arXiv:2305.04826v1 [stat.ME] 8 May 2023

, , 第 1–22 页

doi：DOI HERE提前在线发表日期：日 月 年 论文

论文

基于峰值持久性图从含噪数据估计形状和函数

Woo Min Kim,[1] Sutanoy Dasgupta[2] 和 Anuj Srivastava[1,*]

[1]佛罗里达州立大学统计系，塔拉哈西，佛罗里达州 32306，美国； [2]德克萨斯农工大学统计系，学院站，德克萨斯州 77843，美国

*通讯作者。anuj@stat.fsu.edu 仅供出版方填写 收稿日期：年 月 日；修订日期：年 月 日；录用日期：年 月 日

摘要

从含噪数据中估计其背后的信号，是统计学和工程学中的一项重要问题。根据观测模型和估计准则，文献中已有许多可用的估计量。本文提出一种框架，用于估计未知信号的形状以及信号本身。该方法采用峰持久性图（PPD）这一新颖工具，探索潜在解中的主导峰值，并估计函数的形状，包括内部峰值和谷值的数量。随后，它将此形状约束施加于搜索空间，并根据部分对齐的数据估计信号。该方法兼顾了两种以往的方案：不进行对齐的平均，以及进行完全弹性对齐的平均。从统计学角度看，在同时包含加性噪声和相位噪声或扭曲噪声的模型下，该方法能够实现最优估计量。我们还提出了一种计算效率高的实现方法，并通过若干模拟和真实案例展示其有效性。值得注意的是，这种几何方法的表现优于该领域当前最先进的方法。

关键词：函数型数据分析、函数估计、峰值持久性图、SRVF、峰值约束曲线估计、形状估计。

引言

分析噪声数据以估计底层信号或某些相关属性，是统计学和工程学中的一个基本问题。通常将观测数据建模为受到噪声干扰的实际底层信号，目标是利用数据估计该信号。估计器的选择取决于观测模型、噪声分布、数据结构以及优化准则。经典方法是

函数估计通常是先确定一个函数空间（通常是带有正交归一基的希尔伯特空间），设定一个目标函数，再在该函数空间上对其进行优化。我们的方法略有不同，将重点放在函数的形状上，并引入一种新工具——峰持久性图（PPD），用于搜索估计结果。借助 PPD，我们可以先估计未知函数的形状，再对函数本身进行形状约束估计。

© 作者。由牛津大学出版社出版。保留所有权利。如需授权，请发送电子邮件。
journals.permissions@oup.com

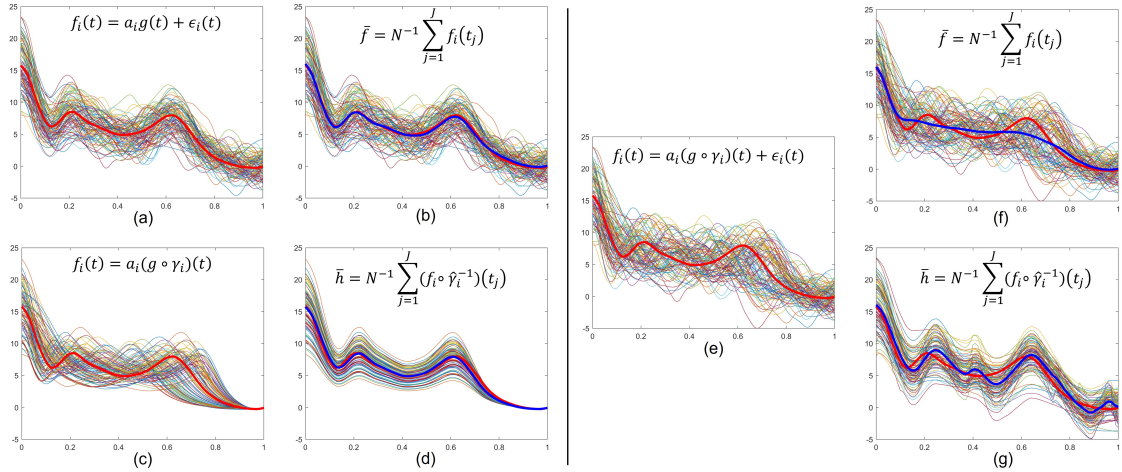


图 1. 粗红线表示 g，(b)、(f) 中的蓝线表示 $\bar{f}$，而 (d)、(g) 中的蓝线表示 $\bar{h}$。(a) 中的函数仅含加性噪声，(c) 中的函数仅含相位噪声。(e) 中的函数同时含有加性噪声和相位噪声。(b) 和 (d) 中的估计效果良好，而 (f) 和 (g) 中的估计效果不佳。

我们先从函数估计问题的传统视角出发，随后论证采用几何（基于形状）方法的理由。函数估计的一种简单传统模型是**加性噪声模型**：

$$f_i(t) = a_i g(t) + \epsilon_i(t) \ , i = 1, 2, \ldots, n \ , \qquad (1)$$

其中 $g \in \mathcal{F}$ 是未知信号，$\epsilon_i \in \mathcal{F}$ 是均值为零的独立噪声，$a_i \in \mathbb{R}$ 是均值为一的独立随机变量。这里，$\mathcal{F}$ 表示所关注的函数空间；最常见的是，$\mathcal{F}$ 是区间上平方可积函数构成的希尔伯特空间。在一种简单情形下，对于密集时间采样，

$$mean\ \bar{f} \overset{g}{=} \frac{1}{n}\sum_{i=1}^{n} f_i \qquad \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \overset{cross\text{-}sectional}{\to} 0$$

的估计由 给出。因为 且 $\frac{1}{n}\sum_{i=1}^{n} a_i \Rightarrow 1$。图 1 左上部分以图示方式展示了这一估计。图 (a) 显示数据 $\{f_i\}$ 和真实函数 g （红色），图 (b) 显示叠加在蓝色曲线上的截面均值。对于稀疏数据或其他挑战，可以使用 g 的光滑基，或在标准最小二乘目标函数中加入粗糙度惩罚（或先验）（Ramsay 和 Silverman（1997）；Fan 和 Gijbels（1996）；Choudhuri 等人（2005）；Green 和 Silverman（1994））。

近年来，人们日益认识到函数型数据中存在另一种噪声或变异，即所谓的 $phase$ 或 $compositional$噪声。这种噪声的观测模型由**相位噪声模型**给出：

$$f_i(t) = g(\gamma_i(t)) \ , i = 1, 2, \ldots, n \ , \qquad (2)$$

其中，γ_i是定义域上的随机时间扭曲函数。（时间扭曲函数的精确定义将在后文给出。）为了使 g可识别，假设相位噪声具有恒等均值，即 $i.\ e.,\ \frac{1}{n}\sum_{i=1}^{n} \gamma_i(t) \to t$。可以验证，在这种情况下，横截面均值 $\bar{f}$ 并不能直接估计 g。这是因为一般而言，$\frac{1}{n}\sum_{i=1}^{n} g(\gamma_i(t))$不会收敛到 $g(t)$，尽管 ($\frac{1}{n}\sum_{i=1}^{n} \gamma_i(t)) \mapsto t$。量 $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(\gamma_i(t))$可近似为$\int g(t + x)\pi(x)\ dx$，其中 π 是正态密度，因此结果是对 g进行高斯模糊。估计方法是显式求解γ_i，然后根据 $\{\tilde{f}_i = f_i \circ \hat{\gamma}_i^{-1}\}$对给定数据进行对齐。文献中也将这种对齐称为 $phase\text{-}amplitude\ separation$（Ramsay 和 Li (1998)；Srivastava 和 Klassen (2016)；Srivastava 等人 (2011)；Marron 等人 (2015)）。随后，可以通过对 $aligned\ functions\frac{1}{n}\sum_{i=1}^{n} \tilde{f}_i$ 进行横截面平均来估计 g （Kurtek 等人 (2011)）。我们将这个平均量称为 $fully\text{-}elastic$ 均值。图 1 左下部分以图示方式展示了这种估计。图 (c) 显示了数据 $\{f_i\}$ 和真实函数 g （红色），图 (d) 则显示了˜对齐函数 $\{f_i\}$ ，并以蓝色叠加显示其均值。文献中有多种函数对齐方法（例如，Kneip 和 Ramsay (2008)；Liu 和 Muller (2004)；Marron 等人 (2015)），其中一种尤为高效的方法基于非参数 Fisher–Rao 度量（ Srivastava 和 Klassen (2016)；Srivastava 等人 (2011)）。该方法使用函数的平方根表示以及动态

规划（DP）算法来估计 $\{\gamma_i\}$，并且在对齐f_i的峰值和谷值方面非常成功。(d) 中所示的示例使用了这种方法。

问题说明

本文聚焦于一个更一般的问题：数据中存在$both$ 加性噪声和相位噪声。现在，观测模型由**加性噪声和相位噪声模型**给出：

$$f_i(t) = a_i(g \circ \gamma_i)(t) + \epsilon_i(t),\ i = 1, 2, \ldots, n\ . \quad (3)$$

模型的各个组成部分与之前相同。接下来，我们讨论处理该模型的当前最先进方法。

以往方法

已有一些研究探讨了在公式 3 所指定的模型下进行函数估计。一种看似自然的估计 g 的方法是采用最小二乘解：

$$\underset{g\in\mathcal{F}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \min_{a_i\in\mathbb{R},\gamma_i\in\Gamma} \left(\|f_i - a_i(g\circ\gamma_i)\|^2 + \kappa\mathcal{R}(\gamma_i) \right) \right), \quad (4)$$

其中，$\mathcal{R}$ 表示对 γ_is施加的粗糙度惩罚。然而，这种优化存在若干问题。首先，对于$\kappa = 0$，关于 γ_is的优化是退化的；这一现象被称为 $pinching\ \ effect$ （Ramsay 和 Silverman（1997）；Srivastava 和 Klassen（2016））。虽然添加惩罚项（设定 $\kappa > 0$）可以避免退化，但并不能彻底解决问题。它会限制解的搜索空间，并可能带来意料之外的后果。此外，确定公式 4 中惩罚参数 κ 的最优值是一个公认的难题。 κ 的不同取值可能导致截然不同的解（第 2 节将给出相关示例）。

一些早期论文（Kneip 和 Ramsay (2008)；Ramsay 和 Li (1998)；Tang 和 Muller (2008)；Earls 和 Hooker (2017)）使用了公式 4，但侧重于函数型数据的对齐。Marron 等人 (2015) 对函数对齐方面的论文进行了综述。另一些研究通过在公式 3 中加入随机效应项，探讨了混合效应模型（Raket 等人 (2014)；Claeskens 等人 (2021)）。图像配准领域也研究了类似问题（Allassonni´ere 等人 (2007)；Simpson 等人 (2012)），不过重点在于保留视觉特征。几乎所有这些论文都采用 $\mathbb{L}^2$ 目标函数，因此必须平衡公式 4 中的两个项。此外，一些论文采用贝叶斯方法，通过对给定数据 $\{f_i\}$施加先验 γ_is 来进行对齐（Kurtek (2017)）。还有一些论文主张先对给定函数进行平滑，然后计算平均值（Cheng 等人 (2013)；Lu 等人 (2017)；Telesca 和 Inoue (2008)）。然而，这些论文并未显式估计信号 g。也可以对 g 施加先验，以构建贝叶斯解法，例如 Matuk 等人 (2021)；Horton 等人 (2020)。要使这种方法有效，需要关于 g的特定先验信息。

我们将采用几何方法，并讨论相关文献。一些论文研究了欧氏曲线“形状”的所谓 Procrustes 均值（St¨ocker 等，2022），但并未提供 g的估计量。此外，还有关于形状约束密度估计的文献，但通常局限于狭窄的形状类别。这一庞大的研究领域始于 Grenander（1956），随后出现了多项相关研究（Wang 和 Berger（2016）；Cheng 等（1999）；Wegman（1970）；Doss 和 Wellner（2016））。这些研究主要局限于单峰或对数凹形状类别，而不是一般形状。关于该领域的最新综述，请参阅专刊（Samworth 和 Sen（2018））。我们注意到，Dasgupta 等（2021）将该问题推广到了通常的形状类别，但仍将估计局限于密度。本文研究任意形状约束下的一般函数。

最后，我们提及有关估计给定数据所对应概率密度的众数数量的文献较为有限。有一些论文（例如参见 Minnotte（1997））研究了密度函数中多个众数的估计，但以往的大多数文献都集中于单峰函数这一狭窄情形。我们的问题与密度估计不同，因为我们处理的是函数型数据。

我们的方法

我们的方法采用一种关注函数形状的几何视角。其动机在于，方程 1、2 是方程 3 的特例。我们旨在在非弹性与完全弹性之间取得平衡，因为这两种方案都不理想。横截面均值会过度平滑数据，而完全弹性均值甚至会对齐噪声伪影，导致估计结果中出现伪峰和谷值。图 1 右侧展示了一个示例：对于 (e) 中的数据，图 (f) 显示非弹性均值，图 (g) 显示弹性均值。我们认为，理想的方案应处于这两个极端之间，并具有自适应性。为此，我们提出采用部分对齐，通过参数 $\lambda \in \mathbb{R}_+$ 控制对齐过程中函数的弹性。由此得到的˜部分对齐函数记为 $f_{\lambda,\ i}$，其均值记为 $\hat{}g_\lambda = \sum_{i=1}^{n} \tilde{f}_{\lambda,\ i}$。一个重要的考量是根据给定数据选择具有自适应性的 λ。此外，我们还必须处理捏缩效应导致的退化问题。为解决这些问题，我们采用

