\boldsymbol{z})\mathrm{d}\boldsymbol{z},$$

其中 $\pi_{\mathbf{F}_j}(\boldsymbol{z}) = \int_{\Sigma\partial f_{\mathbf{F}_j}} \pi_{\boldsymbol{x}^\star}(\boldsymbol{z}+\boldsymbol{w})\mathrm{d}\boldsymbol{w}$。

令 $\boldsymbol{z} = \boldsymbol{x}_j + F_j\boldsymbol{u}$，我们可以写为

$$\pi_{\mathbf{F}_j}(\boldsymbol{u}) \propto \int_{\partial f_{\mathbf{F}_j}} \pi_{\boldsymbol{x}^\star}(\boldsymbol{x}_j + F_j\boldsymbol{u} + \Sigma\boldsymbol{v})\mathrm{d}\boldsymbol{v} = \int_{\partial f_{\mathbf{F}_j} - \boldsymbol{d}} \pi_{\boldsymbol{x}^\star}(\boldsymbol{x}_j + F_j\boldsymbol{u} + \Sigma\boldsymbol{d} + \Sigma\boldsymbol{v})\mathrm{d}\boldsymbol{v}.$$

最优性o 正交性 $F_j\boldsymbol{u}$ 和$\partial f_{\mathbf{F}_j} - \boldsymbol{d}$ 由引理2.9 可知，这意味着对于任意 $\boldsymbol{v} \in \partial f_{\mathbf{F}_j} - \boldsymbol{d}$，

$$\begin{aligned}-2\log\left(\frac{\pi_{\boldsymbol{x}^\star}(\boldsymbol{x}_j + F_j\boldsymbol{u} + \Sigma\boldsymbol{d} + \Sigma\boldsymbol{v})}{\pi_{\boldsymbol{x}^\star}(\boldsymbol{x}_j + F_j\boldsymbol{u} + \Sigma\boldsymbol{d})}\right) &= 2(\boldsymbol{x}_j + F_j\boldsymbol{u} + \Sigma\boldsymbol{d})^T\boldsymbol{v} + \boldsymbol{v}^T\Sigma\boldsymbol{v} \\ &= 2(\boldsymbol{x}_j + \Sigma\boldsymbol{d})^T\boldsymbol{v} + \boldsymbol{v}^T\Sigma\boldsymbol{v},\end{aligned}$$

因此独立于 $\boldsymbol{u}$，由此我们可以得出结论：

$$\pi_{\mathbf{F}_j}(\boldsymbol{u}) \propto \pi_{\boldsymbol{x}^\star}(\boldsymbol{x}_j + F_j\boldsymbol{u} + \Sigma\boldsymbol{d}),$$

其中，$\pi_{\boldsymbol{x}^\star}(\cdot + \Sigma\boldsymbol{d})$ 是均值为 $\boldsymbol{\mu} - \Sigma\boldsymbol{d}$、协方差为Σ的正态分布的密度。

对于第二部分，请注意

$$\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu} + \Sigma\boldsymbol{d}\|^2_{\Sigma^{-1}} = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\Sigma^{-1}} + \boldsymbol{d}^T\boldsymbol{x} - \boldsymbol{d}^T\boldsymbol{\mu} + \frac{1}{2}\|\Sigma\boldsymbol{d}\|^2_{\Sigma^{-1}}.$$

只有前两项依赖于 $\boldsymbol{x}$。此外，根据引理2.9可得 $f|_{\mathbf{F}_i}(\boldsymbol{x}) = \boldsymbol{d}^T\boldsymbol{x} + b_j$，因此对于任意 $\boldsymbol{x} \in \mathbf{F}_i$，

$$\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\Sigma^{-1}} + \boldsymbol{d}^T\boldsymbol{x} = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\Sigma^{-1}} + f(\boldsymbol{x}).$$

因此，我们可以得出结论：

$$\pi_{\mathbf{F}_j}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu} + \Sigma\boldsymbol{d}\|^2_{\Sigma^{-1}}\right) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\Sigma^{-1}} - f(\boldsymbol{x})\right).$$

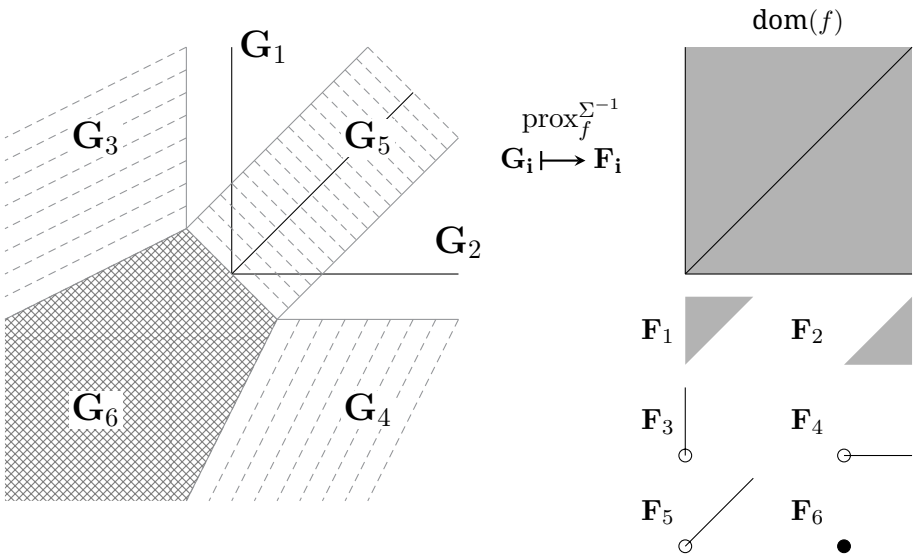

图 2.3：非负全变差正则化的可视化。斜向近端映射将集合 $\mathbf{G}_i$ 映射到相应的集合 $\mathbf{F}_i$。

图 2.3 展示了在二维中非负各向异性全变差对应的密度$\pi_{\mathbf{F}_j}$ 是如何得到的，即 $f(\boldsymbol{x}) = \gamma|x_1 - x_2| + \chi_{\mathbb{R}^2_+}$。包含在$k$维集合 $\mathbf{G}_i := \left[\text{近端算子}_f^{\Sigma^{-1}}\right]^{-1}(\mathbf{F}_i) = (\mathrm{Id} + \Sigma\partial f)(\mathbf{F}_i)$ 中的所有质量都会映射到 $(n-k)$维集合 $\mathbf{F}_i$。对于一维集合$\mathbf{F}_3, \mathbf{F}_4$ 和 $\mathbf{F}_5$，，集合 $\Sigma\partial f$ 以图示方式呈现

在相应集合中的平行虚线 $\mathbf{G}_3$, $\mathbf{G}_4$ 和 $\mathbf{G}_5$表示。对于零维原点 $\mathbf{F}_6$, , 次微分对应于交叉阴影区域 $\mathbf{G}_{6}$。

3. 在线性逆问题中的应用。 本节将讨论如何将正则化高斯分布的概念以及第 2 节中讨论的理论应用于线性逆问题。在第 2.1小节开头，我们将随机正则化线性最小二乘问题

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2+f(\boldsymbol{x})\right\},$$

其中 $A\in\mathbb{R}^{m\times n}$ 且 $\hat{}\,\boldsymbol{b}\sim\mathcal{N}(\boldsymbol{b},\ \lambda^{-1}I)$ 转化为定义 2.2中的近端框架。该转换要求假设 A 的秩为 n。然而，逆问题中的数据往往少于变量，因此无法应用这种转换。通过引入第二个随机线性最小二乘项，可以解决这个问题，从而得到如下形式的问题

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2+\frac{\delta}{2}\|L\boldsymbol{x}-\hat{\boldsymbol{c}}\|_2^2+f(\boldsymbol{x})\right\}, \tag{3.1}$$

其中 $\hat{}\,\boldsymbol{b}\sim\mathcal{N}(\boldsymbol{b},\ \lambda^{-1}I)$, $\hat{}\boldsymbol{c}\sim\mathcal{N}(\boldsymbol{c},\ \delta^{-1}I)$ 并假设 $\mathbf{Null}(A)\cap\mathbf{Null}(L)=\{\mathbf{0}\}$。由于 等价于 (3.1)

$$\operatorname{prox}_f^{\Sigma^{-1}}(\boldsymbol{x}^\star),\quad \text{with } \Sigma^{-1}=\delta L^TL+\lambda A^TA \text{ and } \boldsymbol{x}^\star\sim\mathcal{N}(\Sigma(\lambda A^T\boldsymbol{b}+\delta L^T\boldsymbol{c}),\Sigma), \tag{3.2}$$

该问题符合第 2 节所讨论的框架和理论。

在第 3.1 小节中，我们从贝叶斯视角讨论第 2 节框架适用的这一设定。在第 3.2 小节中，我们讨论随机线性最小二乘项欠定的设定。

3.1. 超定随机化。现在让我们回到从含噪测量 $\boldsymbol{x}\in\mathbb{R}^n$ 中重建信号 $\boldsymbol{b}=A\boldsymbol{x}+\boldsymbol{e}$，其中线性前向算子为 $A:\ \mathbb{R}^n\to\mathbb{R}^m$，并且噪声为独立高斯噪声 $\boldsymbol{e}\sim\mathcal{N}(0,\ \lambda^{-1}I)$。如果我们先验地假设

$$\pi(\boldsymbol{x})\propto\exp\left(-\frac{\delta}{2}\|L\boldsymbol{x}-\boldsymbol{c}\|_2^2\right), \tag{3.3}$$

