arXiv:2303.05043v1 [cs.LG] 9 Mar 2023

基于随机傅里叶特征的可逆核主成分分析

Daniel Gedon , Antˆonio H. Ribeiro、Niklas Wahlstr¨om 和 Thomas B. Sch¨on，高级会员，*IEEE*

摘要——核主成分分析（kPCA）是一种广泛研究的方法，用于在非线性变换后构建低维数据表示。从 kPCA 重建原始输入信号是去噪中的一项重要任务，而现有主流方法需要求解一个监督学习问题。本文提出一种替代方法，使重建过程能够自然地从压缩步骤中实现。我们首先使用随机傅里叶特征近似核。随后，我们利用非线性变换在某个子域内可逆这一事实。因此，该方法被称为可逆核主成分分析（*ikPCA*）。我们在不同数据模态上进行了实验，结果表明，ikPCA 在去噪任务上的表现与采用监督重建的 kPCA 相近，因此是一个很有力的替代方案。

索引词—**去噪、心电图、核主成分分析、原像、随机傅里叶特征、重构。**

I. 引言

PRINCIPAL 成分分析（PCA）涉及求取一个投影矩阵 $\boldsymbol{P}$，该矩阵将给定的输入$\boldsymbol{x} \in \mathbb{R}^p$ 变换为低维表示 $\boldsymbol{z} = \boldsymbol{P}\boldsymbol{x} \in \mathbb{R}^d$，其中 $d < p$。反过来，给定一个低维表示，可以通过逆变换$\hat{\boldsymbol{x}} = \boldsymbol{P}^\top \boldsymbol{z}$重建原始输入空间。数据 $\boldsymbol{x}$ 通常被假定位于低维流形上。在这种情况下，PCA 很有用，因为它能够提取数据中最重要的特征或最大变异方向。该算法是最优的 [1]，即不存在重构矩阵 $\boldsymbol{U}$ 和降维矩阵$\boldsymbol{V}$，能使原始向量与重建向量 $\|\hat{\boldsymbol{x}} - \boldsymbol{U}\boldsymbol{V}\boldsymbol{x}\|_2$之间的平均距离小于使用 $\boldsymbol{U} = \boldsymbol{P}^\top$和 $\boldsymbol{V} = \boldsymbol{P}$时的距离。重要的是，矩阵 $\boldsymbol{P}$ 既可用于降维，也可通过其转置来重建原始输入 $\boldsymbol{P}^\top$。

核主成分分析（kPCA）通过使人们能够研究非线性变换 [2]后的主成分，拓展了传统主成分分析。这使得关于数据位于低维线性流形上的假设可以推广到流形为非线性的情形。在这种情况下，传统主成分分析可能无法提取有用的低维表示 $\boldsymbol{z}$，但核主成分分析可能通过在将输入$\boldsymbol{x}$ 非线性变换 Φ 到一个（可能是无限维的）特征空间 $\mathcal{F}$后应用主成分分析而取得成功

$$\boldsymbol{z} = \boldsymbol{P}\Phi(\boldsymbol{x}). \tag{1}$$

稿件于2023年2月27日提交。此项工作部分由瓦伦堡人工智能、自主系统与软件计划（WASP）资助，该计划由克努特与爱丽丝·瓦伦堡基金会提供资金；由Kjell och M¨arta Beijer基金会资助；并由瑞典研究理事会（VR）通过“物理信息机器学习”项目资助（注册编号：2021-04321）

所有作者均隶属于瑞典乌普萨拉 751 05 乌普萨拉大学信息技术系（电子邮箱：{daniel.gedon,anto- nio.horta.ribeiro, niklas.wahlstrom, thomas.schon}@it.uu.se）

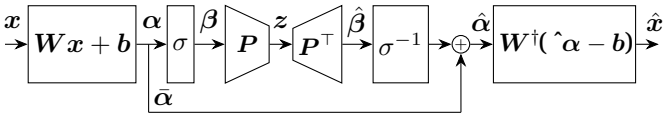


图1：我们提出的可逆核主成分分析方法示意图。

核主成分分析确实是一种自然且有价值的思路。然而，尽管降维可以轻松计算，如何获得重构结果 $\hat{\boldsymbol{x}}$ from $\boldsymbol{z}$却远非显而易见。