形状分析技术。我们通过改变弹性参数来估计 g 和 g的形状，具体如下：

1. **峰持久性图**（PPD）：本阶段的目标是通过识别对相位噪声保持不变的几何特征，估计 g 的形状。为此，我们开发了一种名为峰持久性图（PPD）的新工具。当改变参数 λ时，平均函数$\hat{g}_\lambda$ 的形状会发生变化。通常，随着λ 增大，峰值和谷值等几何特征会逐渐平滑消失。因此，可以通过研究$\hat{g}_\lambda$ 的内部峰值相对于 λ的持久性，并选择最持久的峰值，来确定 λ的合理取值。这种持久性的图形表示称为峰持久性图，即PPD。我们采用一组启发式方法来定义在PPD 中选择最持久峰值的标准。该过程会产生三个量：最优 λ^*、持久内部峰值的数量，以及 $f_{\lambda^*,i}$，后者是对应最优权重 λ^* 的部分对齐数据。

2.**形状约束函数估计**：前一步生成了一个部分对齐的均值 $\hat{g}_{\lambda^*}$。它属于正确的形状类别，但从严格意义上说，可能并非 g 的最优估计。因此，在第二步中，我们旨在找到所选类别中的最优元素。我们将 g 的估计限制为恰好具有 m 个内部峰值，并利用几何方法对 g进行形状约束估计。估计过程涉及使用数据 $\tilde{f}_{\lambda^*,i}$，在适当的约束空间 $\mathcal{F}_m$上最小化平方误差。这里，$\mathcal{F}_m$ 指的是 $\mathcal{F}$ 中所有具有正确形状$i.e.$且有 m 个内部峰值的元素所构成的集合。此方法不同于以往的形状约束估计量，因为它在 $\mathcal{F}_m$中提供的是惩罚最大似然估计，而非 $\mathcal{F}_m$中的任意元素。

本文仅关注 g 的 $internal$峰值和谷值。边界处的潜在峰值也很重要，但可以通过简单的测试单独检测，因此本文不予考虑。

惩罚型 $\mathbb{L}^2$ 估计量

在阐述所提出的框架之前，我们进一步说明基于公式 4 求解时面临的问题。主要有两个问题：（1）对于 $\kappa = 0$，问题是退化的，从而产生 $pinchingeffect$；（2）应如何设定 κ 的值，以避免过度平滑和捏缩？尽管文献中关于如何选择平滑参数已有若干思路，但捏缩效应使这一选择变得复杂。通常需要采用较高的 κ 值来避免捏缩，但这可能导致过度平滑。我们用一个采用一阶粗糙度惩罚 $R(\gamma_i) = \int(1-\sqrt{\dot{\gamma}_i(t)})^2 dt$ 的例子来说明这一点。（详细算法见 Srivastava 和 Klassen（2016）。）图 2 (a) 展示了给定的函数 $\{f_i\}$ 及其横截面均值 $\bar{f}$（以蓝色显示）。红色曲线表示真实值 g，图 (b)-(e) 中的蓝色曲线表示对齐曲线的横截面均值，记为$\hat{g}_{L^2}$。在 $\kappa = 0 \sim 1$时，可以看到估计结果中的捏缩效应；但随着 κ 增大，估计结果变得更加平滑。这个例子表明，需要谨慎选择 κ ，才能获得对 g的良好估计。挑战在于找到一种自动化方法，以实现最优的 κ，同时避免捏缩效应。

我们的基于形状的框架与方程 4 不同：该框架采用 Fisher-Rao 距离而非 $\mathbb{L}^2$ 距离进行构建，并完全避免了挤压问题。不过，如何选择平滑参数的问题依然存在，我们将使用 PPDs 来解决这一问题。在本文余下部分中，我们假设函数型数据定义在紧区间 I 上，且 $\mathcal{F}$ 是所有绝对连续实值函数在I上的集合。我们将考虑方程 3 中所述的模型，其中包含以下相互独立的分量：(1) 缩放噪声：$a_i \in \mathbb{R}^+ \sim N(1, \sigma_a^2)$；(2) 加性噪声：$\epsilon_i \in \mathcal{F}$，一个均值为 $E[\epsilon_i(t)] = 0$（对所有 $t \in I$）的随机函数；(3) 相位噪声：$\gamma_i \in \Gamma$（其中 Γ定义如下），其均值为 $E[\gamma_i(t)] = \gamma_{id}(t) = t$。给定一组独立观测 $\{f_i\}$，我们的目标是估计 g 中的内部峰值数量以及函数 g本身。

步骤 1：使用 PPD 进行形状估计

这里的目标是估计 g、$i.e.$的形状，根据给定的 $\{f_i\}$ 估计 g 中内部峰值和谷值的数量，并形成 g 的初始估计。

背景材料：我们概述函数在 Fisher–Rao度量下进行部分弹性对齐的一种机制，详情请参阅 Srivastava 和 Klassen（2016）。为进行对齐，函数 $f \in \mathcal{F}$ 由其平方根速度函数（SRVF）表示：$q(t) = \text{sign}(\dot{f}(t))\sqrt{|\dot{f}(t)|}$。令 Γ 表示 I的所有保边微分同胚组成的群。任意元素 $\gamma \in \Gamma$ 都是一个光滑函数，且其逆函数也光滑，并保持 I的边界不变。集合Γ 在复合运算 $i.e.$下构成一个群；对于任意γ_1，$\gamma_2 \in \Gamma$，都有 $\gamma_1 \circ \gamma_2 \in \Gamma$。这个群有一个

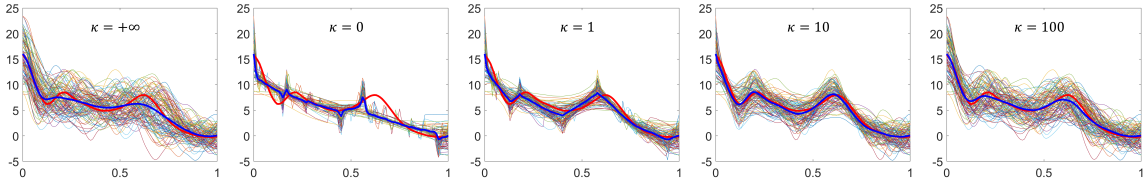


图 2. 根据式 4 对给定函数进行对齐，并估计 g。图 (a) 展示原始函数及其均值（蓝色）。图 (b)-(e) 展示不同 κ 下的真实 g（红色）、部分对齐后的函数集以及估计结果（蓝色曲线）。

单位元 $\gamma_{id}(t) = t$。令函数 $f \in \mathcal{F}$ 经任意 $\gamma \in \Gamma$进行时间扭曲，得到 $f \circ \gamma$。扭曲后函数的 SRVF 为 $(q \circ \gamma)\sqrt{\dot{\gamma}}$，为简洁起见，我们将其记作 $q \star \gamma$。

在此设置下，我们可以如下求解任意两个函数 f_1, $f_2 \in \mathcal{F}$ 的弹性对齐。设q_1, q_2 分别表示 f_1, f_2的 SRVF。则将 f_2 对齐到 f_1 的最优时间扭曲为：$\gamma^* = \operatorname{argmin}_{\gamma\in\Gamma} \|q_1 - (q_2 \star \gamma)\|^2$，其中 $\|\cdot\|$ 表示 $\mathbb{L}^2$ 范数。不过，如果我们想对对齐中的弹性程度进行惩罚，可以使用：$\operatorname{argmin}_{\gamma\in\Gamma} \left(\|q_1 - (q_2 \star \gamma)\|^2 + \lambda\mathcal{R}(\gamma)\right)$。项 $\mathcal{R}(\gamma)$表示对 γ 粗糙度的惩罚，并使其接近恒等元素γ_{id}。（本文选择了一阶惩罚 $\mathcal{R} = \|1 - \sqrt{\dot{\gamma}}\|^2$，但也可以使用其他惩罚项。）常数 $\lambda > 0$ 控制对齐中的弹性程度：$\lambda = 0$ 表示完全弹性，$\lambda = \infty$ 表示非弹性。要对多个函数进行对齐，例如 f_1, f_2,..., f_n，令 q_1, q_2,..., q_n分别表示它们对应的 SRVF。然后，使用以下优化方法对这些进行联合对齐：

$$\min_{q\in\mathbb{L}^2} \left[\sum_{i=1}^{n} \left(\min_{\gamma_i\in\Gamma} \left(\|q - (q_i \star \gamma_i)\|^2 + \lambda\mathcal{R}(\gamma_i)\right)\right)\right] . \tag{5}$$

利用以下事实，我们可以重新整理该方程：$\|q_1 - q_2 \star \gamma\| = \|q_1 \star \gamma^{-1} - q_2\|$，对所有 q_1, $q_2 \in \mathbb{L}^2$ 和 $\gamma \in \Gamma$均成立，且 Γ 是一个群。由此可得：

$$\min_{q\in\mathbb{L}^2} \left[\sum_{i=1}^{n} \left(\min_{\gamma_i\in\Gamma} \left(\|q_i - (q \star \gamma_i)\|^2 + \lambda\mathcal{R}(\gamma_i)\right)\right)\right] ,$$

这可以直接与方程 4 进行比较。（注意，在此我们忽略对 a_i s 的优化，因为在任何基于均值的估计中，它们都会相互抵消。）两种解法的主要区别在于 SRVF 的使用以及弹性度量的不变性。从根本上说，这种差异源于以下事实：对所有 $q \in \mathbb{L}^2$ 和 $\gamma \in \Gamma$，$\|q\| = \|q \star \gamma\|$，而一般情况下 $\|f\| \neq \|f \circ \gamma\|$。

基于 SRVF 的优化（公式 5）不会直接给出 g 的估计值，而是得到

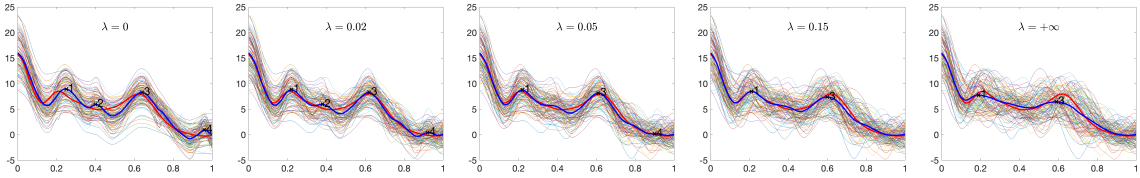
若 $\hat{\gamma}_{\lambda,\,i}$ 表示求和项内部的最优时间扭曲，对于每个i，则 $f_{\lambda,\,i} = f_i \circ \hat{\gamma}_{\lambda,\,i}$ 是所得的部分对齐函数；(2) 令$\hat{g}_\lambda$ 为这些 $\tilde{f}_{\lambda,\,i}$ 的横截面均值。算法 1 总结了这种部分弹性对齐的主要步骤。图 3

算法 1 函数对齐算法

要求： 数据 $\{f_1, f_2, \ldots, f_n\}$
1: **对于** $i = 1, \ldots, n$ **do**
2: $q_i \leftarrow$符号$(\dot{f}_i)\sqrt{|\dot{f}_i|}$
3: **结束循环**
4: $\bar{q} \leftarrow$ 最小化$_{q\in\{q_1,\ldots,q_n\}} \|q - \frac{1}{n}\sum_{i=1}^{n} q_i\|$
5: **</style>当** $\epsilon > tol$ **时执行**
6: **对于** $i = 1, 2, \ldots, n$ **do**
7: $\hat{\gamma}_{\lambda,\,i} \leftarrow \arg\inf_{\gamma\in\Gamma} \left(\|\bar{q} - (q_i \star \gamma)\|^2 + \lambda\mathcal{R}\right)$
8: $f_i^* \leftarrow f_i \circ \hat{\gamma}_{\lambda,i}$, and $q_i^* \leftarrow$ 符号$(\dot{f}_i^*)\sqrt{|\dot{f}_i^*|}$
9: **结束循环**
10: $\bar{q}^* \leftarrow \frac{1}{n}\sum_{i=1}^{n} q_i^*$, $\epsilon \leftarrow \|\bar{q} - \bar{q}^*\|^2$
11: **如果** $\epsilon > tol$ **那么**
12: $\bar{q} \leftarrow \bar{q}^*$
13: **如果结束则**
14: **结束 while**
15: $\bar{\gamma}^{-1} \leftarrow (\frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_{\lambda,i})^{-1}$, $\bar{q}^* \leftarrow (\bar{q}^* \star \bar{\gamma}^{-1})$
16: **对** $i = 1$、2、. . .、n **执行**
17: $\hat{\gamma}_{\lambda,i} \leftarrow$ 参数下确界$_{\gamma_i\in\Gamma} \left(\|\bar{q}^* - (q_i \star \gamma_i)\|^2 + \lambda\mathcal{R}\right)$
18: $f_{\lambda,i} \leftarrow f_i \circ \hat{\gamma}_{\lambda,i}$
19: **结束循环**
20: **返回** $\hat{g}_\lambda = \frac{1}{n}\sum_{i=}^{n}{}_{1} \tilde{f}_{\lambda,i}$