那么我们就可以计算出后验分布的结果（差一个缩放因子），即:

$$\pi(\boldsymbol{x}\,|\,\boldsymbol{b})\propto\pi(\boldsymbol{b}\,|\,\boldsymbol{x})\pi(\boldsymbol{x})\propto\exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2-\frac{\delta}{2}\|L\boldsymbol{x}-\boldsymbol{c}\|_2^2\right).$$

现在，我们可以利用正则化随机线性最小二乘问题定义正则化后验 $\boldsymbol{x}^\star$

$$\boldsymbol{x}^\star=\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2+\frac{\delta}{2}\|L\boldsymbol{x}-\hat{\boldsymbol{c}}\|_2^2+f(\boldsymbol{x})\right\}.$$

在假设 2.7下，根据定理2.10，我们可以得出，该正则化后验在多面体划分（定义 $\mathbf{F}_i$ 2.8）中的集合 上的条件分布满足

$$\pi(\boldsymbol{x} \,|\, \mathbf{F}_i, \boldsymbol{b}) \propto \exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2 - \frac{\delta}{2}\|L\boldsymbol{x}-\boldsymbol{c}\|_2^2 - f(\boldsymbol{x})\right) \propto \pi(\boldsymbol{b} \,|\, \boldsymbol{x})\pi(\boldsymbol{x})\exp(-f(\boldsymbol{x})). \tag{3.4}$$

注意，对于 $\pi(\boldsymbol{b} \,|\, \boldsymbol{x}) = \pi(\boldsymbol{b} \,|\, \boldsymbol{x},\ \mathbf{F}_i)$，若 $\boldsymbol{x} \in \mathbf{F}_i$，则可以计算与正则化后验相对应的条件先验，具体而言，

$$\pi(\boldsymbol{x} \,|\, \mathbf{F}_i) \propto \frac{\pi(\boldsymbol{x} \,|\, \mathbf{F}_i, \boldsymbol{b})}{\pi(\boldsymbol{b} \,|\, \boldsymbol{x}, \mathbf{F}_i)} \propto \pi(\boldsymbol{x})\exp(-f(\boldsymbol{x})) \propto \exp\left(-\frac{\delta}{2}\|L\boldsymbol{x}-\boldsymbol{c}\|_2^2 - f(\boldsymbol{x})\right). \tag{3.5}$$

因此，在给定集合 $\mathbf{F}_i$ 的条件下，先验与我们在整个 $\mathbb{R}^n$ 上选取先验 $\exp\left(-\frac{\delta}{2}\|L\boldsymbol{x}-\boldsymbol{c}\|_2^2 - f(\boldsymbol{x})\right)$ 时相同。然而，正则化后验会为稀疏集合赋予正概率。此外，这表明，正则化后验相当于将变维模型作为先验施加于 $\boldsymbol{x}$，其中模型与集合 $\mathbf{F}_i$ 相关联。因此，我们的方法隐式地定义了模型先验 $\pi(\mathbf{F}_i)$，它通常取决于前向算子和数据。

3.2. 欠定随机化。在本小节中，我们考虑线性逆问题欠定且不添加显式先验的情形，即考虑如下形式的优化问题

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{\frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2 + f(\boldsymbol{x})\right\}, \tag{3.6}$$

其中 $A \in \mathbb{R}^{m\times n}$ 的秩小于 n。

如果 $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{\delta}{2}\|L\boldsymbol{x}-\boldsymbol{c}\|$ 对于某个 L 满足 $\mathbf{Null}(A) \cap \mathrm{Null}(L) = \{\mathbf{0}\}$，仍然是正常下半连续凸函数，那么优化问题（3.6）等价于

$$\boldsymbol{z}^\star = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{\frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2 + \frac{\delta}{2}\|L\boldsymbol{x}-\boldsymbol{c}\|_2^2 + g(\boldsymbol{x})\right\}, \tag{3.7}$$

也就是说，这与添加显式先验时的优化问题相同，但不对先验最小二乘项进行随机化。如果 f 是强凸函数，则可能会出现这种情况。

因此，$\boldsymbol{z}^\star$ 可以写成

$$\boldsymbol{z}^\star = \mathrm{prox}_g^{\lambda A^TA+\delta L^TL}\left(\boldsymbol{x}^\star_{\mathrm{deg}}\right),$$

其中 $\boldsymbol{x}^\star_{\mathrm{deg}} = (\lambda A^TA + \delta L^TL)^{-1}(A^T\hat{\boldsymbol{b}} + L^T\boldsymbol{c})$ 是退化高斯分布 .

由于近端算子 $f^{\lambda A^TA+\delta L^TL}$ 仍是连续函数，根据引理 B.1 可得，$\boldsymbol{z}^\star$ 不一定支撑在 dom 的整个范围上(f)，因此，$\boldsymbol{z}^\star$ 的支撑可能不包含真实信号 $\boldsymbol{x}_0 \in \mathrm{dom}(f)$。以下引理刻画了哪些信号可以被重构。

引理 3.1. 假设 f 满足条件，使得优化问题 (3.6) 对于任意扰动$\hat{}$ $\boldsymbol{b}$都是适定的，即存在唯一解，且该解关于扰动连续依赖$\hat{}\boldsymbol{b}$。那么，对于任意 $\boldsymbol{x}_0 \in \mathbb{R}^n$，存在 $\boldsymbol{q} \in \mathbb{R}^m$ 使得

$$\boldsymbol{x}_0 = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{q}\|_2^2 + f(\boldsymbol{x}) \right\}, \tag{3.8}$$

当且仅当 $\partial f(\boldsymbol{x}_0) \cap$ 值域$(A^T) \neq \varnothing$。

此外，若 $\hat{}\boldsymbol{q} \in \mathbb{R}^m$ 是一个随机变量，且满足 支撑集$(\hat{}\boldsymbol{q}) = \mathbb{R}^m$，则随机变量

$$\boldsymbol{x}^\star = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|A\boldsymbol{x} - \hat{\boldsymbol{q}}\|_2^2 + f(\boldsymbol{x}) \right\}, \tag{3.9}$$

满足

$$\operatorname{supp}(\boldsymbol{x}^\star) = \operatorname{cl}\left(\left\{\boldsymbol{x} \in \operatorname{dom}(f) \,|\, \partial f(\boldsymbol{x}) \cap \operatorname{range}(A^T) \neq \varnothing\right\}\right).$$

证明。 第一部分可由以下观察得出：式（3.8）的最优性条件可以写成

$$A^T(\boldsymbol{q} - A\boldsymbol{x}_0) \in \partial f(\boldsymbol{x}_0).$$

对于第二部分，根据假设，解映射 $\boldsymbol{q} \mapsto$ 极小值点$_{\boldsymbol{x}} \left\{\frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{q}\|_2^2 + f(\boldsymbol{x})\right\}$是连续的，因此结合引理B.1即可由第一部分得出结论。 ■

虽然对于一般的正则化函数 f，集合 $\operatorname{supp}(\boldsymbol{x}^\star)$ 可能非常复杂，但以下推论描述了几种 $\operatorname{supp}(\boldsymbol{x}^\star)$ 表现良好的情形。

推论 3.2. 在假设 2.7下，考虑定义 $\{\mathbf{F}_i\}_{i\in\mathbf{J}}$ 中的多面体划分2.8。假设 (3.8) 是适定的，那么如果 $\partial f(\boldsymbol{x}) \cap \operatorname{range}(A^T) \neq \varnothing$ 对任意 $\boldsymbol{x} \in \mathbf{F}_i$成立，则它$\tilde{}$对所有 $\boldsymbol{x} \in \mathbf{F}_i$都成立。因此，存在 $\mathbf{J} \subset \mathbf{J}$ 使得

$$\operatorname{supp}(\boldsymbol{x}^\star) = \bigcup_{j\in\tilde{\mathbf{J}}} \operatorname{cl}\left(\mathbf{F}_j\right).$$

此外，若 f 对称且正齐次，则支撑$(\boldsymbol{x}^\star)$ 是一个线性子空间；若 $f = \chi_{\mathbf{C}} + g$，其中$\mathbf{C}$ 是一个闭凸集，且 g 对称且正齐次，则支撑$(\boldsymbol{x}^\star)$ 是一个凸集。

例如，根据推论3.2的第二部分，支撑 $\operatorname{supp}(\boldsymbol{x}^\star)$ 是一个线性子空间。$\operatorname{supp}(\boldsymbol{x}^\star)$ 的凸性保证某些点估计（如经验均值）仍位于支撑内。不过请注意，支撑并不依赖于用于重建的数据，因为受扰动的数据具有全支撑。

尽管采用低维随机化的这些方法仍能取得良好结果，但分布的低维支撑无法保证真值附近具有正概率质量；如第 3.1 小节开头所述，加入随机化线性最小二乘正则化即可保证这一点。

4. 贝叶斯层次模型。考虑第 3.1 小节中 $\boldsymbol{c} = \mathbf{0}$的设定。为使模型更加稳健，我们将在超参数 λ 和δ上加入先验分布。更具体地，令条件分布 $\pi(\boldsymbol{x} \,|\, \lambda, \delta, \boldsymbol{b})$ 定义为

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x} - \hat{\boldsymbol{b}}\|_2^2 + \frac{\delta}{2}\|L\boldsymbol{x} - \hat{\boldsymbol{c}}\|_2^2 + \sqrt{\delta} f(\boldsymbol{x}) \right\}, \tag{4.1}$$