这个逆重建问题称为预像问题。其解决方案包括基于梯度下降的非线性优化 [3]，或利用特征空间中的距离约束 [4]。Bakır 等人提出的传播最广的解决方案 [5]，是应用 (1) 构建一个数据集 [6]，$(\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i=1}^n$，其中包含原始输入及其低维表示。目标是找到一个非线性函数 f 映射回 $\boldsymbol{z}_i$。例如，scikit-learn $\boldsymbol{x}_i$或 Julia 的多元统计软件包 [7]中均提供了这种方法。不过，这种方法存在一些缺点：与主成分分析不同，重建并非核主成分分析的直接副产品，而是需要求解一个监督学习（SL）问题。这里，我们将这种组合记为核主成分分析 [8]SL。此外，由于函数 +需要是非线性的，因此，寻找 f 与 $\boldsymbol{z}$ 之间映射的监督学习问题通常会导致一个非凸优化问题。事实上，同时实现降维和重建的直接非线性方法——例如自编码器 $\boldsymbol{x}$ 和变分自编码器 [9], [10] —— 其性能可能显著优于核主成分分析 [11], [12]SL。深度自编码器虽然是生成模型中常用的组成部分，但需要求解一个非凸优化问题。与之相反，核方法和主成分分析已得到充分研究，并被广泛用作预处理步骤。

我们提出了一种新的核主成分分析表述方法，可直接将重建方法作为其副产品。该方法适用于任何平移不变核。正如我们将在第 II 节中讨论的那样，这类核都可以通过如下类型的特征映射来近似 $\Phi(x) = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$，其中非线性函数为 σ，且 $\boldsymbol{W} \in \mathbb{R}^{r \times p}$。在这里，维度通过以下计算步骤依次降低

$$\boldsymbol{\alpha} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \tag{2a}$$
$$\boldsymbol{\beta} = \sigma(\boldsymbol{\alpha}), \tag{2b}$$
$$\boldsymbol{z} = \boldsymbol{P}\boldsymbol{\beta}. \tag{2c}$$

我们提出的方法是逐步反转各项操作，如图 1所示。如果某项操作$\sigma(\boldsymbol{\alpha})$ 无法反转，我们就将向量 $\boldsymbol{\alpha}$ 分解为两个

分量 $\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}}$ 和 $\bar{\boldsymbol{\alpha}}$，使得 $\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}}$ 属于 σ 可逆的定义域。我们可以使用主成分分析对第一分量进行压缩和解压缩，而第二分量则直接跳过。这样，我们就避免了任何非线性监督问题，并可直接进行重建。

II. 背景

设，$k(\boldsymbol{x},\ \boldsymbol{y})$ 是一个半正定核

$$k(\boldsymbol{x},\boldsymbol{y})=\langle\Phi(\boldsymbol{x}),\Phi(\boldsymbol{y})\rangle=\sum_i\phi_i(\boldsymbol{x})\phi_i(\boldsymbol{y}),\tag{3}$$

其中 $\Phi(\boldsymbol{x})=(\phi_1(\boldsymbol{x})\ \phi_2(\boldsymbol{x})\ldots)$ 表示一个将输入映射到特征空间 $\mathcal{F}$的数值序列。

A. 核主成分分析

对于一组观测值 $\{\boldsymbol{x}_i\}_{i=1}^n$，$\mathcal{F}\times\mathcal{F}$ 中的经验协方差矩阵定义为

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{i=1}^n\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_i)^\top.$$

该矩阵的谱分解给出

$$\hat{\boldsymbol{\Sigma}}=\sum_i\lambda_i\boldsymbol{v}_i\boldsymbol{v}_i^\top,$$

使得 $\lambda_1\geq\lambda_2\geq\ldots$。我们将投影到前 d 个分量定义为 $\boldsymbol{P}=\begin{bmatrix}\boldsymbol{v}_1&\cdots&\boldsymbol{v}_d\end{bmatrix}^\top$。由此，我们可以得到低维表示 $\boldsymbol{z}=\boldsymbol{P}\Phi(\boldsymbol{x})$。

B. 无限维特征映射

在实践中，核技巧使我们能够处理无限维特征空间。然而，我们在此提出的方法旨在用于有限维特征空间。因此，处理无限维特征映射时，我们将采用近似方法。具体而言，我们将使用截取至前 r 个分量的特征映射，记作 $\tilde{\Phi}(\boldsymbol{x})=(\phi_1(\boldsymbol{x})\ \phi_2(\boldsymbol{x})\ldots,\phi_r(\boldsymbol{x}))$ 来近似核，$\tilde{\ }$ $\tilde{\ }$这意味着我们可以写成 $k(\boldsymbol{x},\boldsymbol{y})\approx\langle\Phi(\boldsymbol{x})\ \Phi(\boldsymbol{y})\rangle$。

我们沿用 [13] 利用随机傅里叶特征来近似平移不变核的做法，即形式为 $k(\boldsymbol{x},\ \boldsymbol{y})=g(\boldsymbol{x}-\boldsymbol{y})$的核。博赫纳定理保证，当且仅当 $g(\delta)$ 是某个概率分布的傅里叶变换$p(\omega)$（可能经过重新缩放）时，该核连续且半正定。