展示了算法 1 输出结果的示意图。图中给出了多个结果，每个结果都展示了在不同 λ 值下的部分对齐函数及其横截面均值。原始数据显示在右下方的面板中。极端情况下，$\lambda = 0$ 会使峰值和谷值完美对齐。相反，$\lambda = \infty$ 则完全不进行对齐。

在此背景下，接下来的任务是找到一种自动化方法，从现有数据中选出最佳的 λ。这可以通过利用



(a) (b) (c) (d) (e)

图 3. 根据公式 5，在不同的 λ 取值下进行部分弹性对齐。每幅图展示真实的 g （红色）、部分对齐后的集合 $\{\tilde{f}_{\lambda,i}\}$以及它们的横截面均值$\hat{g}_\lambda$ （蓝色曲线）。

下文将介绍的峰值持久性图来实现。

峰值持久性图

随着 λ 从 0 变为 ∞, $\hat{g}_\lambda$ 中的一些峰值和谷值开始减弱，甚至完全消失。有时也会产生新的峰值。我们使用峰值持久性图对这一现象进行形式化描述。

定义 1 一组函数 $\{f_i, i = 1, 2, \ldots, n\}$ 的峰值持久性图（PPD），是对其部分弹性均值$\hat{g}_\lambda$ 中**显著的内部峰值**相对于 λ所作的可视化呈现。从本质上说，PPD 用于识别$\hat{g}_\lambda$中显著内部峰值的存在、大小和位置。

PPD 可以生成多种图示。只绘制 λ取值范围内的峰值指标，可得到 PPD 条形图。如果还关注峰值位置，则可以使用 PPD 曲面图。一个需要探讨的重要问题是，什么样的峰值才算 $significant$峰值；我们将在下一小节讨论这一问题。此前，显著峰值的追踪已应用于密度估计（Chaudhuri 和 Marron（1999）），不过是在基于核的方法的带宽选择这一背景下。PPD 类似于拓扑数据分析中的持久同调（Zomorodian 和 Carlsson（2005）），后者追踪不同数据分辨率下拓扑特征的存在情况。相比之下，PPD 追踪标量函数的几何特征（峰值）。

图 4 展示了一个针对 4(a) 所示数据的 PPD 示例。在面板 (b) 中，PPD 条形图显示了对于不同 λ 值所检测到的峰值。 x 轴表示 λ，而 y 轴表示峰值标签。面板 (c) 和 (d) 显示了三维曲面图，其中根据高度的不同，$\hat{g}_\lambda(t)$ 的值以从蓝色到黄色的颜色表示。 y 轴表示 λ， x 轴表示 t，标记的黑线表示峰值的位置

随着 λ 变化。在 $\lambda = 0$ 处, $\hat{g}_\lambda$ 呈现出四个内部峰值，并且这些峰值会随着 λ 增大而变化。值得注意的是，标记为 2 和 4 的峰值在 $\lambda \sim 0.05$ 时消失，而标记为 1 和 3 的峰值则持续存在。

PPD 是研究未知函数在一系列 λ 上的几何或形状特征的有用工具。持续范围较长的峰值被认为是显著的，而较小或持续时间不长的峰值则归因于噪声或对齐伪影。不过，判断哪些显著、哪些不显著可能颇具挑战。下一节将讨论这一问题及其他相关问题。

峰值的显著性与持续性

判断 PPD 中的峰值是真实峰值还是伪影是一个复杂的过程，需要根据最终目标做出一些 $ad\ hoc$ 选择。本文着重关注更大、更全局的结构，并接受舍弃一些较小、更精细特征所带来的损失。

峰值的显著性：什么样的峰值才算显著？在$\hat{g}_\lambda$ 中会出现一些较小的峰值，它们可能由噪声或计算误差造成，并不存在于原始 g 中。例如，在定义域中 g 为常数的某些区域，噪声的对齐可能会产生伪峰。对于位于 t_0 的峰值，我们将该峰值的 $strength$ 定义为 $\bar{}_{\hat{g}''_\lambda(t_0)}\|\hat{g}''_\lambda\|$ 。这个量衡量点 t_0 处的曲率，并经过适当归一化。如果该量小于预先设定的阈值（例如 τ），那么位于 t_0 的峰值就被视为 $\overline{insignificant}$，并予以舍弃。否则，该峰值被认为是显著的，并保留在 PPD 中。本文采用了一个较为保守的阈值 $\tau = 0.03$；该阈值经大量实验确定，在各种数据集上均表现良好。

峰值的持久性:
对于标记为 k 的峰值，我们将其 $persistence$ 定义为$\hat{g}_\lambda$ 中的 $p_k = |\{\lambda \geq 0 \mid$ 峰值 k 是否显著$\}|$, 其中$|\cdot|$ 表示区间的长度。接下来的问题是确定峰值达到何种最小值才算 $persistent$。为此，我们选择持久性最高的峰值，记为 k_0，并定义其他峰值相对于

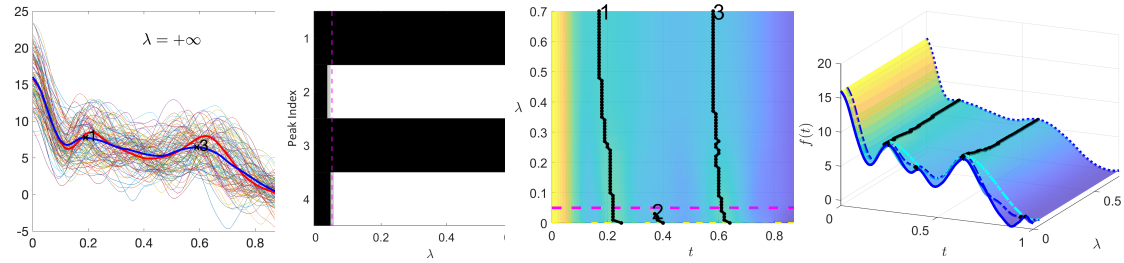


图 4. (a) 原始函数型数据，(b) PPD 峰值持久性图柱状图，以及 (c)、(d) 通过对不同 λ 下的$\hat{g}_\lambda$ 进行插值而生成的 PPD 曲面图，以及峰值的移动情况。

p_{k_0}的相对持久性。如果 k^{th} 峰值 p_k/p_{k_0} 的相对持久性大于某个阈值（例如 θ），则认为该峰值 $persistent$。通过实验，我们发现 $\theta \approx 0.28$ 在模拟数据和真实数据上表现最佳。内部显著峰值的数量为：$m = |\{k \geq 0 \mid p_k/p_{k_0} > \theta\}|$。该数量 m 决定了 g 的估计形状，并构成估计 g 时的一项约束。

总之，确定显著峰值和持久峰需要两个超参数 τ 和 θ。结果表明，对于本文所采用的参数选择，结果相对稳定。

λ 的选择与 g 的初始估计

选出显著且持久的内部峰值后，下一步是确定 λ的最优值，例如 λ^*，以便后续处理。我们选择能使$\hat{g}_\lambda$ 中出现 m 个显著峰值的最小 λ。这一选择基于如下观察：随着 λ 增大，$\hat{g}_\lambda$ 的极值点通常会减弱。对于图 4 所示的示例，我们选择 $m = 2$ 和 $\lambda^* = 0.05$。对于这个λ^*，我们还计算横截面均值$\hat{g}_{\lambda^*} = \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{i,\lambda^*}$。

–

形状估计量性质

研究估计量 m 在 g中内部峰值或众数数量的统计性质将会很有意义。过去已有多篇论文研究了密度估计背景下的众数估计问题，尤其是在使用非参数核估计量时，参见 $e.g.$、Minnotte（1997）；Chaudhuri 和 Marron（1999）以及其中的参考文献。我们的估计量基于部分对齐函数横截面均值$\hat{g}_{\lambda^*}$ 的几何性质。鉴于黎曼弹性度量具有非欧几里得性质，且峰值持久性的几何抽象性，我们尚未对此进行任何理论研究。

估计器。相反，我们使用模拟数据集和真实数据集，对这一方法进行了广泛的实验验证。

实验结果：形状估计

现在，我们展示一些基于PPD的形状估计示例，数据来自模拟和真实的函数型数据集。

示例 1：第一个模拟实验从一个双峰函数生成数据，该函数标记为 g_1 （红色曲线）；随后，我们通过添加几条三峰曲线g_2 （洋红色曲线）对数据进行污染。完整数据集包含针对 g_1 的 80 个随机扰动，以及针对 g_2的 20 个随机扰动。我们的目标是研究如何从这些含噪且受污染的数据中估计 g_1 的形状。

图 5 展示了估计结果。图 (a) 显示原始数据，图 (b) 显示完整弹性对齐的结果，均值曲线$\hat{g}_0$以蓝色绘出。这条均值曲线有三个峰值，其中第三个相对较小。图 (c) 显示了部分弹性对齐的结果，其中 $\lambda^* = 0.02$ 是通过 PPD 选出的；相应的 PPD 显示在图 (d)、(e) 和 (f) 中。图 (d) 中的条形图显示了两个显著且持久的峰值（1 和 2），与 g_1一致。这个例子表明，即使数据受到另一种形状的污染，PPD方法仍能成功估计底层信号的形状。

示例 2：在本例中，与第一次模拟相比，我们调换了 g_1 和 g_2 的角色。具体而言，我们有 80 个三峰函数样本和 20 个双峰函数样本。图 6 中的第四幅图 (d) 表明，三个峰值（1、2 和 3）均显著且持久，并且 $\lambda^* = 0.01$ 是最优对齐。这些发现表明，PPD方法具有鲁棒性，能够准确识别含噪信号中的峰值数量。

示例 3：加利福尼亚州空气质量：加利福尼亚州经常遭受野火侵袭，野火会显著

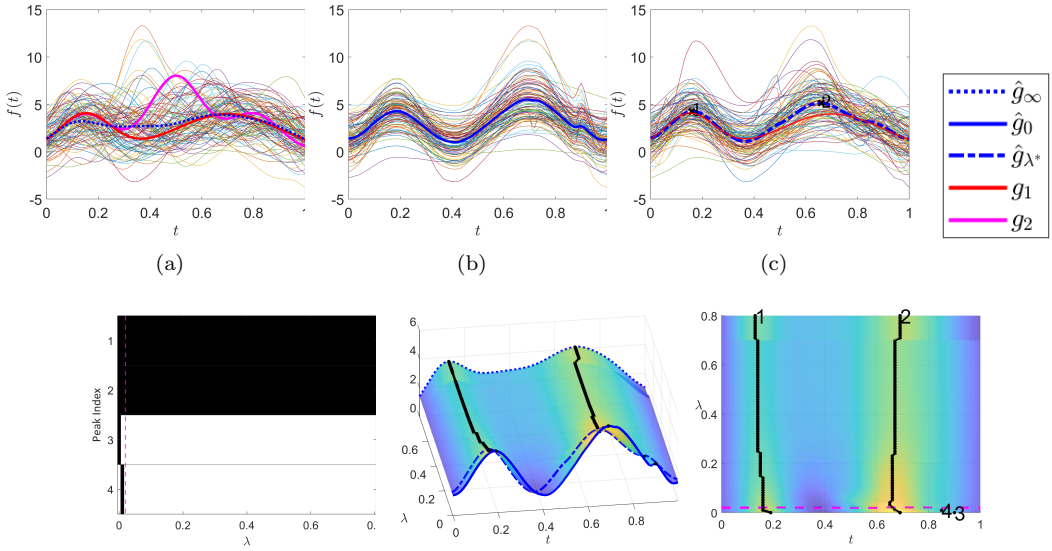


图 5. 示例 1：数据由分别根据 g_1 和 g_2 生成的 80 个和 20 个函数组成。图 (a)、(b) 和 (c) 分别显示$\hat{g}_\infty$、$\hat{g}_0$ 和$\hat{g}_{\lambda^*}$。从 (d) 到 (f) 的 PPDsin 表明，两个峰值（1 和 2）显著且持久。

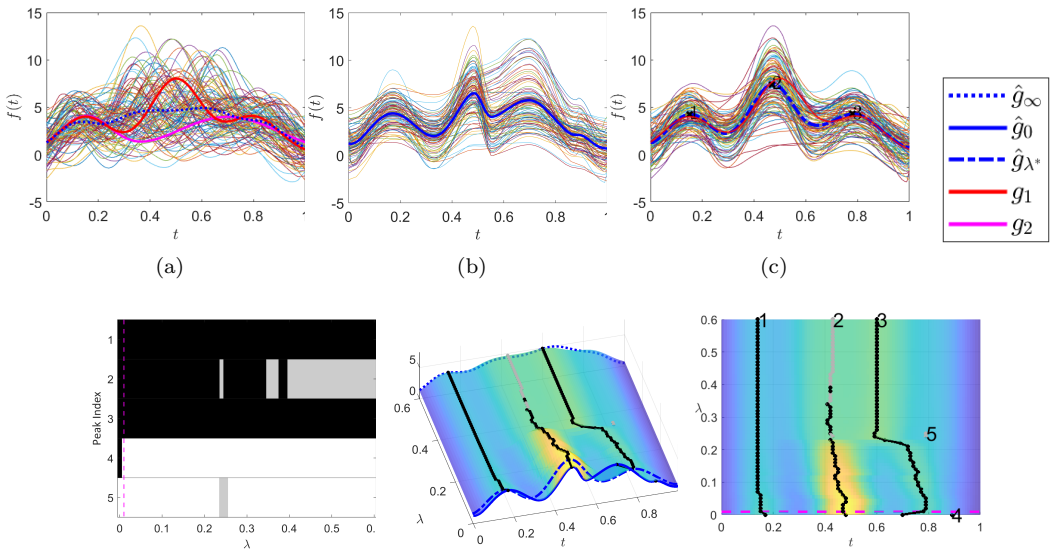


图 6. 示例 2：此图的布局与图 5 类似。在本例中，80 个函数有三个峰值，其余 20 个有两个峰值。从 (d) 到 (f) 的 PPD 表明，三个峰值（1、2 和 3）显著且持久。