其中 $\hat{}\, \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{b},\ \lambda^{-1} I)$ 以及 $\hat{}\boldsymbol{c} \sim \mathcal{N}(\mathbf{0},\ \delta^{-1} I)$。此外，令 $\alpha_\lambda,\ \alpha_\delta,\ \beta_\lambda,\ \beta_\delta > 0$，并定义超先验

$$\lambda \sim \Gamma(\alpha_\lambda, \beta_\lambda) \quad \text{and} \quad \delta \sim \Gamma(\alpha_\delta, \beta_\delta),$$

或者等价地

$$\pi(\lambda) \propto \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda) \quad \text{and} \quad \pi(\delta) \propto \delta^{\alpha_\delta - 1} \exp(-\beta_\delta \delta). \tag{4.2}$$

为了计算 δ 在给定 $\boldsymbol{x}$时的条件分布，我们需要知道条件先验（3.5）如何依赖于 δ，而这种依赖关系隐藏在比例关系中。在对正则化函数 f施加某些限制后，可以计算出这个比例关系，如下引理所述。

引理 4.1。 假设 f 的上图除相加常数外是一个多面锥。等价地， f满足假设2.7且为正齐次，即 $f(\gamma\boldsymbol{x}) = \gamma f(\boldsymbol{x})$对于$\gamma \geqslant 0$。那么条件先验 $\pi(\boldsymbol{x} \,|\, \delta,\ \mathbf{F}_j)$ 满足

$$\pi(\boldsymbol{x} \,|\, \delta, \mathbf{F}_j) \propto \delta^{-\dim(\mathbf{F}_j)/2} \exp\left(-\frac{\delta}{2}\|L\boldsymbol{x}\|_2^2 - \sqrt{\delta} f(\boldsymbol{x}) \right). \tag{4.3}$$

证明。 通过计算归一化常数 δ ，我们可以得到对 K的依赖关系。利用任意集合 $\mathbf{F}j$ 在乘以正数后保持不变这一性质，

$$\begin{aligned} 1 = \int_{\mathbf{F}_j} \pi(\boldsymbol{x} \,|\, \delta, \mathbf{F}_i) \mathrm{d}\boldsymbol{x} &= K \int_{\mathbf{F}_j} \exp\left(-\frac{\delta}{2}\|L\boldsymbol{x}\|_2^2 - \sqrt{\delta} f(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \\ &= K \int_{\mathbf{F}_j} \exp\left(-\frac{1}{2}\|L\sqrt{\delta}\boldsymbol{x}\|_2^2 - f(\sqrt{\delta}\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \\ &= K\delta^{-\dim(\mathbf{F}_j)/2} \int_{\mathbf{F}_j} \exp\left(-\frac{1}{2}\|L\boldsymbol{x}\|_2^2 - f(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x}. \end{aligned}$$

因此我们得到

$$\pi(\boldsymbol{x} \,|\, \delta, \mathbf{F}_j) \propto \delta^{\dim(\mathbf{F}_j)/2} \exp\left(-\frac{\delta}{2}\|L\boldsymbol{x}\|_2^2 - \sqrt{\delta} f(\boldsymbol{x}) \right).$$

■

请注意，在优化问题（4.1）和条件先验（4.3）中，正则化函数均由超参数 δ 缩放。尽管对 f 的上图为多面锥这一附加条件可能相当严格，但许多重要的正则化函数都满足这一条件，例如非负约束以及 l_1基于范数的正则化，如 $\|Dx\|_1$ ，其中任意矩阵 D。

U利用引理 4.1，我们可以得到给定 $\mathbf{F}^j$ 时（超）参数的分布，

$$\begin{aligned}\pi(\boldsymbol{x},\lambda,\delta\,|\,\boldsymbol{b},\mathbf{F}_j) &\propto \mathbf{1}\{\boldsymbol{x}\in\mathbf{F}_j\}\,\pi(\boldsymbol{b}\,|\,\boldsymbol{x},\lambda,\delta)\pi(\boldsymbol{x}\,|\,\delta,\mathbf{F}_j)\pi(\lambda)\pi(\delta)\\ &\propto \mathbf{1}\{\boldsymbol{x}\in\mathbf{F}_j\}\,\lambda^{m/2+\alpha_\lambda-1}\delta^{\dim(\mathbf{F}_j)/2+\alpha_\delta-1}\\ &\quad\times\exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2-\frac{\delta}{2}\|L\boldsymbol{x}\|_2^2-\sqrt{\delta}f(\boldsymbol{x})-\beta_\lambda\lambda-\beta_\delta\delta\right),\end{aligned}\tag{4.4}$$

其中，$\mathbf{1}\{\boldsymbol{x}\in\mathbf{F}_j\}$ 在 $\boldsymbol{x}\in\mathbf{F}^j$ 时为 $\mathbf{1}$，否则为 $\mathbf{0}$。

给定 $\boldsymbol{x}$，将包含 $\boldsymbol{x}$ 的集合记为 $\mathbf{F}$ $\mathbf{F}(\boldsymbol{x})$，则可得到超参数的以下条件分布，

$$\pi(\lambda\,|\,\boldsymbol{x},\boldsymbol{b})\propto\lambda^{m/2+\alpha_\lambda-1}\exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2-\beta_\lambda\lambda\right),\ \text{for }\lambda>0\quad\text{and}\tag{4.5}$$

$$\pi(\delta\,|\,\boldsymbol{x},\boldsymbol{b})\propto\delta^{\dim(\mathbf{F}(\boldsymbol{x}))/2+\alpha_\delta-1}\exp\left(-\frac{\delta}{2}\|L\boldsymbol{x}\|_2^2-\sqrt{\delta}f(\boldsymbol{x})-\beta_\delta\delta\right),\ \text{for }\delta>0,\tag{4.6}$$

其中，第一个条件分布满足 $\lambda\,|\,\boldsymbol{x},\boldsymbol{b}\sim\Gamma(m/2+\alpha_\lambda,\frac{1}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2+\beta_\lambda)$。

这些条件分布可用于推导算法 4.1中所述的多面锥上图层次吉布斯采样器。

Input: $\boldsymbol{x}^0,\alpha_\lambda,\beta_\lambda,\alpha_\delta,\beta_\delta,k_{\max}$
for $k=1$ **to** $k_{\max}$ **do**
  Compute $(\lambda^k,\delta^k)\sim\pi(\lambda,\delta\,|\,\boldsymbol{x},\boldsymbol{b})$ as follows:
    $\lambda^k\sim\Gamma\left(m/2+\alpha_\lambda,\frac{1}{2}\|A\boldsymbol{x}^{k-1}-\boldsymbol{b}\|_2^2+\beta_\lambda\right)$,
    $\delta^k\sim\pi(\delta\,|\,\boldsymbol{x}^{k-1},\boldsymbol{b})$ defined in (4.6).
  Compute $\boldsymbol{x}^k\sim\pi_{\boldsymbol{x}|\boldsymbol{b},\lambda^k,\delta^k}$ using (4.1).
end for
return $\{(\boldsymbol{x}^k,\lambda^k,\delta^k)\}_{k=1,\dots,k_{\max}}$

算法 **4.1**：用于 $(\boldsymbol{x},\lambda,\delta$ 的多面锥上图层次吉布斯采样器).

与从 $\lambda\,|\,\boldsymbol{x},\boldsymbol{b}$ 中采样相比，从 $\delta\,|\,\boldsymbol{x},\boldsymbol{b}$中采样相对容易，因为前一种分布能够利用伽马分布和高斯分布之间的共轭性。有一种情况下，条件分布 $\delta\,|\,\boldsymbol{x},\boldsymbol{b}$ 可以简化为伽马分布。如果正则化函数是多面锥的示性函数，那么

$$\delta\,|\,\boldsymbol{x},\boldsymbol{b}\sim\Gamma(\dim(\mathbf{F}(\boldsymbol{x}))/2+\alpha_\delta,\frac{1}{2}\|L\boldsymbol{x}\|_2^2+\beta_\delta).$$

对于这种受约束的情形，算法 4.1 可简化为 [5,Sampler3.1]。

另一种条件分布更简单的模型如下。若 $\mathbf{Null}(A)=\{\mathbf{0}\}$，则优化问题（$\frac{\delta}{2}\|L\boldsymbol{x}-\hat{}\boldsymbol{c}\|_2^2$**4.1**）中无需使用随机最小二乘项 来采用近端框架。因此，考虑以下替代优化

问题

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2 + \gamma f(\boldsymbol{x}) \right\}, \tag{4.7}$$

其中 $\hat{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{b},\ \lambda^{-1}I)$。然后，我们可以对噪声水平 λ和正则化强度 γ都设置超先验。更确切地说，令

$$\lambda \sim \Gamma(\alpha_\lambda, \beta_\lambda) \quad \text{and} \quad \gamma \sim \Gamma(\alpha_\gamma, \beta_\gamma), \tag{4.8}$$

其中 $\alpha_\lambda,\ \alpha\gamma,\ \beta_\lambda,\ \beta_\gamma > 0$。

运用与引理 4.1中类似的论证，可得条件分布

$$\lambda \,|\, \boldsymbol{x}, \boldsymbol{b} \sim \Gamma(m/2+\alpha_\lambda, \frac{1}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \beta_\lambda) \tag{4.9}$$

$$\gamma \,|\, \boldsymbol{x}, \boldsymbol{b} \sim \Gamma(\dim(\mathbf{F}(\boldsymbol{x})) + \alpha_\gamma, f(\boldsymbol{x}) + \beta_\gamma). \tag{4.10}$$

注意，当 $L = 0$ 时，无法直接从式（4.4）推导出这些分布。通过将算法第 5 行中 δ 对应的条件分布替换为条件分布（4.10），即可得到一个与算法 4.1 类似的吉布斯采样器。