以高斯核 $k(\boldsymbol{x},\ \boldsymbol{y})=\exp^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2}}$ 为例，只有在考虑无限维特征空间时，才能得到分解（3）。为使用随机傅里叶特征近似该特征空间，令 $\boldsymbol{W}\in\mathbb{R}^{r\times p}$ 为一个随机矩阵，其元素从分布 $p(\omega)$ 中独立同分布地抽取；再令 $\boldsymbol{b}\in\mathbb{R}^r$ 为一个从 $\mathcal{U}(-\pi,\ \pi)$中独立同分布抽取的向量。于是，我们得到

$$\tilde{\Phi}(\boldsymbol{x})=\sqrt{2}\sin(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}),$$

其中 正弦函数 是逐元素应用的。文献 [13] 证明了，$\tilde{\ }$ $\tilde{\ }\langle\Phi(\boldsymbol{x})\ \Phi(\boldsymbol{y})\rangle$ 一致收敛于 $k(\boldsymbol{x},\ \boldsymbol{y})$。此外，收敛速度关于 r是指数级的。

III. 可逆核主成分分析

考虑如下类型的特征映射

$$\Phi(\boldsymbol{x})=\sigma(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}),$$

其中 $\boldsymbol{W}\in\mathbb{R}^{r\times p},\ \boldsymbol{b}\in\mathbb{R}^r$ 且 σ 是逐元素应用的非线性函数。上一节的讨论说明，这些特征映射可用于近似与任意平移不变核相关联的空间。接下来，在降维是按照（2）计算的前提下，我们详细说明如何逆转这些操作。一个关键难点是，激活函数 σ通常不可逆。下面介绍我们应对这些问题的解决方案。

A. 不可逆激活函数

在大多数相关情形中，非线性函数 $\sigma:\mathbb{R}\to\mathbb{R}$,$\sigma:\ \alpha\mapsto\beta$ 在整个定义域 $\mathbb{R}$ 上不可逆，但它可能在某个子域 $\mathcal{X}\subset\mathbb{R}$ 上可逆。将函数 σ 在 $\mathcal{X}$ 上的限制记为 $\sigma_{\mathcal{X}}$，则逆函数 $\sigma_{\mathcal{X}}^{-1}$ 定义良好。令 $\bar{\alpha}=\alpha-\sigma_{\mathcal{X}}^{-1}\circ\sigma(\alpha)$。考虑两个例子：首先，对于 ReLU 激活函数 $\beta=\sigma(\alpha)=\max(\alpha,0)$，其可逆定义域为 $\mathcal{X}=[0,\infty)$。因此，$\sigma_{\mathcal{X}}^{-1}(\beta)=\beta$，且$\bar{\alpha}=\min(\alpha,0)$。其次，对于随机傅里叶特征中使用的 $\beta=\sigma(\alpha)=$正弦函数 α，我们有 σ 在 $\mathcal{X}=(-\pi/2,\pi/2]$上可逆。因此，$\sigma_{\mathcal{X}}^{-1}(\beta)=\arcsin\beta$，且对于某个 $k\in\mathbb{Z}$，$\bar{\alpha}=(-1)^k\alpha+\pi k$。

B. ikPCA

重建方法逐步逆转（2）中的运算。我们可以写成

$$\widehat{\boldsymbol{\beta}}=\boldsymbol{P}^\top\boldsymbol{z},\tag{4a}$$

$$\widehat{\boldsymbol{\alpha}}=\sigma_{\mathcal{X}}^{-1}(\widehat{\boldsymbol{\beta}})+\bar{\boldsymbol{\alpha}},\tag{4b}$$

$$\widehat{\boldsymbol{x}}=\arg_{\boldsymbol{x}}\min\|\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}-\widehat{\boldsymbol{\alpha}}\|_2^2+\lambda\|\boldsymbol{x}\|_2^2.\tag{4c}$$

第一步（4a）逆转降维，其依据与主成分分析相同：投影矩阵 $\boldsymbol{P}$使重构误差 $\|\hat{}\boldsymbol{\beta}-\boldsymbol{P}^\top\boldsymbol{z}\|_2$ 达到最小。第二步（4b）逆转子域 $\mathcal{X}$ 上的非线性函数 σ，并补上被绕过的不可逆部分 $\bar{\boldsymbol{\alpha}}$。最后一步（4c）通过求解一个岭回归问题，逆转线性映射 $\boldsymbol{x}\mapsto\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}$。请注意，对于 $\lambda\to0^+$，最后一步可简化为 $\widehat{\boldsymbol{x}}=\boldsymbol{W}^\dagger(\widehat{\boldsymbol{\alpha}}-\boldsymbol{b})$，其中 $\boldsymbol{W}^\dagger$是 $\boldsymbol{W}$的伪逆，如图 1所示。