影响其空气质量。为研究这种影响，我们分析了环境保护局（EPA）公开发布的数据，数据记录了2018年加利福尼亚州58个县中42个县的每日细颗粒物（PM 2.5）浓度。图 7（a）展示了

通过平滑原始测量值获得的时间索引数据。我们采用 *Lowess*（局部加权散点图平滑）方法进行平滑，窗口大小为 50 天。(d)、(e) 和 (f) 中的 PPD 显示，估计数量

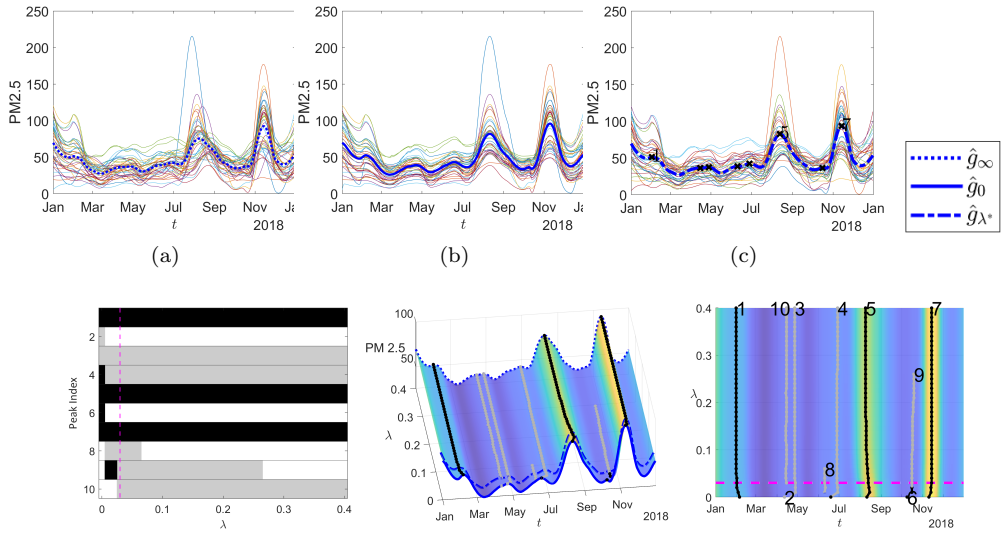


图 7. 示例 3：此图的布局与图 5 类似。(d) 中的条形图表明，尽管$\hat{g}_0$ 有七个峰值，但只有三个峰值（1、5 和 7）显著且持久。

真实值信号中的峰值数量 g 为三个（1、5 和 7），其中 $\lambda^* = 0.03$。除了 12 月至 1 月（边界）外，内部峰值时段还出现在 2 月、8 月下旬和 11 月下旬。2 月的峰值较小但显著，而另外两个内部峰值则明显占主导地位。

示例 4：欧洲国家出生率变化：我们分析了 Roser 等人（2013）收集的 1950 年至 2021 年间 50 个欧洲国家出生率历史变化的数据集。此处的响应变量为出生率的年度变化，其中出生率表示每年每 1,000 人中的活产婴儿数。首先使用 15 年的窗口大小对原始数据进行平滑处理。PPD方法发现了四个显著且持久的峰值。该峰值的最优参数为 $\lambda^* = 0.09$。图 (b) 表明，$\hat{g}_0$ 呈现五个峰值，但随着 λ 的增大，第五个峰值很快消失。另一方面，(a) 和 (d) 显示$\hat{g}_\infty$ 有四个峰值。然而，(c) 显示$\hat{g}_{\lambda^*}$ 有四个显著峰值，即编号 1 至 4 的峰值。**示例 5：OECD 国家消费者信心指数**：本例分析了经济合作与发展组织（OECD）公开发布的 2008 年至 2022 年间 38 个 OECD 国家的月度消费者信心指数（CCI）。CCI 根据家庭对财务状况的预期和经济情绪调查得出。指数高于 100 表示信心增强且消费意愿上升，低于 100 则表示悲观情绪和储蓄倾向并增加支出。作为预处理，我们采用窗口大小为 24 个月的 Lowess 平滑方法。

如图 9 所示，使用 PPD 方法分析该数据集发现，(b) 显示$\hat{g}_0$ 有四个峰值。然而，(d) 至 (f) 中的 PPD 图表明，$\lambda^* = 0.04$ 有三个峰值（1、2 和 3），且在 (c) 中的$\hat{g}_{\lambda^*}$ 中被识别为显著峰值。

步骤 2：峰值约束曲线估计

下一步是利用预先定义的形状类别，并采用惩罚最大似然准则来估计 g 。此处的目标是将探索范围限制在所需的形状类别内，并通过几何搜索确定该类别中的最优元素。这项技术改编自 Dasgupta 等人（2021）提出的形状约束密度估计方法，并针对函数估计问题进行了调整。该方法通过调整初始估计$\hat{g}_{init}$ ，同时最大化惩罚对数似然函数来实现。（我们利用$\hat{g}_{\lambda^*}$中的峰值和谷值（包括位置和高度），以及它们之间的平滑插值，构建初始估计$\hat{g}_{init}$。）优化过程覆盖具有指定形状的函数的完整范围。

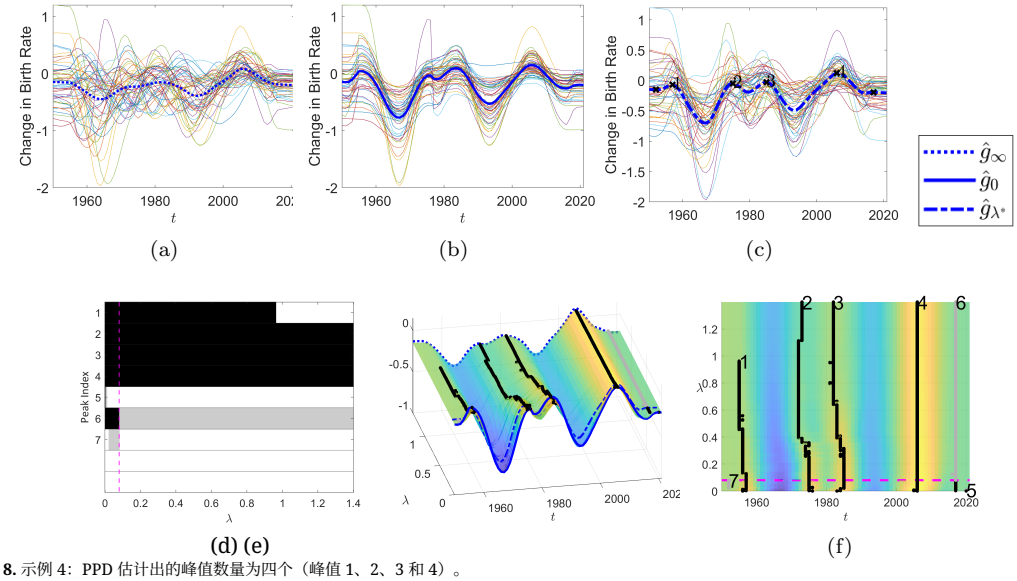


图 8. 示例 4：PPD 估计出的峰值数量为四个（峰值 1、2、3 和 4）。

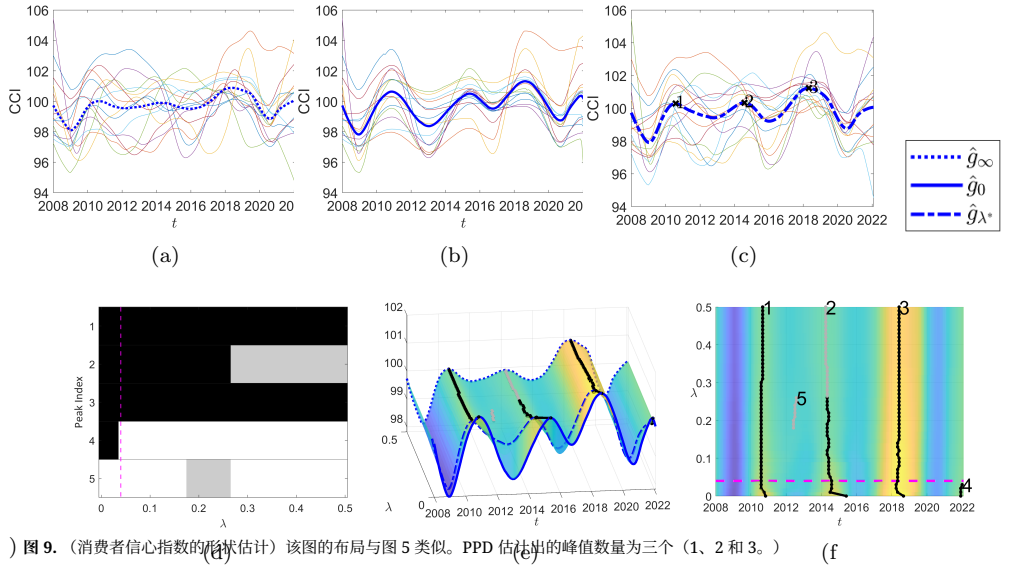


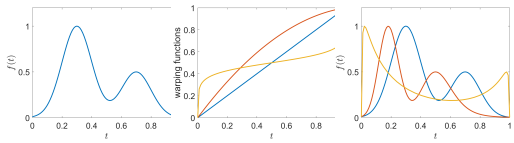
图 9. （消费者信心指数的形状估计）该图的布局与图 5 类似。PPD 估计出的峰值数量为三个（1、2 和 3。）

惩罚最大似然估计方法

设 M 为$\hat{g}_{init}$中 $extrema\ points$ 的数量，其中包括内部峰值、谷值和两个端点。 M 的取值范围为 $2m+1$ 至 $2m+3$ （含端点），其中 m 为估计的内部峰值数量。例如，如果两个端点都是极小值，并且有 $m-1$ 个内部谷值，则 M 等于 $2m+$ 1.

形状约束函数集：设函数 f 在 I 中有 M 个极值点。我们将 $\mathcal{F}_m \subset \mathcal{F}$ 定义为在 I 上具有 m 个内部峰值的所有函数的集合。 $\mathcal{F}_m$ 中任意两个元素的极值点位置和高度都不同。

相应地，我们将定义两组变量。



(a) (b) (c)

图 10。函数的时间扭曲。（a）为原始函数，（b）展示了三个扭曲函数，（c）为由（b）中相应的 γ's 进行扭曲后的函数。

- **位置**：我们将使用时间扭曲函数—— Γ 的元素——来改变极值点的位置。也就是说，对于一个 $\gamma \in \Gamma$，复合 $\hat{g}_{init} \circ \gamma$ 会在保持极值点的高度和顺序不变的同时，改变极值点的位置。图 10 展示了函数的峰值数量和高度对于时间扭曲是不变的。

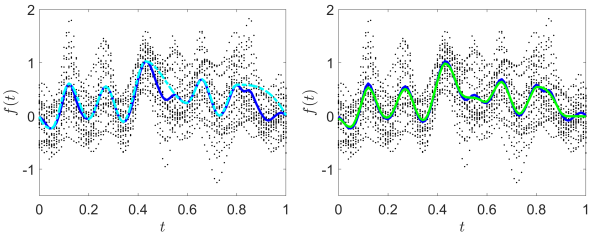
- **高度**：我们还将 $\mathbf{s} = \{s_1, s_2, \ldots, s_M\}$ 定义为 $\hat{g}_{init}$ 中极值点的高度。对 $\mathbf{s}$ 的取值存在一些自然约束。谷的高度应低于其相邻峰值的高度。令 $\mathcal{S}$ 为满足这些约束的所有向量 $\mathbf{s}$ 的集合。我们将使用 $f_{\mathbf{s}} \in \mathcal{F}_m$ 表示一个函数，该函数的极值点高度向量为 $\mathbf{s}$ 。

从初始估计 $\hat{g}_{init}$ 开始，我们调整极值点的位置和高度，以探索集合 $\mathcal{F}_m$。最终估计为：$\hat{g} = (f_{\mathbf{s}^*} \circ \gamma^*)$，其中

$$(\gamma^*, \mathbf{s}^*) = \underset{\gamma \in \Gamma, \mathbf{s} \in \mathcal{S}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \int_0^1 (f_{\mathbf{s}}(\gamma(t)) - \tilde{f}_{\lambda^*,i}(t))^2 dt + \rho \int_0^1 \ddot{f}_{\mathbf{s}}(\gamma(t))^2 dt \right) \quad (6)$$

这里 $\rho > 0$ 是用于偏好平滑函数的无穷小权重（$\sim 10^{-8}$）。该解与公式 4 在几个方面有所不同。这里我们只对一个 γ 进行优化，而公式 4 对每个观测值都使用一个 γ_i 。此外，我们使用偏好 m 个峰值的部分对齐数据，而不是使用原始数据。