算法 4.1 要求对任意 $\boldsymbol{x} \in \mathrm{dom}(f)$ 计算 $\dim(\mathrm{F}(\boldsymbol{x}))$ 。对于许多正则化函数，这个量都可以高效计算。每个面 F 都包含于描述稀疏子集的线性方程组 $S\boldsymbol{x} = \boldsymbol{s}$ 的解空间中；此外，$\dim(\mathrm{F}) = \dim(\mathrm{Null}(S)) = n - \mathrm{rank}(S)$。对于 $f(\boldsymbol{x}) = \chi_{\mathbb{R}^n_+}(\boldsymbol{x})$或 $f(\boldsymbol{x}) = \gamma\|\boldsymbol{x}\|_1$，以及 $\boldsymbol{x} \in \mathbb{R}^n_+$，与 $\mathrm{F}(\boldsymbol{x})$ 对应的矩阵 S 由单位矩阵中与 $\boldsymbol{x}$中的零元素相对应的行组成。类似地，对于 $f(\boldsymbol{x}) = \gamma\|D\boldsymbol{x}\|_1$，与 $\mathrm{F}(\boldsymbol{x})$对应的矩阵 S 由 D 中满足 $\boldsymbol{d}^T\boldsymbol{x} = 0$的行 $\boldsymbol{d}$ 组成。虽然这种方法适用于任何稀疏模式，但下面的命题针对前述的一些例子，给出了 $\dim(\mathrm{F}(\boldsymbol{x}))$ 的若干简洁表达式。

命题 4.2。 设 $\gamma > 0$，则各个面的维数可表示为：

$$\dim(\mathbf{F}(\boldsymbol{x})) = \begin{cases} \|\boldsymbol{x}\|_0, & \text{if } f(\boldsymbol{x}) = \chi_{\mathbb{R}^n_+}(\boldsymbol{x}) \text{ or } f(\boldsymbol{x}) = \gamma\|\boldsymbol{x}\|_1, \\ \|D\boldsymbol{x}\|_0 + n - \mathrm{rank}(D), & \text{if } f(\boldsymbol{x}) = \gamma\|D\boldsymbol{x}\|_1, \\ N_{flat}(\boldsymbol{x}), & \text{if } f(\boldsymbol{x}) = \gamma\|L\boldsymbol{x}\|_1, \\ N_{nz\text{-}flat}(\boldsymbol{x}), & \text{if } f(\boldsymbol{x}) = \gamma\|L\boldsymbol{x}\|_1 + \chi_{\mathbb{R}^n_+}(\boldsymbol{x}), \end{cases}$$

其中，$D \in \mathbb{R}^{k\times n}$ 满足 $k \leqslant n$ 且为满秩，L 是任意有限维空间中各向异性全变分的有限差分矩阵，$N_{flat}(\boldsymbol{x})$ 是 $\boldsymbol{x}$中连通平坦区域的数量，$N_{nz\text{-}flat}(\boldsymbol{x})$ 是 $\boldsymbol{x}$中非零连通平坦区域的数量。

证明。 对于 $f(\boldsymbol{x}) = \chi_{\mathbb{R}^n_+}(x)$或 $f(\boldsymbol{x}) = \gamma\|x\|_0$，与 S对应的矩阵 $\mathbf{F}(\boldsymbol{x})$具有如下各行$\boldsymbol{e}_i$ 对于 i ，使得 $\boldsymbol{x}_i = 0$。因此，S 的秩等于 $\boldsymbol{x}$中零元素的数量，即秩$(S) = n - \|\boldsymbol{x}\|_0$，因此，维数$(\mathbf{F}(\boldsymbol{x})) = \|\boldsymbol{x}\|_0$。

同理，对于 $f(\boldsymbol{x}) = \gamma\|D\boldsymbol{x}\|_1$，$S$ 中对应于$\mathbf{F}(\boldsymbol{x})$ 的行，是指满足 $\boldsymbol{d}^T\boldsymbol{x} = 0$ 的那些 D中的行 $\boldsymbol{d}$ 。由于 D 是满秩的，因此 S 也是满秩的，其秩为$(S) = \mathrm{rank}(D) - \|D\boldsymbol{x}\|_0$。

对于 $f(\boldsymbol{x}) = \gamma\|L\boldsymbol{x}\|_1$，令 $\boldsymbol{x} \in \mathbb{R}^n$和 P为一个置换矩阵，使得 $P\boldsymbol{x}$可分为 $N_{\text{flat}}(\boldsymbol{x})$ 个连续分量，每个分量对应于 $\boldsymbol{x}$中的一个连通平坦区域。对应于$\mathbf{F}(\boldsymbol{x})$ 的矩阵 S，对于 $\boldsymbol{x}$ 中每对相邻且值相同的数值都有一行，因此 PSP^T是一个分块矩阵，其对角线上有 $N_{\text{flat}}(\boldsymbol{x})$个矩形块，其余位置均为零。若 l_j 是分块 B_j中的列数，也就是说，l_j 是 B_j 所表示的平坦分量中的数值个数，则 $\mathsf{rank}(B_j) = l_j - 1$。因此，

$$\operatorname{rank}(S) = \operatorname{rank}(PSP^T) = \sum_{j=1}^{N_{\text{flat}}(\boldsymbol{x})} \operatorname{rank}(B_j) = \sum_{j=1}^{N_{\text{flat}}(\boldsymbol{x})} (l_j - 1) = n - N_{\text{flat}}(\boldsymbol{x}),$$

因此，我们可以得出 $\mathsf{dim}(\mathbf{F}(\boldsymbol{x})) = N_{\text{flat}}(\boldsymbol{x})$。

关于 $f(\boldsymbol{x}) = \gamma\|D\boldsymbol{x}\|_1 + \chi_{\mathbb{R}^n_+}(\boldsymbol{x})$ 的论证几乎完全相同，唯一的区别是，对于任意一个由零分量构成的块 B_j，其秩满足 $\mathsf{rank}(B_j) = l_j$，因此 $\mathsf{rank}(S) = n - N_{\text{nz-flat}}(\boldsymbol{x})$。

尽管命题 4.2 指出，$\mathsf{dim}(\mathbf{F}(\boldsymbol{x}))$ 可以高效计算，但在给定超参数的情况下，如何从（4.1）中计算样本是第 5节讨论的问题。

5. 数值示例。现在，我们将第 3 节和4 节的方法应用于去模糊和计算机断层扫描示例。第5.1 小节重点探讨正则化的影响，第 5.2 小节重点介绍第 4节中的吉布斯采样器。最后，在第5.3小节中，我们运用第 3.2 小节的思路，说明在欠定情形下也能获得有效结果。在所有实验中，我们都使用交替方向乘子法（ADMM），通过固定的迭代次数近似求解（4.1）。该算法的简要说明见附录 A。

5.1. 正则化的影响。在本小节和下一小节中，我们考虑由下式定义的一维高斯去模糊问题：

$$\boldsymbol{b} = A\boldsymbol{x} + \boldsymbol{e},$$

对于真实信号 $\boldsymbol{x} \in \mathbb{R}^n$、噪声 $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0},\ \lambda^{-1}I)$ 超参数为 $\lambda = 1000$，前向算子 A 由以下托普利兹矩阵定义：

$$A_{ij} = \frac{h}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{h(i-j)}{\sigma}\right)^2\right), \quad \text{for } i,j = 1,\dots,n,$$

其中 $m = n = 128$，$h = 1/n$ 以及 $\sigma = 0.02$。真实信号 $\boldsymbol{x}$和含噪数据 $\boldsymbol{b}$ 如图 5.1所示。

前向算子 A 是一个采用零边界条件的高斯模糊算子，且 A为满秩。因此，无需额外添加随机最小二乘项即可应用第 3.1 小节和第 4节的理论，并将考虑以下随机优化问题

$$\underset{\boldsymbol{x}\in\mathbb{R}^{128}}{\operatorname{argmin}} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x} - \hat{\boldsymbol{b}}\|_2^2 + \gamma\|L\boldsymbol{x}\|_1 \right\} \tag{5.1}$$

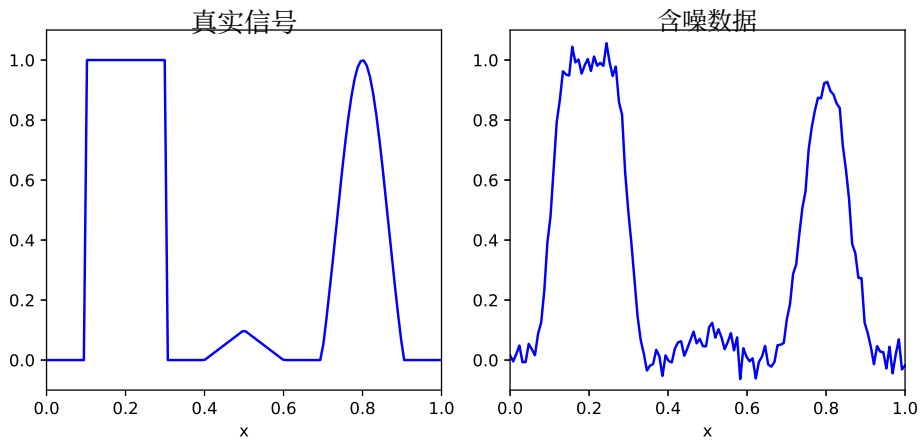

图5.1：真实信号（左）和含噪数据（右），其中含有 $\lambda = 1000$ 。

其中 $\hat{} \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{b},\ \lambda^{-1}I)$ 和 $\gamma > 0$，其中 $L \in \mathbb{R}^{(n-1)\times n}$ 是一个有限差分矩阵。对于 $\hat{}$ 未扰动问题，即将 $\boldsymbol{b}$ 替换为 $\boldsymbol{b}$，确定性解对应于采用拉普拉斯差分先验所得到的后验分布的最大后验估计。该后验分布的形式为

$$\pi(\boldsymbol{x} \mid \boldsymbol{b}) \propto \exp\left(-\frac{\lambda}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 - \gamma\|L\boldsymbol{x}\|_1\right). \tag{5.2}$$