使用 MATLAB 中的 `fmincon`函数求解公式 6 中的优化问题。然而，由于 Γ是一个无限维非线性流形，直接优化颇具挑战。为此，我们采用 SRVF 映射，随后使用逆指数映射，在向量空间中表示扭曲函数，并使用正交基通过系数表示 γ 。感兴趣的读者可参阅 Dasgupta等（2017）了解优化细节。算法2概述了对 g进行形状约束估计的步骤。图 11 给出了一个示例。在图 (a) 中，我们以点状标记显示部分对齐函数 $f_{\lambda^*,i}$ 并以青色显示初始估计 $\hat{g}_{init}$ 。图 (b) 以绿色显示算法2的结果。深蓝色线条表示真实值 g。



(a) (b)

图 11. （形状约束函数估计）虚线点表示部分对齐函数 $\{\tilde{f}_{\lambda^*,i}\}$ ，蓝线表示其横截面均值。青色线和绿色线分别是 g 的初始函数估计值和最终函数估计值。

算法2 峰值约束曲线估计

要求： 数据：$\{\tilde{f}_{\lambda^*,i}\}_{i=1}^n$，初始形状估计 $\hat{g}_{init}(t)$

1:初始化的极值点的高度 $\hat{g}_{init}$: $\{s_1, s_2, \ldots, s_M\}$,

2: 初始化 $\gamma \leftarrow \gamma_{id}$。

3: 通过 `fmincon` 求解方程 6，得到$\hat{\gamma}$ 和 $\hat{\mathbf{s}}$ 。

4: **返回** $\hat{f}(t) = f_{\hat{\gamma},\hat{\mathbf{s}}}(t)$

估计置信带的自助法

为评估估计值$\hat{g}$ 的性能，我们计算估计器的逐点置信带，并将其与估计结果一同展示。对于模拟数据，我们检查这些估计的置信带是否包含实际信号，以验证我们的结果。算法 3 列出了计算这些置信带的步骤。

函数估计器的性质

形状约束函数估计的基本方法沿用 Dasgupta 等人（2021）提出的方法，不过群 Γ的作用有所不同，因为那篇论文关注的是估计 pdfs。当前估计量$\hat{g}$ 的渐近性质

算法 3 构造逐点函数置信带

要求： $\{f_i\}_{i=1}^n$，初始形状估计，$\hat{g}_{init}$，估计的平滑参数，λ^*

1: **对** $j = 12, \ldots, B$ **执行**

2: 有放回地重采样 n 个函数，以获得一个函数自助抽样集，$\{f_i^{(j)}\}_{i=1}^n$。

3: 获取部分地对齐的函数 with 平滑参数，λ^*, $\{\tilde{f}_{\lambda^*,i}^{(j)}\}_{i=1}^n$

4: 给定$\hat{g}_{init}$ 和 $\{\tilde{f}_{\lambda^*,i}^{(j)}\}_{i=1}^n$，获得最佳通过算法2得到的 g 估计值: $\hat{g}^{(j)}$

5: **结束于娱乐主管**

6: 对于 $\{\hat{g}^{(j)}\}_{j=1}^B$，**返回** $(\alpha/2)$ 和 $(1-\alpha/2)$ 分位数以设定……的下限和上限的逐点置信界。

与早期研究的结果相似，详细内容请参阅该论文。

实验结果：函数估计

本节通过一系列模拟数据集和真实数据集评估所提出方法的性能。我们的目标是将估计值$\hat{g}$ 与两种先前的方法进行比较，其中包括：(1) 未对齐函数的均值，$\bar{f} = \hat{g}_\infty$；(2) 完全弹性均值，$\hat{g}_0$。此外，对于模拟数据，我们还研究了使用公式 4 得到的结果。然而，由于最优 κ 值未知（如第 2 节所述），我们展示了若干个 κ 值对应的结果。我们使用均方根误差（RMSE）作为比较估计误差的指标，并提供 95% 逐点自助法置信带，以展示估计量的变异性。

模拟研究

我们在定义域 $I = [0, 1]$ 上尝试了四种不同的模拟情景。在每项实验中，我们选择不同的 g，并根据公式 3 生成 $n = 100$ 个样本。为计算并比较不同的结果，我们针对每种情景进行 50 次独立模拟，并计算估计量的均方根误差统计量。

- 模拟数据集***1***

图12展示了第一个数据集的实验结果。原始数据如面板 (a) 所示，真实函数 g 以红色绘出。图 (a) 和 (b) 分别展示了在 λ 的极端取值 ∞ 和 0 下函数的对齐结果。每种情况下，横截面均值均以蓝色显示。该数据的 PPD 条形图见 (d)，它估计出两个显著内部峰值。最优参数 $\lambda^* = 0.05$ 以 (d) 中洋红色的竖直点线表示。图 (e) 和 (f) 为 PPD 曲面图展示了随着 λ 增大，$\hat{g}_\lambda$ 的渐进变化。曲面上的黑线是显著峰值的轨迹。图 (f) 展示了曲面的俯视图。对于这个 λ^*，(c) 展示了部分对齐函数，$\{\tilde{f}_{\lambda^*,i}\}_{i=1}^n$（以点云形式显示，以便看清）。其均值$\hat{g}_{\lambda^*}$ 以蓝色虚线曲线表示，相比$\hat{g}_0$ 和 $\hat{g}_\infty$，它能更好地捕捉 g 的几何特征。青色虚线表示初始估计$\hat{g}_{init}$。

接下来，我们使用这些量来估计 g，具体方法如第 5 节所述。(g) 中的绿色虚线曲线是最终估计$\hat{g}$。我们还使用灰色区域展示了逐点自助法置信带。可以看出，在较大的区间内，函数估计（绿色）和置信带（灰色）均可与$\hat{g}_0$ 和 $\hat{g}_\infty$ 区分开来。此外，从 $\mathbb{L}^2$ 距离的角度来看，$\hat{g}$ 是最接近真实值 g 的函数。稍后将对结果进行详细的定量评估。

- 模拟数据集***2***

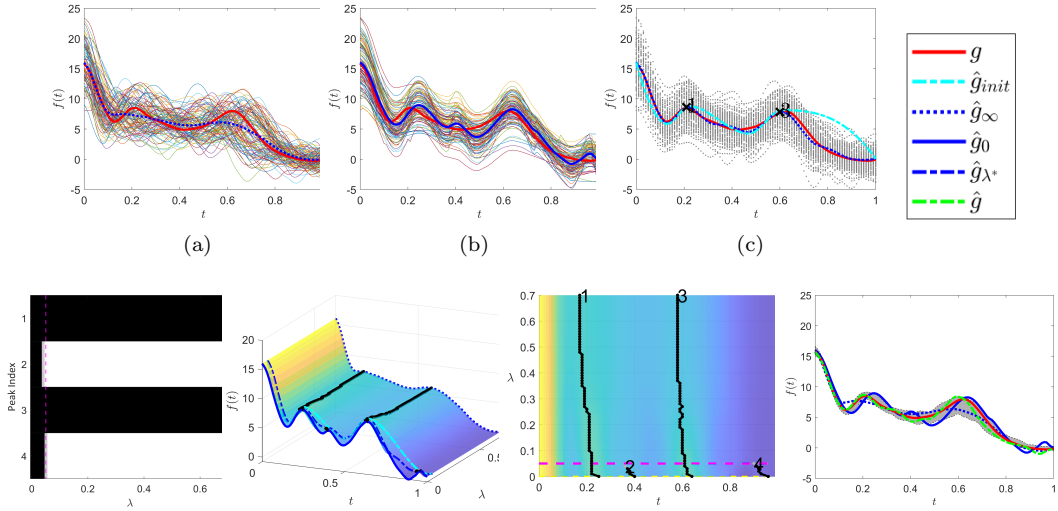
本实验对g使用锯齿波函数，数据如图13 (a)所示。该图的其余部分与图12的布局类似。图 (d) 中的 PPD 条形图成功检测出两个显著的内部峰值，分别标记为 3 和 6，并估计了 $\lambda^* = 0.04$。峰值的演变过程通过 (e) 和 (f) 中的 PPD 曲面图展示。(c) 图展示了~部分对齐函数、$\{f_{\lambda^*,i}\}_{i=1}^n$、它们的均值$\hat{g}_{\lambda^*}$，以及在 $\lambda^* = 0.04$ 处的初始估计$\hat{g}_{init}$。(g) 图展示了最终估计$\hat{g}$（绿色）及其灰色的自助法置信带。如图所示，弹性均值$\hat{g}_0$（蓝色实线）存在伪峰，而$\hat{g}_\infty$（蓝色虚线）低估了峰值和谷值的高度。我们的估计$\hat{g}$（绿色）对g给出了出色的估计。

- 模拟数据集 ***3***

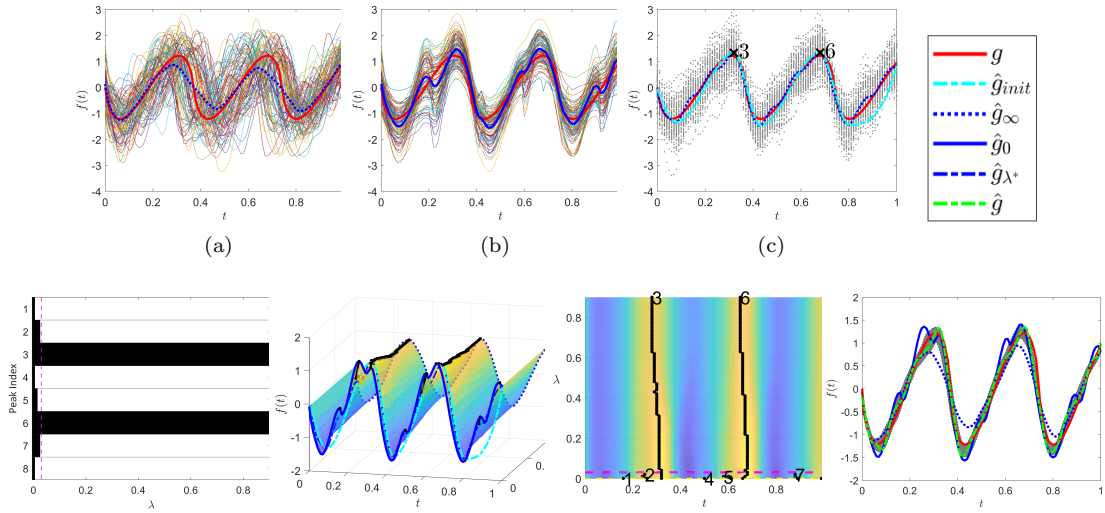
在这种情况下，我们选择一个具有两个峰值且峰值之间存在较大常数区域的函数 g。由于噪声可能在平坦区域内错误地引入伪峰，因此该常数区间难以估计，如图 14 的图 (b) 所示。这组数据包含一些持续时间较短的峰值，以及其他未通过显著性检验的峰值。PPD 条形图检测出峰值 2 和峰值 5 具有显著性和持久性，由此得到$\lambda^* = 0.03$。(e) 和 (f) 中的 PPD 曲面图展示了这些峰值的演变过程。图 (g) 展示了最终估计（绿色）及其逐点置信带（灰色），该结果与$\hat{g}_0$（蓝色实线）和$\hat{g}_\infty$（蓝色虚线）存在显著差异。

- 模拟数据集***4***

在这个例子中，我们使用一个斜率适中、相位变化较大的宽广单峰函数，来处理



(d) (e) (f) (g) **图 12**。（数据集1）图 (a) 和 (b) 分别绘制了 $\lambda = \infty$ 和 0 处的对齐函数。这些图还显示了$\hat{g}_0$ （蓝色粗线）、$\hat{g}_{0.05}$ （蓝色粗虚线）和$\hat{g}_{\infty}$ （蓝色点线）。真实值 g、初始估计$\hat{g}_{init}$和最终估计$\hat{g}$ 分别以红色、青色和绿色绘制。图 (d) 为 PPD，其中 $\lambda^* = 0.05$。图 (e) 和 (f) 展示了$\hat{g}_{\lambda}$ 随 λ 的变化。曲面上的黑线表示内部峰值，(f) 是 (e) 的俯视图。图 (c) 显示了 λ^*处的对齐函数。图 (g) 比较了未对齐均值、完全弹性均值、我们的估计值和真实值， g。95% 自助法置信带以灰色区域表示。



(d) (e) (f) (g) **图13**。(模拟数据集2)此图的布局与图 12 类似。我们的估计$\hat{g}$ 最接近 g ，且 95% 置信带完全包含 g 。

数据。图15 (a) 显示原始数据。 (d) 中的峰值持久性图柱状图成功筛除了$\lambda^* = 0.11$处的伪峰。图 (g) 显示最终估计值$\hat{g}$。我们

我们发现，$\hat{g}$ 能够很好地估计 g峰值的位置和高度。

- 估计性能的量化

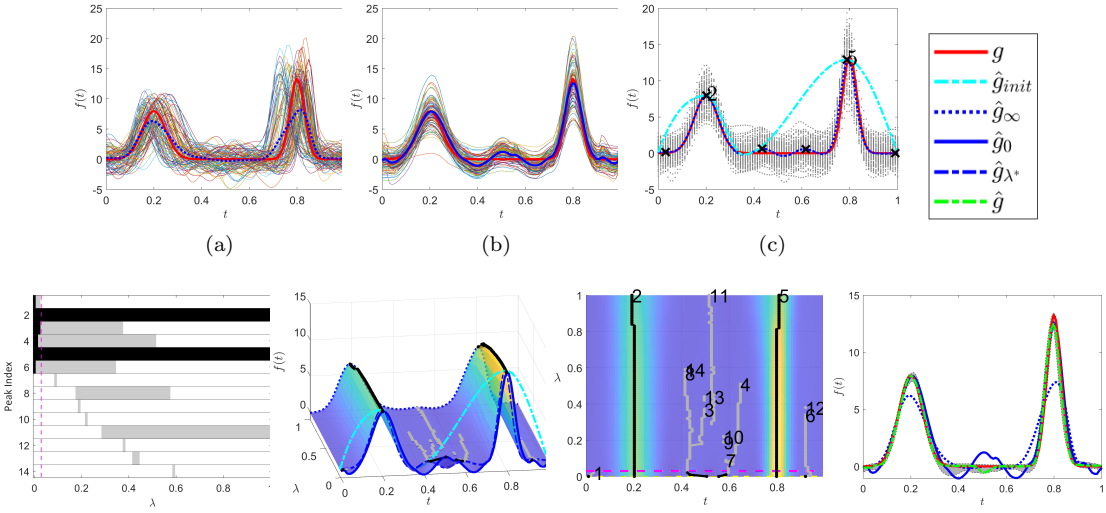


图 14.(模拟 数据集 3) (d) 中的峰值持久性图柱状图显示有两个峰值，编号为 2 和 5。因此，虚线青色线条位于 (c) 和 (e) 中，表示 g 的估计形状。最终估计$\hat{g}$ 显示在 (g) 中。

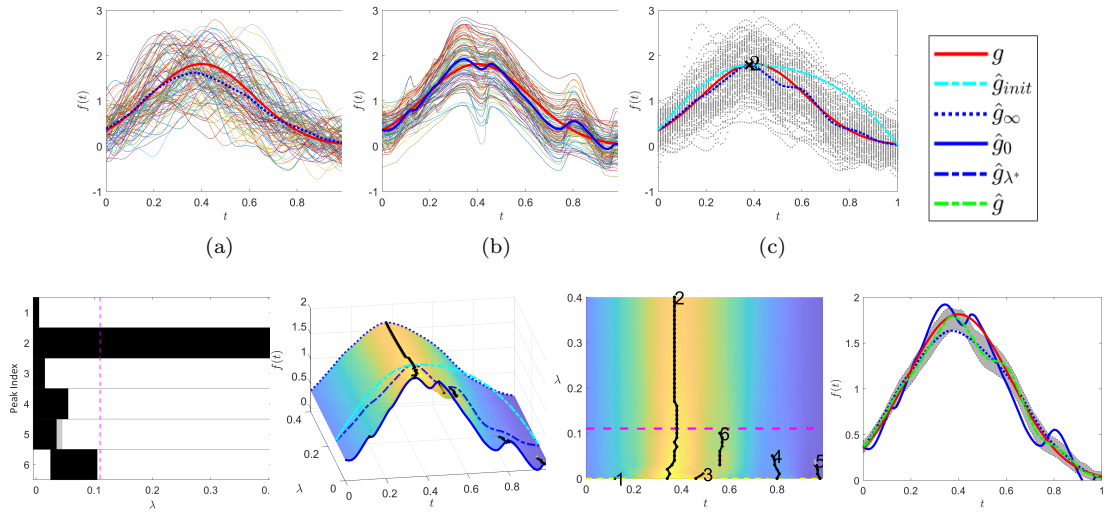


图15。 （模拟数据集4）：详细信息与图 1 类似

2.

到目前为止，我们通过图示分析了不同方法得出的估计结果。为了量化这些估计的准确性，我们对每项实验进行了 50 次重复，并计算了估计误差。图16展示了四项模拟实验中 $\mathbb{L}^2$ 误差均值和标准差的箱线图。结果表明，$\hat{g}$（箱线图中的第三个箱体）的均方根误差最低。相比之下，未对齐均值（$\hat{g}_{\infty}$）在估计方面表现不佳

峰值和谷值的高度，以及完全对齐的均值（$\hat{g}_0$）会高估峰值的数量。为强调形状在估计中的作用，图17展示了不同模拟数据集所估计的峰值数量直方图。上排直方图显示了部分对齐均值$\hat{g}_{\lambda^*}$的横截面均值中的峰值数量，其中既包括显著峰值，也包括不显著峰值。下排则显示了数量直方图

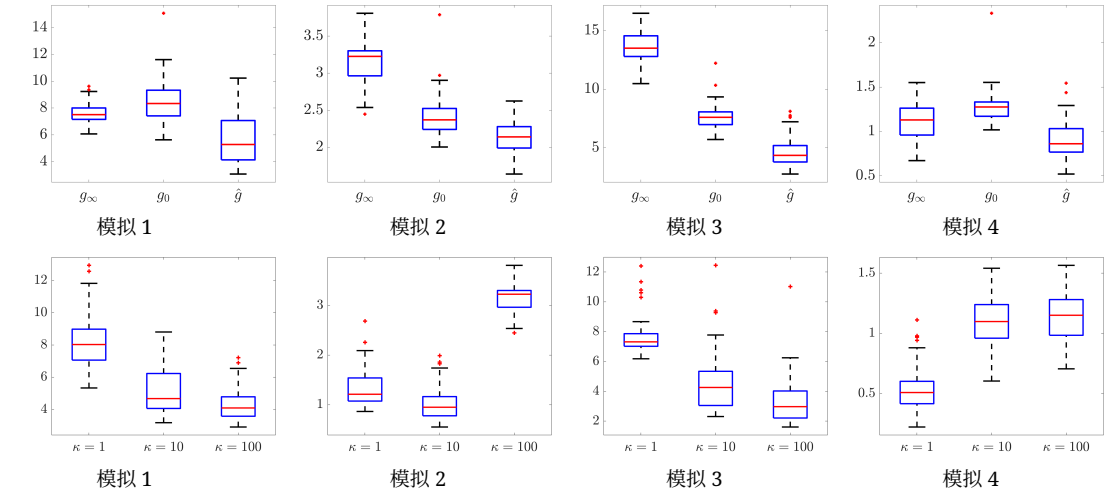


图16。上：惩罚后的 四种模拟场景下三种不同估计量的 $\mathbb{L}^2$ 误差箱线图。下：$\mathbb{L}^2$ 不同 κ 值对应的估计量。

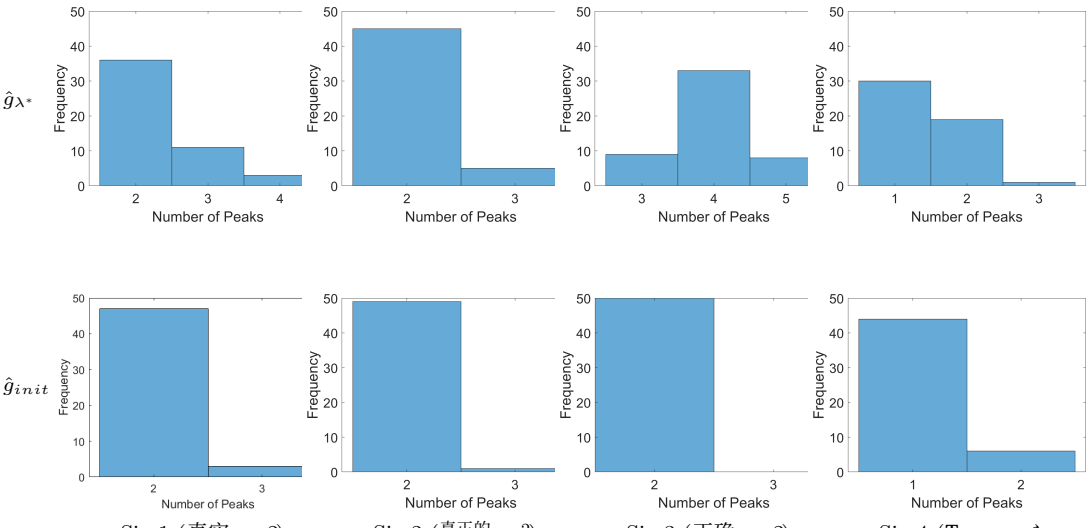


图17。上：H s 的数量 s 估计的峰值数量（不显著和显著）直方图，基于 50 次重复实验。底部：直方图使用 PPD 估计的显著峰值。

PPD 选出的显著峰值。结果表明，PPD 方法在大多数情况下都能正确估计出 g 中峰值的数量。这有助于提升所提出的形状约束函数估计器的性能。

最后，我们分析了与以往方法相比，我们的估计流程的计算成本。表 1 列出了执行不同算法步骤所需的时间。该成本是在 CPU（Intel(R)）i7-11800H @ 2.30GHz 上计算得出的。我们省略了未对齐的成本。

均值，因为它不需要任何函数对齐。计算完全弹性均值的成本（在 50 次重复实验中取平均）列于第一列。最后，我们方法不同步骤的成本——部分对齐、计算 PPD 以及函数估计的成本——列于右侧三列中。大部分时间都消耗在针对不同 λ 值对函数进行部分对齐上。

表 1。我们流程中各个步骤所需的计算时间（秒）。

	完全弹性	部分对齐		
	函数对齐	函数对齐（第 1 步）	计算 PPD（第 1 步）	函数估计（第 2 步）
模拟 1	15.29	1868.4	0.13	17.87
模拟 2	17.12	2021.6	0.12	18.92
模拟 3	14.59	2205.0	0.14	38.16
模拟 4	15.83	1564.8	0.09	17.01

真实数据研究

在本节中，我们将应用我们的方法来估计一些真实数据集背后的函数 g 。然而，在这些情况下，我们无法将我们的估计结果与真实值函数进行比较以评估其准确性。因此，必须将结果与其他估计量进行解读和比较，以评估它们的性能。

伯克利生长率数据:
首先，我们研究经典的伯克利生长研究数据（Ramsay 和 Silverman（1997））， 这组数据常用于展示函数型数据中的相位变异。女性数据集包含 54 名女孩 1 至 18 岁期间的身高生长速率。生长曲线中的峰值称为 $growth\ spurts$ ，并且备受关注。科学家和医生使用生长图表分析个体的生长情况，并提出这样的问题：男性或女性平均会经历几次生长突增？该类别人群的平均生长剖面是什么样的？图 18（a）展示女性受试者的原始数据，（b）展示完全对齐的函数。未对齐均值（（a）中的蓝色曲线）几乎抹平了所有生长突增，因此可能无法很好地代表总体。完全对齐均值（（b）中的蓝色曲线）显示了几次强烈的生长突增，也有一些相对较弱的生长突增。在生成的 13 个峰值中，PPD 图表选出四个内部峰值（2、3、5 和 8），认为它们显著且持久，由此得出结论：女性生长有三次主要的生长突增。（g）图中的曲线展示了估计函数$\hat{g}$ ，颜色为绿色。

欧洲 COVID-19 数据分析
人们对分析新冠疫情数据及相关感染率激增现象抱有浓厚兴趣。尽管不同社群经历新冠峰值（或波次）的时间并不同步，但新冠发病情况和结局背后存在共同规律。例如，如果我们关注该地区不同国家的新冠结局，可以说，疫情期间总体呈现出波次规律。

•**每日住院率曲线**：我们考虑了2020年4月至2021年7月期间25个欧洲国家与 COVID-19 相关的每日医院床位占用人数（Mathieu 等人（2020））。图 19 (a) 和 (b) 分别展示了感染率的函数型数据及其完全对齐函数，以及它们的均值$\hat{g}_\infty$和$\hat{g}_0$。如图所示，$\hat{g}_\infty$ 和$\hat{g}_0$ 在峰值的形状和高度方面存在很大差异。(e) 和 (f) 中的 PPD 曲面图展示了$\hat{g}_\lambda$ 相对于 $\lambda \in [0, 0.2]$的渐进变化，而 (d) 中的条形图则选出了具有三个持久峰（1、4 和 6）的 $\lambda^* = 0.025$ 。面板 (g) 中的最终估计（绿色）突出了这一估计形状。(g) 中的 95% 逐点自助法置信带（灰色）范围较宽，这可能归因于较小的样本量$n = 25$。

- **每日感染率曲线**：COVID-19 的每日感染率曲线可用于研究某种变异株的传播速度或疫苗的有效性。我们从 OWID（Mathieu 等人（2020））收集了 2020年4月至2022年3月期间 25 个欧洲国家的感染率曲线。图 20 (a) 和 (b) 分别展示了函数型数据和完全对齐函数。尽管 (b) 中的函数已很好地对齐，但 I 前半部分的一些峰值似乎并不明显。事实上，PPD方法将 2021年7月 之前的所有峰值都判定为不显著，仅保留了之后的三个峰值。图 (g) 展示了我们的最终估计（绿色），其中前半部分呈现出一段较长的平坦区间。这表明，在 2021年秋季德尔塔变异株暴发期间出现了显著波次。它还检测到 2022年1月 左右的一个显著峰值，而这个峰值并未出现在$\hat{g}_\infty$中。

- **每日死亡率曲线**：我们还研究了2020年4月至2022年3月这25个欧洲国家的新冠肺炎每日死亡率。图21（a）和（b）中的标准估计值$\hat{g}_\infty$ （蓝色虚线）和$\hat{g}_0$ （蓝色实线）差异显著：$\hat{g}_0$ 有几个显著的峰值，而$\hat{g}_\infty$ 则有