从小节 3.1中，我们知道，从隐式定义的概率分布（5.1）中抽取的样本以正概率呈现分段常数行为。与显式定义的后验分布（5.2）中的样本不同。图 5.2 展示了隐式分布（5.1）和显式分布（5.2）在不同 γ取值下的若干（近似）独立样本，即正则化强度和先验强度。显式分布（5.2）的样本通过随机游走MCMC算法计算，而隐式分布（5.1）的样本则通过交替方向乘子法计算。图5.2 突显了连续分布的样本与通过变维模型得到的样本在正则性方面的差异。

尽管（5.2）的最大后验估计对应于（5.1）的未扰动解，但其他点估计表现出不同的行为。图 5.3展示了不同 γ取值下两个模型的逐分量中位数。显式模型（5.2）比隐式模型（5.1）表现出更高的光滑性。

图 5.4 展示了两个模型的95%可信区间宽度。对于相同的参数 γ，正则化分布具有较低的不确定性，但两个模型具有相似的特征。两个模型对信号左侧的不连续跳跃给出了相似的结果，但在信号中心的小峰值周围，隐式模型显示出更明显的峰值。尽管两个模型都对信号右侧抛物线两侧赋予了较高的不确定性，但隐式模型的不确定性更加均匀，这可能与跳跃中的不确定性有关，如图5.2所示。

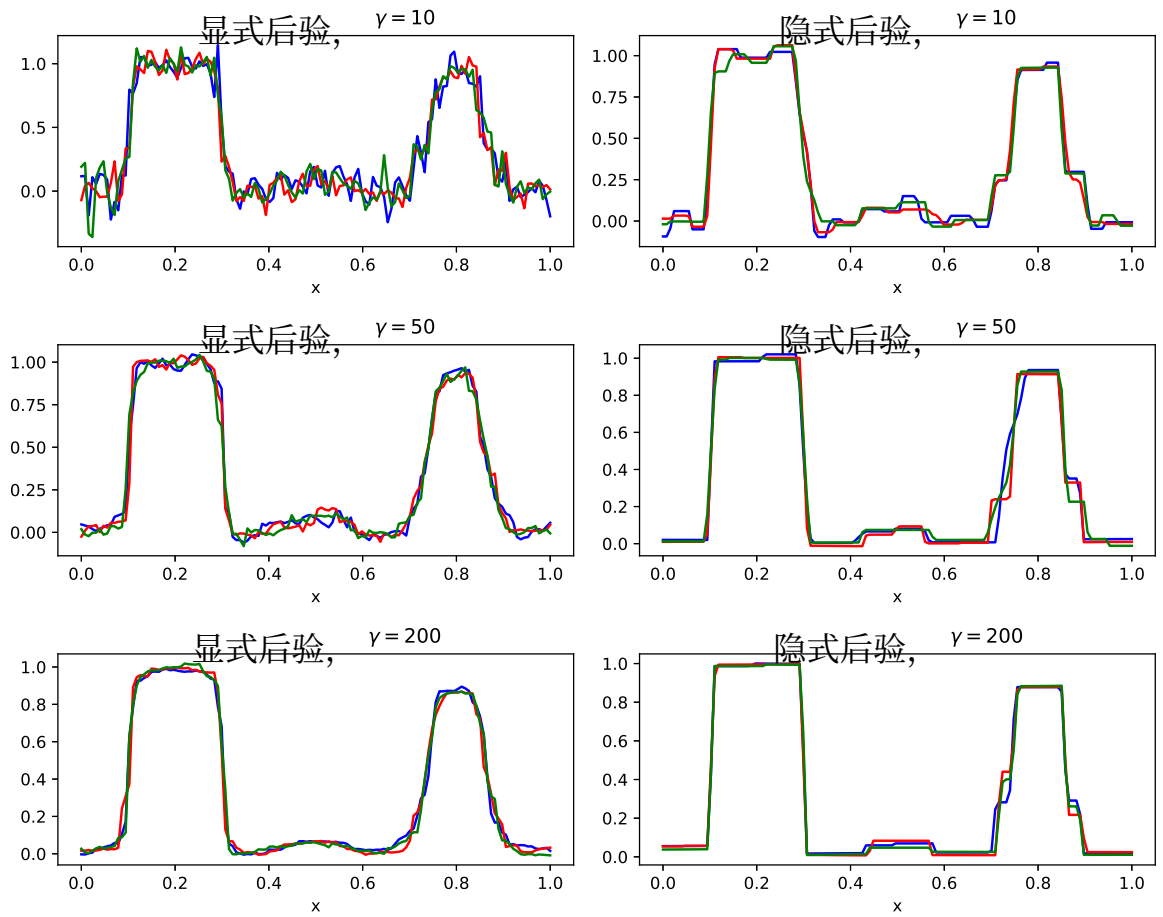

图 5.2：来自具有拉普拉斯差分先验的后验分布的样本（左），以及来自全变分正则化随机线性最小二乘问题的样本（右）。

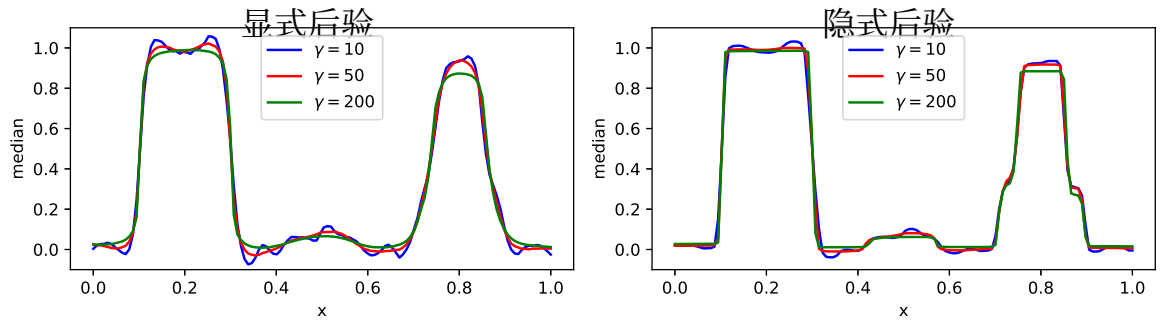

图 5.3：具有拉普拉斯差分先验的后验分布的中位数（左），以及全变分正则化随机线性最小二乘问题的中位数（右）。

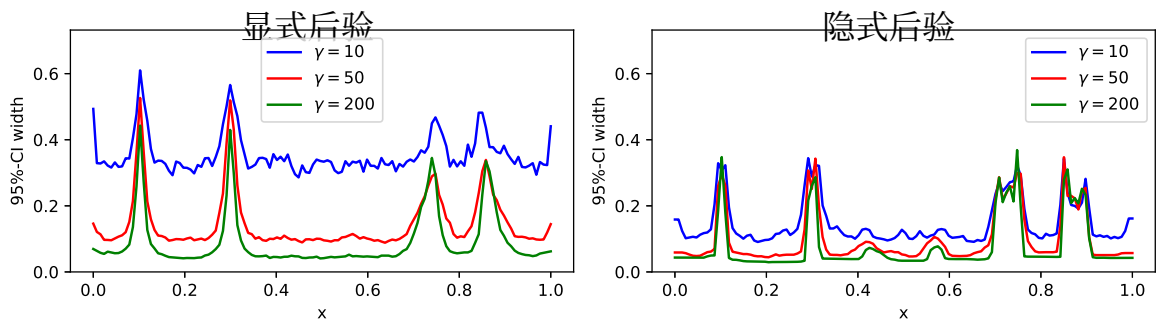

图5.4：采用拉普拉斯差分先验的后验分布（左）以及全变分正则化随机线性最小二乘问题（右）的95%可信区间宽度。

5.2. 吉布斯采样器。 同样的去模糊问题也可以使用第 4 节提出的吉布斯采样器求解。具体而言，我们考虑以下带非负约束和全变差正则化的问题

$$\underset{\boldsymbol{x}\in\mathbb{R}_{+}^{128}}{\operatorname{argmin}} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2 + \gamma\|L\boldsymbol{x}\|_1 \right\}, \tag{5.3}$$

其中 $\boldsymbol{e} \sim \mathcal{N}(0,\lambda^{-1}I)$ 以及超先验 $\lambda \sim \Gamma(1,10^{-4})$ 和 $\gamma \sim \Gamma(1,10^{-4})$。使用算法 4.1 的一个变体可获得该层次贝叶斯模型的样本。由于计算一个样本的代价较高，我们研究对（5.3）进行近似求解。其缺点是样本描述的是略有不同的分布，而且不准确的解也可能对条件分布（4.9）和（4.10）产生较大影响。图5.5展示了超参数 λ 和 γ的分布、正则化强度γ/λ, 的分布，以及以 $\dim(\mathrm{F}(\mathrm{x}))$ 表示的样本稀疏性分布。注意， λ 的分布近似独立于迭代次数。这可以解释为，交替方向乘子法能够较快地将目标函数最小化至较低精度，因此可在数量级上得到 $\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2$。其他链的衰减速度相当快，经过相对较少的迭代便已给出一致的结果。主要问题在于，交替方向乘子法需要更多迭代才能精确求解该优化问题，而这对于准确计算 $\dim(\mathrm{F}(\mathrm{x}))$ 是必需的。不过，由于问题规模较小，它很快就能收敛到正确的数量级，足以获得稳定的正则化效应。