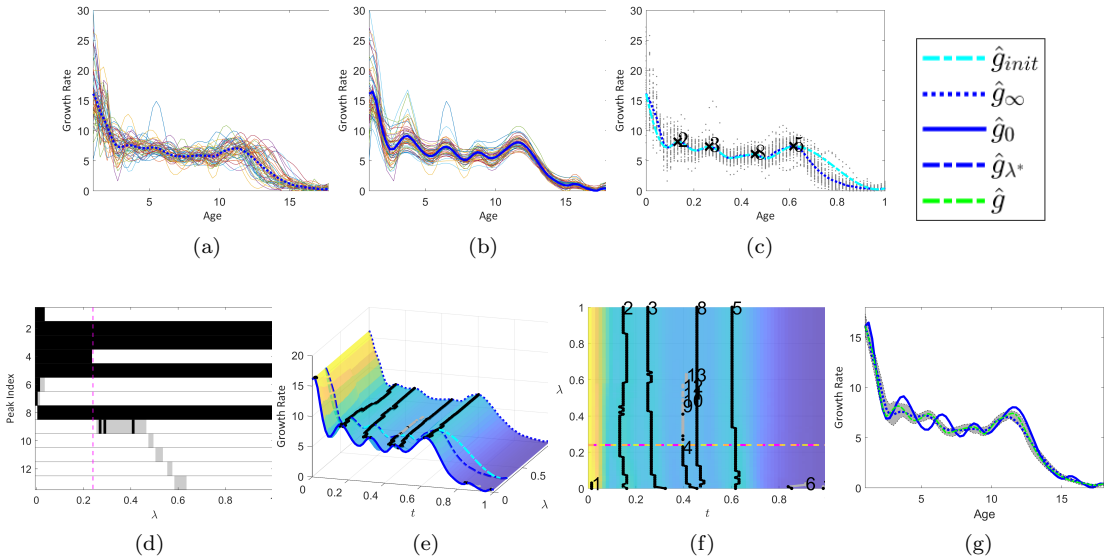


图 18.（女性增长率数据）(a) 为未对齐的原始数据，(b) 为完全弹性对齐结果。(d) 中的 PPD 表明五个峰值（2、3、5 和 8）具有持久性。(c) 展示了 g 的部分对齐点云和初始形状估计（青色曲线）。图 (e) 和 (f) 展示了随着 λ 增大，对齐函数及其峰值过程行为的逐步变化。(g) 展示了底层函数的估计结果。

更少。具体而言，$\hat{g}_0$ 中 2021年1月附近的峰值似乎是人为产生的，因为没有任何国家的原始数据中出现峰值，也没有理由将该峰值对齐到这一特定时间。(d)、(e) 和 (f) 中的 PPD 估计 $\lambda^* = 0.025$，随着 λ 的增加，标记为 4 的峰值消失。如图 (g) 所示，最终估计$\hat{g}$（绿色）与$\hat{g}_\infty$ 和$\hat{g}_0$均不相同。尽管函数的部分对齐与原始数据相差不大，但我们的估计$\hat{g}$ 在 2021年1月之前的新冠肺炎早期阶段识别出了三个明显不同的峰值。

家庭电力消费数据

下一项实验旨在分析佛罗里达州塔拉哈西特定社区家庭住户的半小时电力消费数据 Dasgupta等人（2019）。本研究聚焦于数据的两个子集，分别对应1月和7月，包含968个和1，219个函数。图22（a）和（b）展示了1月数据的原始函数和对齐后的函数。PPD峰值持久性图柱状图显示，五个峰值（5、7、10、11和12）显著，且 $\lambda^* = 0.245$。面板（g）显示，不同估计量之间存在巨大差异: $\hat{g}$ （绿色）、$\hat{g}_0$ （实线蓝色）和$\hat{g}_\infty$（蓝色虚线）。图23展示了7月子集的结果。PPD峰值持久性图柱状图显示有两个突出的峰值（2和5），且 $\lambda^* = 0.18$；而 g_0 和 g_∞分别显示三个和一个峰值。

讨论

本文呈现的实验结果表明，我们的方法能够成功地：（1）估计函数型数据中的峰值数量；（2）以形状约束的方式估计底层未知函数。一项重要成果是利用PPD自动选择调优参数 λ 。使用PPD不仅直观，而且能有效确定 g中峰值的数量和位置，从而为平滑参数 λ选择一个合理的值。此外，形状约束函数估计通过消除PPD判定为不显著的峰值，改进了估计值$\hat{g}$。

这里有一个相关问题：为什么不在原始函数空间中定义 PPD 的概念，而要使用 SRVF 表示？ 图 24 通过一个示例说明了这一问题。该示例试图针对之前研究过的模拟数据之一构造 PPD，并得出 κ =0 是最优的结论。当然，这是不正确的，因为在 $\kappa = 0$ 处可以清楚地看到捏缩效应。在 SRVF 表示下，即使对于 $\lambda = 0$，捏缩效应也完全避免了。

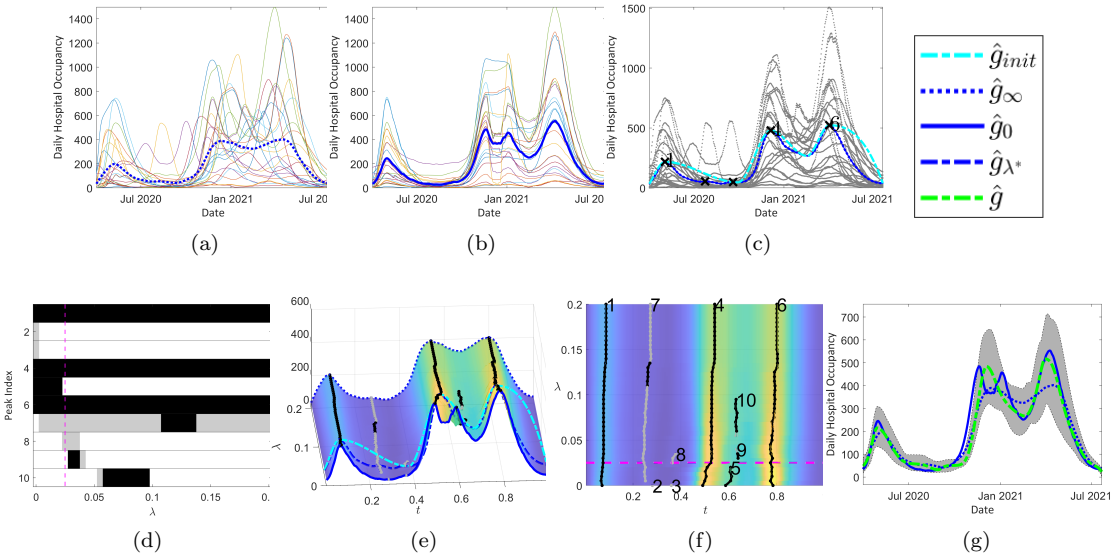


图 19。（欧洲每日医院占用率）(d) 中的 PPD 表明，三个峰值（1、4 和 6）是真实值的真实特征。在图 (b) 中存在四个峰值，其中峰值 5 是$\hat{g}_\infty$ 和$\hat{g}_0$之间的显著差异。结合 (d) 中的 $\lambda^* = 0.025$ ，图 (g) 中的最终估计（绿色）与其他估计$\hat{g}_\infty$ 和$\hat{g}_0$存在明显区别。由于样本量较小，95% 自助法置信带似乎较宽。

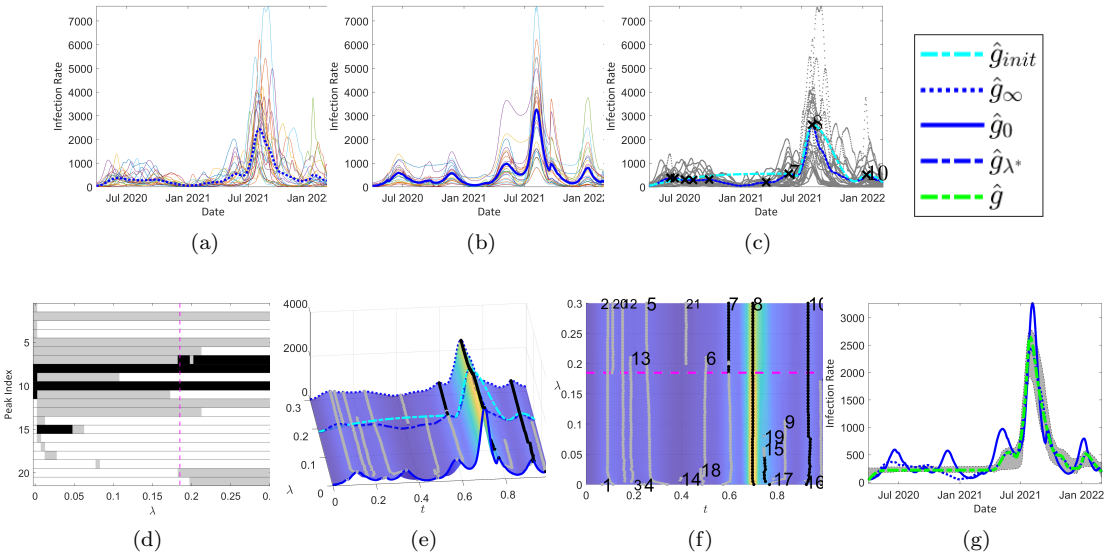


图20。（欧洲每日感染率）(d) 中的 PPD 将三个峰值（7、8 和 10）标记为显著。我们的最终估计（(g) 中的绿色曲线）消除了早期阶段的所有小峰值，并选出了持续时间最长的三个波次。

结论与未来工作

要理解抽样函数型数据的群体行为，就需要估计实际的底层信号 g。在许多情况下，数量，极值点的位置或高度本身也可能直接引起人们的兴趣。在相位噪声和加性噪声的影响下，经典非对齐均值$\hat{g}_\infty$会失去 g 的几何特征，而完全弹性均值$\hat{g}_0$ 会产生伪峰。本文

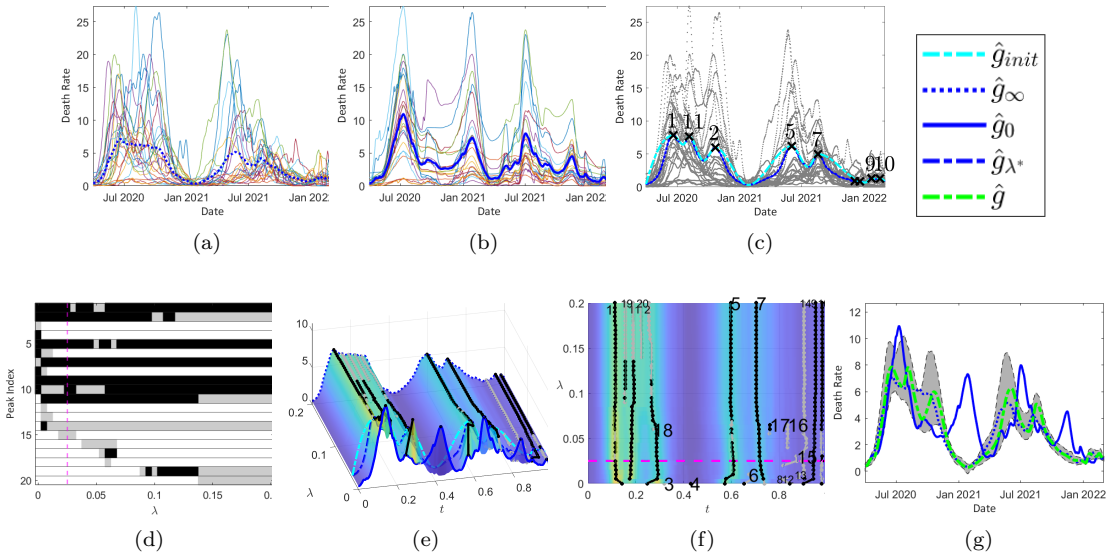


图 21.（欧洲的死亡率）(d) 中的 PPD 表明，七个峰值（1、2、5、7、9、10 和 11）显著。我们的最终估计（(g) 中的绿色曲线）显示有五个主要峰值和两个次要峰值，中间有一个较大的谷。

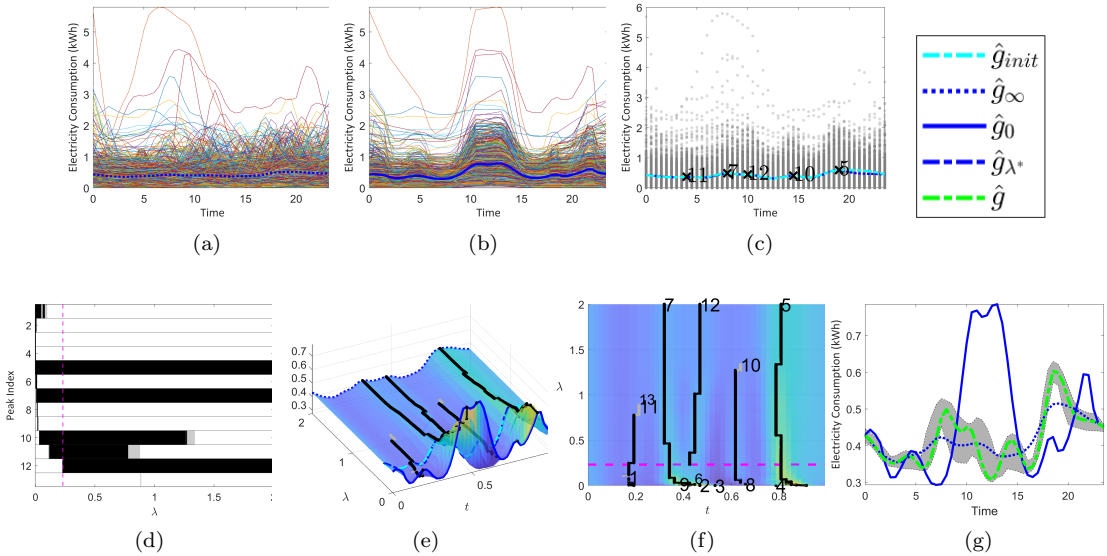


图 22.（2015 年 1 月佛罗里达州塔拉哈西的电力消耗）图 (a) 和 (b) 分别显示家庭数据和完全对齐函数。(d) 中的 PPD 提供了 $\lambda^* = 0.235$，以及五个显著峰值（5、7、10、11、12）。(g) 中的最终估计 $\hat{g}$（绿色）与其他估计 $\hat{g}_\infty$ 和 $\hat{g}_0$ 明显不同。

提出了一种几何方法，用于估计函数 g的形状特征和图像。该方法通过研究$\hat{g}_\lambda$在一系列平滑参数$\lambda \in [0, \infty)$下的几何特征来探索解空间。本文还引入了一种名为峰值持久性图（PPD）的新工具，用于研究这个几何空间以及聚焦于持久峰。峰值的这种持久性有助于我们舍弃不显著峰值，估计 g 的形状并达到最优的 λ^*。后两个量导出了对 g的形状约束估计。这一估计改进了$\hat{g}_{\lambda^*}$ 并产生了一个最优估计，即$\hat{g}$

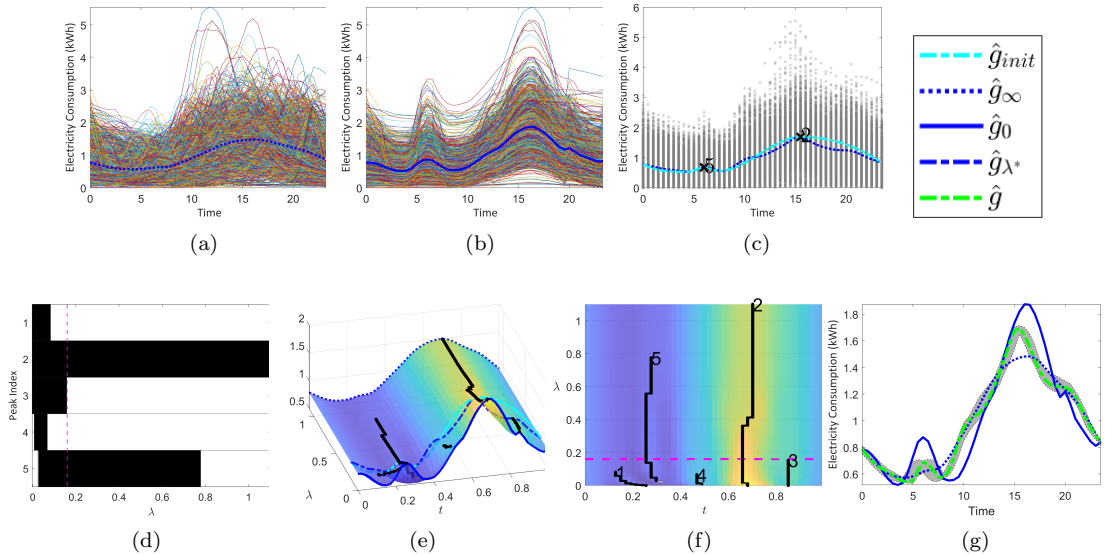


图 23. （塔拉哈西 2015年7月的电力消耗）图 (a) 和 (b) 分别展示了原始数据和完全对齐函数。图 (d) 中的 PPD 提供了 $\lambda^* = 0.18$，并显示了两个显著峰值（2，5）。最终估计$\hat{g}$（绿色）在图 (g) 中与其他估计$\hat{g}_\infty$ 和$\hat{g}_0$ 截然不同。

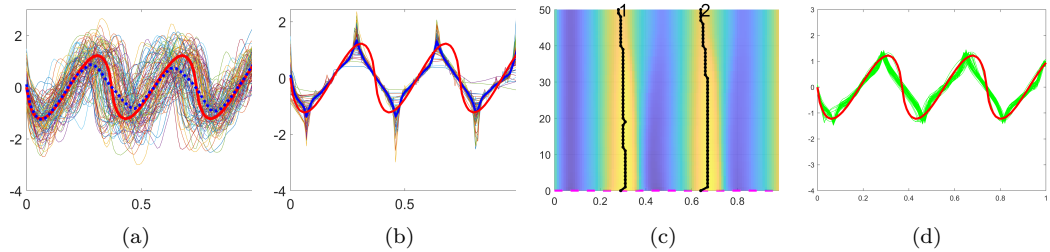


图 24. $\hat{g}_{L2}$ 的 PPD）图 (a) 展示了原始函数、横截面均值（蓝色虚线）和$\hat{g}_{L2}$ （红色）。图 (b) 展示了使用 $\mathbb{L}^2$ 度量对函数进行对齐的结果，并显示其均值（蓝色），其中 $\kappa = 0$。图 (c) 展示了后验概率密度（PPD）图，表明 $\kappa = 0$ 是最优值。图 (d) 展示了 50 次重复模拟的结果，估计值$\hat{g}_{L2}$ 以绿色绘出。

在惩罚最大似然估计下。最重要的是，$\hat{g}$ 中的高度、位置和极值点数量都具有可解释性，并且有数据支持。相比之下，$\hat{g}_\infty$ 低估了峰值高度，而$\hat{g}_0$ 不仅高估了高度，还高估了极值点数量。

作为后续研究，我们希望将真实底层信号的趋势和季节性分离开来，即 $f_i(t) = g_1(t) + (g_2 \circ \gamma_i)(t) + \epsilon_i(t)$，其中 g_1 和 g_2 是带有时间扭曲的趋势函数和季节性函数，$\{\gamma_i\}$，以及加性噪声，$\{\epsilon_i\}$。将数据分解为两个主要信号或许能让我们对其进行更深入的考察。例如，在研究气候变化数据集（如 CO_2排放或全球气温）时，我们可以预期存在一个趋势 g_1，并伴有季节性波动 g_2。通过分离趋势与季节性，我们可以对趋势进行统计检验。另一方面，我们可以拓展“存在单一真实来源”这一假设。例如，在现实场景中，很难确定采样得到的函数 $\{f_i\}$ 是否由单一信号 g 生成。

参考文献

Allassonni´ere, S., Amit, Y. 和 Trouv´e, A.（2007）迈向稠密可变形模板估计的连贯统计框架。 *Journal of the Royal Statistical Society (B)*,**69**, 3–29。Chaudhuri, P. 和 Marron, J. S.（1999）用于探索曲线结构的 Sizer。 *Journal of the*

American Statistical Association,**94**, 807–823。Cheng, M.-Y., Gasser, T. 和 Hall, P.（1999）单峰性和单调性约束下的非参数密度估计。*J. Comput. Graph.Stat.*,**8**, 1–21。Cheng, W., Dryden, I. 和 Huang, X.（2013）函数和曲线的贝叶斯配准。*BayesianAnalysis*, **2015**。Choudhuri, N., Ghosal, S. 和 Roy, A.（2005）函数估计的贝叶斯方法。载于 *Bayesian Thinking*（编者：D. Dey 和 C. Rao）, *Handbook ofStatistics*第25卷，373–414。Elsevier。Claeskens, G., Devijver, E. 和 Gijbels, I.（2021）使用 B 样条对函数型数据进行非线性混合效应建模与时间扭曲。*Electronic Journal ofStatistics*, **15**。Dasgupta, S., Pati, D., Jermyn, I. H. 和 Srivastava, A.（2021）通过可变形模板进行模态约束密度估计。 *Technometrics*,**63**, 536–547。Dasgupta, S., Pati, D. 和 Srivastava, A.（2017）密度建模的两步几何框架。Dasgupta, S., Srivastava, A., Cordova, J. 和 Arghandeh, R.（2019）使用弹性形状分析对家庭用电负荷曲线进行聚类。载于 *2019 IEEE MilanPowerTech*, 1–6。Doss, C. R. 和 Wellner, J. A.（2016）对数凹和 S 凹密度最大似然估计的全局收敛速率。 *Annals of Statistics*,**44**, 954–981。Earls, C. 和 Hooker, G.（2017）函数型数据配准、平滑与预测的变分贝叶斯方法。 *Bayesian Analysis*,**12**, 557 – 582。Fan, J. 和 Gijbels, I.（1996） *Local polynomial modelling and its applications*。佛罗里达州博卡拉顿：Chapman & Hall/CRC。Green, P. 和 Silverman, B.（1994） *Nonparametric regression and generalized linear models: a roughness penalty approach*。英国：Chapman and Hall。Grenander, U.（1956）论死亡率测量理论：第二部分。*Scand. Actuar. J.*Horton, W., Page, G., Reese, C., Lepley, L. 和 White, M.（2020）贝叶斯曲线配准中的模板先验。*Technometrics*,**63**, 1–13。Kneip, A. 和 Ramsay, J. O.（2008）函数模型中配准与拟合的结合。 *Journal of American Statistical Association*,**103**。Kurtek, S.（2017）使用重要性抽样对函数型数据进行成对贝叶斯对齐的几何方法。 *Electronic Journal of Statistics*, **11**, 502 – 531。Kurtek, S., Srivastava, A. 和 Wu, W.（2011）随机时间扭曲与非线性信号对齐下的信号估计。载于 *Advances in Neural InformationProcessing Systems*（编者：J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira 和 K. Weinberger）, 第24卷。Curran Associates, Inc. Liu, X. 和 Muller, H. G.（2004）时间扭曲随机曲线的函数型凸平均与同步。*J. American Statistical Association*, **99**, 687–699。Lu, Y., Herbei, R. 和 Kurtek, S. (2017) 基于高斯过程先验的函数贝叶斯配准。*Journal of Computational and Graphical Statistics*, **26**, 894–904。Marron, J. S., Ramsay, J. O., Sangalli, L. M. 和 Srivastava, A. (2015) 振幅与相位变化的函数型数据分析。*Statistical Science*, 468–484。Mathieu, E., Ritchie, H., Rod´es-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E. 和 Roser, M. (2020) 冠状病毒大流行（COVID-19）。*Our World in Data*。Matuk, J., Bharath, K., Chkrebtii, O. 和 Kurtek, S. (2021) 针对含噪、稀疏且不完整函数型数据的同步配准与估计的贝叶斯框架。 *Journal of the American Statistical Association*, **0**, 1–17。Minnotte, M. C. (1997) 众数存在性的非参数检验。*The Annals of Statistics*, **25**, 1646–1660。Raket, L. L., Sommer, S. 和 Markussen, B. (2014) 用于函数型数据同步平滑与配准的非线性混合效应模型。*PatternRecognition Letters*, **38**, 1–7。Ramsay, J. 和 Silverman, B. (1997) *Functional DataAnalysis*。《统计学 Springer 系列》。Springer。Ramsay, J. O. 和 Li, X. (1998) 曲线配准。*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**, 351–363。Roser, M., Ritchie, H., Ortiz-Ospina, E. 和 Rod´es- Guirao, L. (2013) 世界人口增长。 *Our Worldin Data*。Samworth, R. J. 和 Sen, B. (2018) 编者按：形状约束下的非参数推断特刊。*Statistical Science*, **33**, 469–472。Simpson, I. J., Schnabel, J. A., Groves, A. R., Andersson, J. L. 和 Woolrich, M. W. (2012) 非刚性配准中正则化的概率推断。*NeuroImage*, **59**, 2438–2451。Srivastava, A. 和 Klassen, E. (2016) *Functional andShape Data Analysis*。《统计学 Springer 系列》。Springer New York。Srivastava, A., Wu, W., Kurtek, S., Klassen, E. 和 Marron, J. S. (2011) 使用 Fisher–Rao度量进行函数型数据配准。St¨ocker, A., Pfeuffer, M., Steyer, L. 和 Greven, S. (2022) 通过厄米协方差平滑对平面曲线进行弹性完全 Procrustes 分析。Tang, R. 和 Muller, H. G. (2008) 函数型数据的成对曲线同步。 *Biometrika*, **95**, 875–889。Telesca, D. 和 Inoue, L. Y. T. (2008) 贝叶斯分层曲线配准。*Journal of theAmerican Statistical Association*, **103**, 328–339。Wang, X. 和 Berger, J. O. (2016) 使用高斯过程估计形状约束函数。*SIAM/ASA Journal on Uncertainty Quantification*, **4**, 1–25。

Wegman, E. J. (1970) 单峰密度的最大似然估计，II。 *Ann. Math. Stat.*，**41**，2169–2174。Zomorodian, A. 和 Carlsson, G. (2005) 计算持久同调。 *Discrete and Computational Geometry*，**33**，249–274。