图5.6 显示，交替方向乘子法能够迅速达到较低精度，但要达到较高精度则需要更多的迭代次数，这体现了分量级 95%可信区间的宽度。尽管 50 次迭代似乎足以确定不确定性相对较高的区域，但对应于大幅跳变的峰值（如图 5.2所示）仍在显著增长。不过，在实际应用中，这样的精度并非总是必要的。

.....的值 γ/λ 会极大地影响样本的规则性，进而影响测得的不确定性。图5.7 显示了平均正则化强度 γ/λ与施加于 γ的伽马先验的期望和方差之间的函数关系 γ。图中显示，与.....类似 λ,

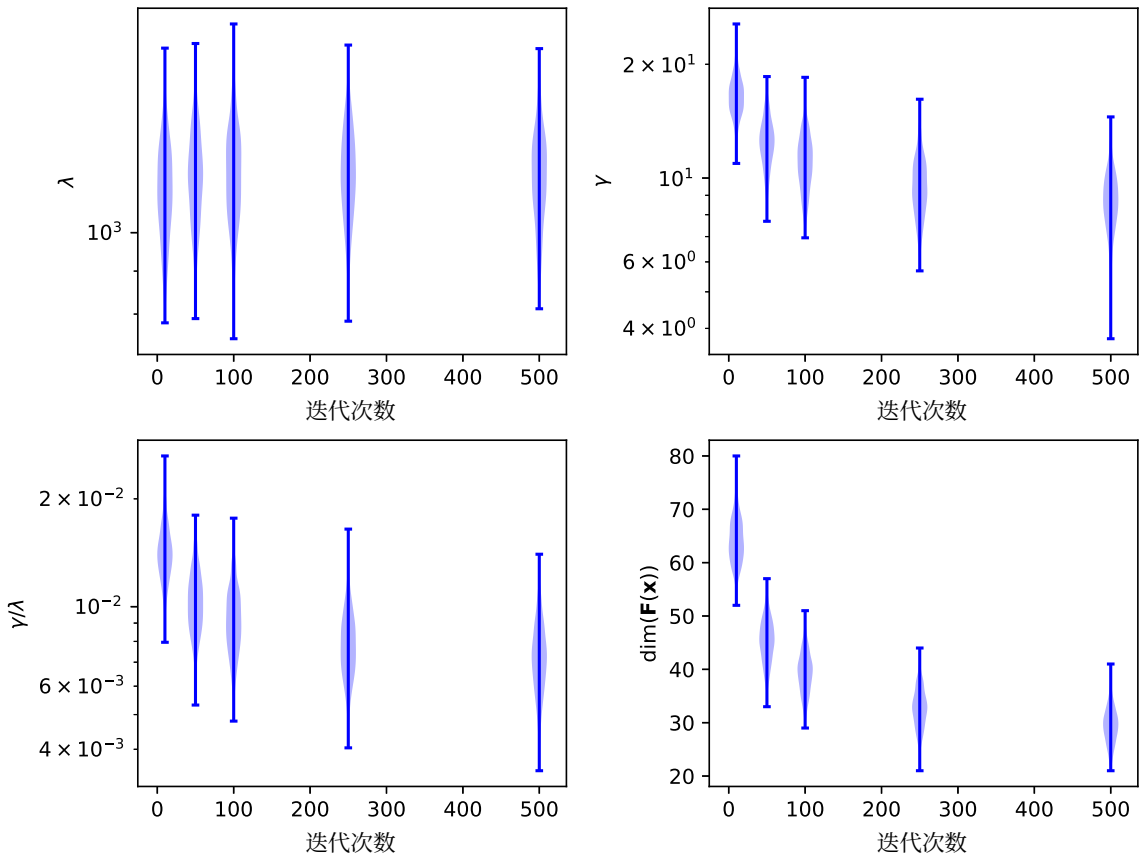

图 5.5：随着精度/迭代次数增加，超参数 λ 和 γ、正则化强度 γ/λ 以及稀疏性 $\dim(\mathbf{F}(\mathbf{x}))$ 的分布。

只要方差足够大， γ/λ 链就是稳定的。方差过小时，期望会主导 γ，从而导致超先验不稳定。

5.3. 欠定线性系统。考虑一个小规模 CT问题 [10] ，其形式为

$$\boldsymbol{b} = A\boldsymbol{x} + \boldsymbol{e},$$

其中，真实信号 $\boldsymbol{x} \in \mathbb{R}^{100\times100}$ 为 Shepp-Logan体模，噪声满足$\boldsymbol{e} \sim \mathcal{N}(0,\ \lambda^{-1}I)$ 其中 $\lambda = 10$。此外，前向算子 A 是一个离散化拉东变换，在 0 到 180 度之间取 20 个等间隔角度，每个角度有 120 条射线，并采用平行束几何。该算子使用 AIR Tools II [11]生成。真实信号$\boldsymbol{x}$ 和含噪正弦图 $\boldsymbol{b} \in \mathbb{R}^{20\times120}$ 如图5.8所示。

请注意，这是一个稀疏角度断层成像问题，包含 10.000 个变量，且只有 2.400 个测量值。为求解 CT问题并量化不确定性，我们采用第3.2小节所述的欠定随机化框架。具体而言，我们考虑

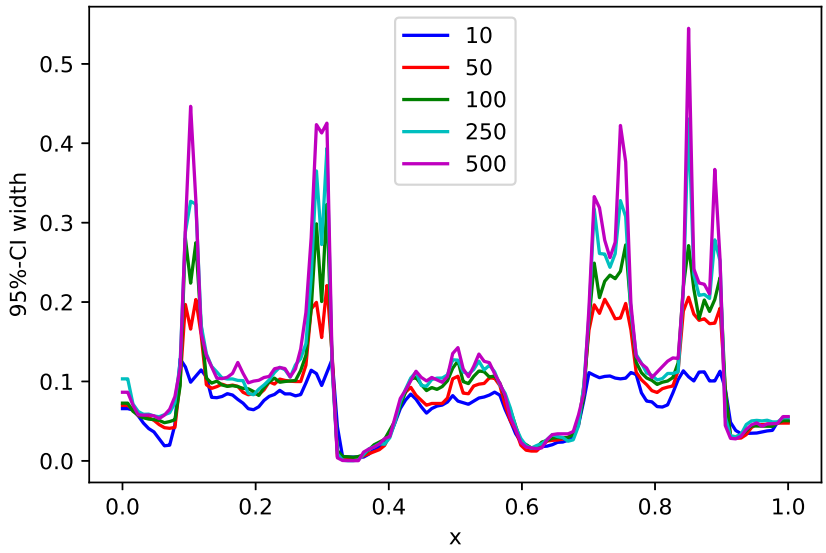

图 5.6：随着精度/迭代次数增加，95%可信区间的宽度。

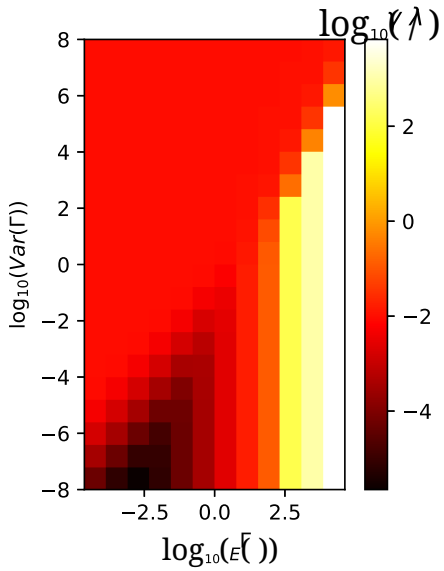

图 5.7：平均正则化强度 γ/λ 作为关于 γ 的超先验期望和方差的函数。

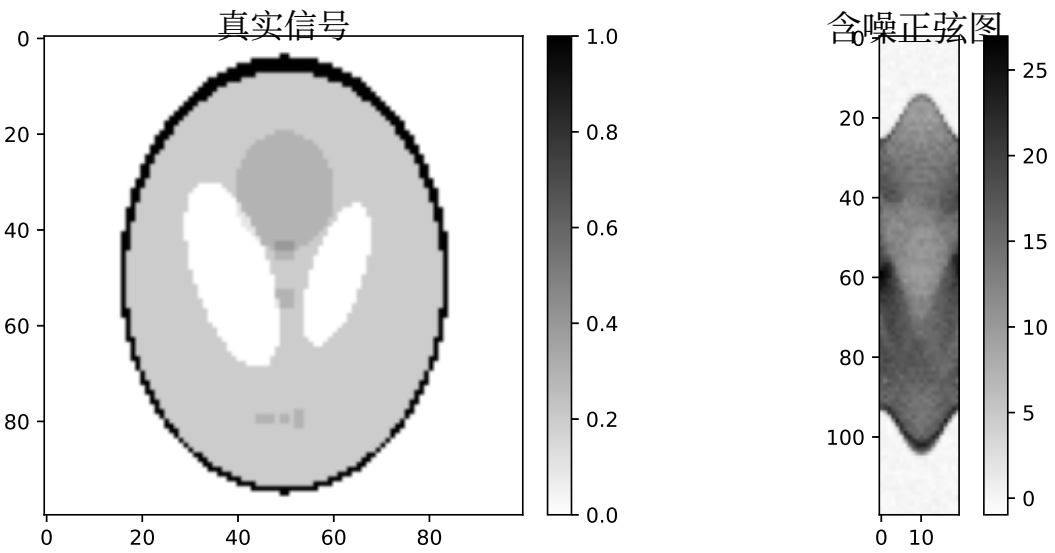

图 5.8：Shepp-Logan体模（左）以及含有 $\lambda = 10$ 的含噪正弦图（右）。

由下式定义的概率分布

$$\underset{\boldsymbol{x}\in\mathbb{R}_{+}^{100\times 100}}{\operatorname{argmin}} \left\{ \frac{\lambda}{2}\|A\boldsymbol{x} - \hat{\boldsymbol{b}}\|_2^2 + \gamma\|L\boldsymbol{x}\|_1 \right\},$$

其中 $\hat{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{b}, \lambda^{-1}I)$, $\gamma > 0$ 和 L 是二维有限差分算子。因此，我们求解一个带各向同性全变差正则化的随机化非负约束线性最小二乘问题。

图 5.9 显示了通过交替方向乘子法对 200 次迭代中的 500 个独立样本进行计算后得到的 $\gamma = 10$、50 和 100 的中位数及 95%可信区间宽度，其中计算过程使用了$\rho = 200$。请注意，对于 γ的所有取值，整体形状都得到了适当重建，但当 γ较大时，一些细节几乎被完全平滑掉了。这一点在可信区间宽度缺乏不确定性方面尤为明显。根据推论 3.2，概率分布的支撑与 γ 的选择无关。然而，正如全变分正则化所预期的那样，较大的 γ 会降低小细节出现的概率。

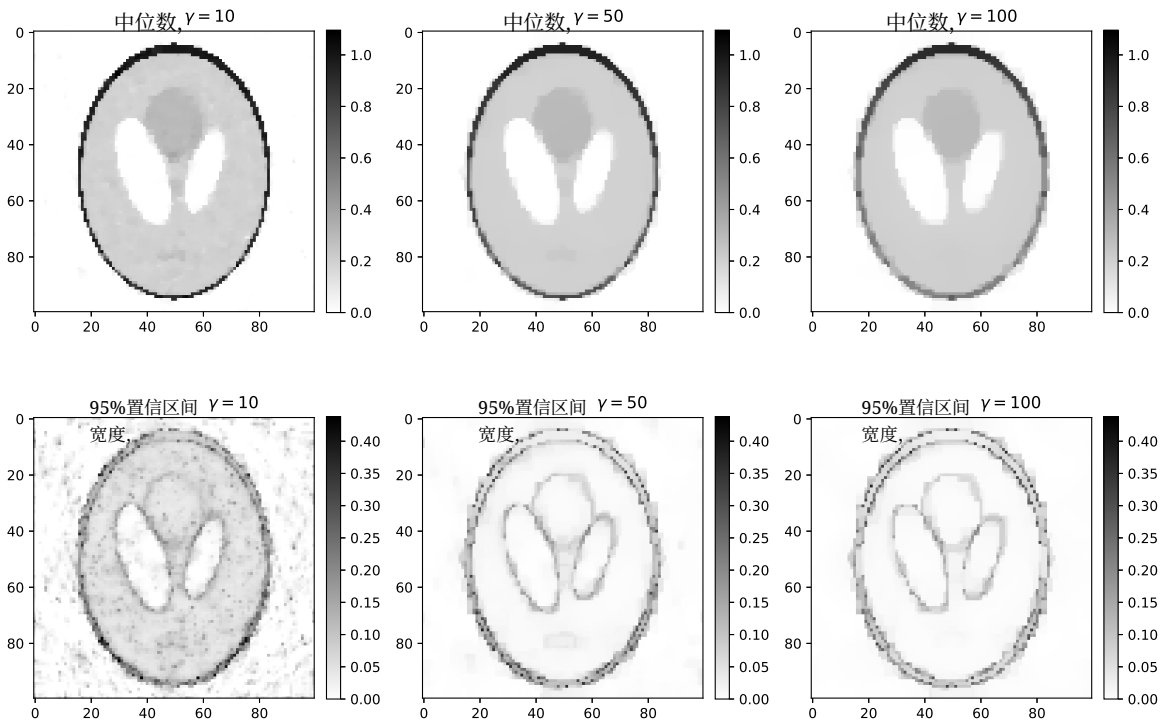

图5.9：三种不同正则化强度下各分量的中位数及95%可信区间宽度 γ。

6. 结论与进一步研究。 我们提出了一个将正则化与高斯分布相结合的框架。通过针对数据的不同高斯扰动反复求解正则化线性最小二乘问题，我们从一个概率分布中获得样本；该分布等价于使用近端算子对高斯后验分布进行后处理。如果正则化函数足够光滑，则该分布是连续分布。然而，对于稀疏性诱导正则化函数（例如全变差和套索），正则化分布会对与正则化函数所建模的稀疏性相对应的低维子空间赋予正概率。这类分布的行为类似于变维模型，但其由随机优化问题的解隐式定义。

我们将该理论应用于贝叶斯线性逆问题，并通过考察其底层先验分布，讨论了如何将正则化高斯分布解释为隐式定义的后验分布。基于这种隐式先验视角，我们为一些包含正则化高斯分布的贝叶斯层次模型推导出了吉布斯采样器；这些模型所依据的正则化函数的上图是多面锥，例如非负全变差。对于线性最小二乘项欠定的一般情形，我们还给出了一些关于该分布支撑的结果。

由于从该分布中采样只需求解正则化线性最小二乘问题，因此获取样本只需要优化理论中的工具。通过一些小规模的去模糊和计算机断层扫描示例，我们展示了我们的框架与连续分布之间的差异，并强调了用于贝叶斯层次模型的吉布斯采样器的稳健性。

从正则化分布中获取单个独立样本的计算成本，大致相当于求解一个类似的正则化变分逆问题。因此，获取大量精确的独立样本在计算上代价高昂。不过，我们通过一个小型数值示例表明，即使样本不精确，也仍能以低得多的计算成本获得良好结果。为了使这一方法更具实用性，后续研究可以更深入地探讨，不精确地求解优化问题会如何影响样本的分布。

在这项工作中，我们并未考虑所有可能的正则化函数。理论中的很大一部分聚焦于低维子空间，而对于这些子空间，我们仅给出了适用于足够光滑的函数，或次微分性质足够良好的非可微函数的结果。此外，我们所给出的理论大多依赖于线性最小二乘项为强凸的假设，因此对于一般情形，只能得出一些简单性质。进一步研究这些性质以及正则化分布更一般的性质，是未来研究的一个方向。

参考文献

[1] H. H. Bauschke、P. L. Combettes 等。 *Convex analysis and monotone operator theory in Hilbert spaces*，第408卷。Springer，2011年。[2] S. Boyd、N. Parikh、E. Chu、B. Peleato、J. Eckstein 等。通过乘子交替方向法进行分布式优化与统计学习。

Foundations and Trends in Machine ®，（）：–。*learning* 3 1 1 122, 2011

[3] C. Chen、B. He、Y. Ye 和 X. Yuan。用于多块凸最小化问题的 ADMM 直接扩展并不一定收敛。*Mathematical Programming*，155(1):57–79，2016年。[4] L. Condat。一维全变分去噪的直接算法。*IEEE Signal Processing Letters*，20 11 1054 1057，2013（）：–。[5] J. M. Everink、Y. Dong 和 M. S. Andersen。基于投影密度的贝叶斯推断。*arXiv preprint arXiv:2209.12481* 2022，。[6] K. Ewald 和 U. Schneider。论套索估计量在高低维情形下的分布、模型选择性质及唯一性。*Electronic Journal of Statistics*，14(1)，2020年。[7] S. Foucart 和 R. Holder。*A Mathematical Introduction to Compressive Sensing*。Birkh¨auser，2013年。[8] P. J. Green。可逆跳跃马尔可夫链蒙特卡罗计算与贝叶斯模型判定。*Biometrika*，82(4):711–732，1995年。[9] P. C. Hansen。*Discrete inverse problems: insight and algorithms*。工业与应用数学学会，2010年。[10] P. C. Hansen、J. Jørgensen 和 W. R. Lionheart。

Computed Tomography: Algorithms Insight and Just Enough Theory，，。工业与应用数学学会，2021。[11] P. C. Hansen 和 J. S. Jørgensen。AIR Tools II：代数迭代重建方法的改进实现。*Numerical Algorithms*，79(1):107–137，2018年。[12] F. Orieux、O. F´eron 和 J.-F. Giovannelli。针对一般线性逆问题的高维高斯分布采样。 *IEEE Signal Processing Letters*，19(5):251–254，2012年。[13] N. Parikh、S. Boyd 等。近端算法。*Foundations and trends in Optimization* 1 3 127 239,

®，（）：–。2014[14] C. P. Robert 和 G. Casella。 *Monte Carlo statistical methods*，第2卷。Springer，1999年。[15] R. T. Rockafellar 和 R. J.-B. Wets。 *Variational analysis*，第317卷。施普林格科学与商业媒体，2009年。

[16] W. Rudin。 *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc.，美国，1987年。[17] A. Schreck、G. Fort、S. Le Corff 和 E. Moulines。用于贝叶斯变量选择的收缩阈值梅特罗波利斯调整朗之万算法。 *IEEE Journal of Selected Topics in Signal Processing*，10 2 366( ): –375，2015年。[18] S. Waldmann。 *Topology: an introduction*。Springer，2014年。[19] Z. Wang、J. M. Bardsley、A. Solonen、T. Cui 和 Y. M. Marzouk。基于 l_1 先验的贝叶斯反问题：先随机化再优化方法。 *SIAM Journal on Scientific Computing*，39(5):S140–S166，2017年。

附录 A. 算法。

考虑求解该优化问题

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \sum_{i=1}^{k} f_i(L_i\boldsymbol{x}) \right\}, \tag{A.1}$$

其中，函数 $f_i : \mathbb{R}^{c_i} \to \mathbb{R}$ 是正常下半连续函数，且其近端算子 $\operatorname{prox}_{f_i}^{\sigma}$ 可以高效计算。

对于本文所述的方法，反复求解 (A.1)，并处理不同的函数 f_i，是主要的计算问题。在第 5 节的数值实验中，我们使用了交替方向乘子法（ADMM）的一种变体 [2]。为此，将无约束问题 (A.1) 改写为以下约束问题：

$$\begin{aligned} &\text{Minimize} \quad \left\{ \frac{1}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \sum_{i=1}^{k} f_i(\boldsymbol{y}_i) \right\}, \\ &\text{subject to} \quad \begin{bmatrix} L_1 \\ \vdots \\ L_k \end{bmatrix} \boldsymbol{x} - \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_k \end{bmatrix} = \boldsymbol{0}. \end{aligned} \tag{A.2}$$

与 (A.2) 对应的 ADMM 算法需要计算 $\sum_{i=1}^{k} f_i(\cdot)$ 的近端算子。由于它是可分离和，因此可以将其简化为分别计算各个 f_i 项的近端算子 [13, 3]。由此得到的算法总结于算法 A.1。

Input: $A, \boldsymbol{b}, k, \{(\operatorname{prox}_{f_i}^{\sigma}, L_i)\}_{i=1}^{k}, \rho, \{(\boldsymbol{y}_i, \boldsymbol{u}_i)\}_{i=1}^{k}$
while Convergence criterion not met **do**
  Compute $\boldsymbol{x} \leftarrow \operatorname{argmin}_{\boldsymbol{z}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \frac{\rho}{2}\sum_{i=1}^{k}\|L_i\boldsymbol{x}-\boldsymbol{y}_i+\boldsymbol{u}_i\|_2^2 \right\}$
  Compute $\boldsymbol{y}_i \leftarrow \operatorname{prox}_{f_i}^{\sigma/\rho}(L_i\boldsymbol{x}+\boldsymbol{u}_i)$ for $i = 1, \ldots, k$
  Compute $\boldsymbol{u}_i \leftarrow \boldsymbol{u}_i + L_i\boldsymbol{x} - \boldsymbol{y}_i$ for $i = 1, \ldots, k$
end while
return $\boldsymbol{x}$

算法 A.1：用于可分离正则化项的交替方向乘子法。

为高效计算近端算子 $\text{prox}_{f_i}^{\sigma}$，我们有以下几种情况。如果 f_i 是闭凸集的示性函数，则近端算子可简化为对该集合的欧几里得投影。对于各种实际集合，例如具有逐分量界限的集合和球，计算此类投影都很高效。如果 $f_i(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$，则近端算子对应于软阈值算子 [1]。

全变差，即 $g(\boldsymbol{x}) = \|L_i\boldsymbol{x}\|_1$ 与一阶有限差分矩阵 $L_i \in \mathbb{R}^{(n-1)\times n}$，可以通过不同方式建模。如果考虑 $g(\boldsymbol{x}) = f_i(L_i\boldsymbol{x})$，那么它符合上文提到的 l_1 情形。或者，对于一维信号，近端算子 $\text{prox}_{\|L_i\cdot\|_1}^{\sigma}$可以高效计算，参见 [4]。

附录 B. 杂项引理与证明。

引理 B.1. 设 μ 是定义在 $\mathbb{R}^n$ 上的概率测度，且 $g\colon\ D \subset \mathbb{R}^n \to \mathbb{R}^n$ 连续，则

$$\text{supp}(\mu \circ g^{-1}) = \text{cl}(g(\text{supp}(\mu))).$$

证明。 设 $\boldsymbol{y} = g(\boldsymbol{x}) \in g(\text{支撑}(\mu))$，并设 $\mathbf{N}_{\boldsymbol{y}}$ 是 $\boldsymbol{y}$的任意开邻域，则

$$(\mu \circ g^{-1})(\mathbf{N}_{\boldsymbol{y}}) = \mu(g^{-1}(\mathbf{N}_{\boldsymbol{y}})) = \mu(\hat{\mathbf{N}}_{\boldsymbol{x}}) > 0,$$

其中 $\hat{\mathbf{N}}_{\boldsymbol{x}}$ 是 $\boldsymbol{x}$的一个开邻域。由于 $\mu \circ g^{-1}$ 的支撑是闭集，我们得到 $\text{cl}(g(\text{支撑}(\mu))) \subseteq \text{支撑}(\mu \circ g^{-1})$。

设 $\boldsymbol{y} \in \text{supp}(\mu \circ g^{-1})$，并假设 $\boldsymbol{y} \notin \text{cl}(g(\text{supp}(\mu)))$，则存在一个围绕 $\boldsymbol{y}$ 的开邻域 $\mathbf{N}_{\boldsymbol{y}}$，使得 $\mathbf{N}_{\boldsymbol{y}} \cap g(\text{supp}(\mu)) = \varnothing$。因此，$g^{-1}(\mathbf{N}_{\boldsymbol{y}}) \cap \text{supp}(\mu) = \varnothing$。这意味着 $(\mu \circ g^{-1})(\mathbf{N}_{\boldsymbol{y}}) = 0$，与 $\boldsymbol{y} \in \text{supp}(\mu \circ g^{-1})$ 矛盾。因此我们可以得出结论：$\text{supp}(\mu \circ g^{-1}) \subseteq \text{cl}(g(\text{supp}(\mu)))$。 ■

引理 B.2。 设 $f\colon\ \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是一个局部 *Lipschitz* 连续函数，且 $\text{dom}(f)$为开集。设 $\mathbf{U} \subseteq \text{dom}(f)$ 满足 $\boldsymbol{Leb}(\mathbf{U}) = 0$，则 $\boldsymbol{Leb}(f(\mathbf{U})) = 0$，即局部 *Lipschitz* 连续函数将勒贝格零测集映射为勒贝格零测集。

证明。 由 [16, 引理 725] 可知，Lipschitz 连续函数保持勒贝格零测集不变。若 f 不是 Lipschitz 连续函数，则考虑覆盖域 $(f) = \cup_{i\in\mathbb{N}}\mathbf{K}_i$，其中 $\mathbf{K}_i$ 是一个递增的紧集序列。由于

$$\text{Leb}(f(\mathbf{U})) \leqslant \sum_{i\in\mathbb{N}} \text{Leb}(f(\mathbf{U} \cap \mathbf{K}_i)),$$

只需证明对于任意紧集 $\mathbf{K}$，$\text{Leb}(f(\mathbf{U} \cap \mathbf{K})) = 0$ 即可。

由于 f 是局部 Lipschitz 连续函数，对于每个 $\boldsymbol{u} \in \mathbf{U} \cap \mathbf{K}$，都存在一个开球$\mathbf{B}(\boldsymbol{u}, \epsilon)$ 在该开球上是 Lipschitz 连续的。此外，由于所有这些球构成了紧集 $\text{cl}(\mathbf{U} \cap \mathbf{K}) \subset \text{dom}(f)$的一个开覆盖，因此存在一个有限子覆盖 $\{\mathbf{B}(\boldsymbol{u}_i, \epsilon_i)\}_{i=1}^{k}$。因此，我们可以得出

$$\text{Leb}(f(\mathbf{U} \cap \mathbf{K})) \leqslant \sum_{i=1}^{k} \text{Leb}(f(\mathbf{U} \cap \mathbf{K} \cap \mathbf{B}(\boldsymbol{u}_i, \epsilon_i))) = 0.$$

■

推论 B.3。 设 f： $\mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是一个局部 *Lipschitz* 连续函数，且 $\mathrm{dom}(f)$是开集，并设 μ 是上的博雷尔测度。如果 $\mu \ll$ 是勒贝格测度，那么 $\mu \circ f \ll$ 也是勒贝格测度，即如果 μ 相对于勒贝格测度具有密度，那么 $\mu \circ f$也具有密度。

引理 *B.4.* 设 $U \subseteq \mathbb{R}^k$ 和 $V \subseteq \mathbb{R}^l$ 为带有连续函数 $\phi : U \times V \to \mathbb{R}^n$ 和 $\psi : U \times V \to \mathbb{R}^n$ 的紧集。若由 $h(\boldsymbol{u}、\ \boldsymbol{v}) = \phi(\boldsymbol{u}、\ \boldsymbol{v}) + \psi(\boldsymbol{u}、\ \boldsymbol{v})$ 定义的 $h : U \times V \to \phi(U,V) + \psi(U, V)$ 是单射，则 h 是 $U \times V$ 与闵可夫斯基和 $\phi(U,V) + \psi(U, V)$ 之间的同胚。

证明。 注意， h 是从紧集 $\mathbf{U} \times \mathbf{V}$ 到豪斯多夫空间 $\phi(\mathbf{U}、\ \mathbf{V}) + \psi(\mathbf{U}、\ \mathbf{V}) \subset \mathbb{R}^n$的双射连续函数，因此根据 [18, 命题 5.2。5]， h 是一个同胚。 ■