IV. 数值示例

本节概述了用于评估所提出的 ikPCA 方法性能的实验。我们重点研究对各种模态的输入进行去噪这一任务。定量而言，我们评估去噪后的测试信号与真实无噪声测试信号之间的均方误差（MSE）。鉴于 ikPCA、PCA 和 kPCA+SL 在结构上相似，我们将 ikPCA 与这两种方法进行比较。为确保公平比较，我们未考虑具有层次化深度模型结构的去噪自编码器。在讨论中，我们将详细说明如何将我们的方法扩展到神经网络，以及如何设置实验以便与自编码器进行更恰当的比较。

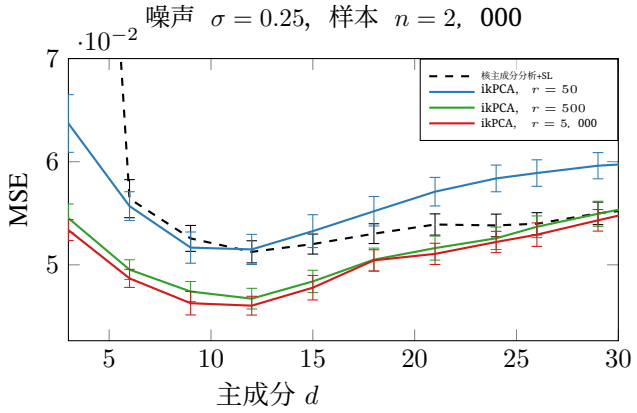


图 2：S 曲线玩具示例。ikPCA 采用不同数量的随机特征时的重构均方误差。

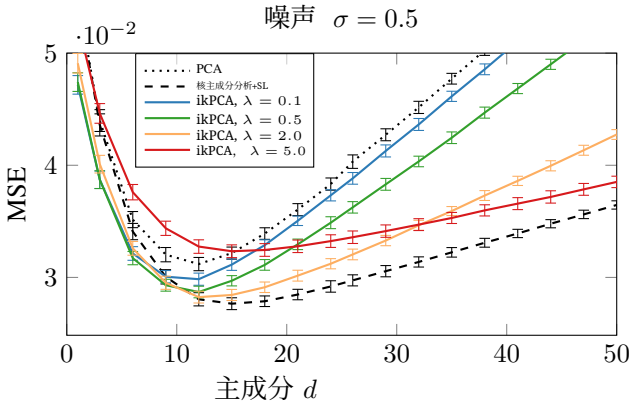


图 3：USPS 数据。正则化参数 λ的影响。

在所有实验中，我们均采用高斯核（或其随机傅里叶特征近似），并报告 20 次随机运行结果的均值和标准差。实验中固定的方法参数通过超参数网格搜索进行优化。为确保结果可复现，我们公开发布了代码[1]。

A. 合成玩具数据：*S* 曲线

我们使用 S 曲线玩具问题，生成了形状如字母“S”的合成三维数据点。对于训练和测试，我们生成 $n = 2{,}000$ 个数据点，并添加标准差为 $\sigma = 0.25$的高斯噪声。对于核主成分分析+SL，我们将核宽度设为 $\gamma = 1$，并将重构岭回归强度设为 $\lambda = 1$；对于 ikPCA，我们设置 $\gamma = 0.5$ 和 $\lambda = 1$。结果见图2。我们提出的 ikPCA 方法能够以与核主成分分析+SL 相当的方式对数据进行去噪。对于这个问题，我们发现，只需使用 $r = 50$ 个随机傅里叶特征，就足以达到与核主成分分析+SL 相当的性能。这使 ikPCA 相较于核主成分分析+SL 具有计算优势，因为后者需要求逆一个 $n \times n$ 矩阵。

B. USPS 图像

我们使用 USPS 数据集，其中包含以灰度格式呈现、尺寸为 16×16 的手写数字图像，并添加标准差为 $\sigma = 0.5$的高斯噪声。我们使用 $n = 1{,}000$ 张图像进行训练，并使用400 张图像进行测试。对于核主成分分析+SL，我们设置 $\gamma = 5 \cdot 10^{-3}$和 $\lambda = 10^{-2}$；对于 ikPCA，我们使用30,000 个随机傅里叶特征，并设置 $\gamma = 10^{-4}$。在图3中，我们改变 ikPCA 的正则化参数 λ。同样，我们的结果表明，当主成分数量+达到最优值时，ikPCA 的表现与核主成分分析 dSL 相似。图中还显示，无岭回归重构的性能与主成分分析相当。然而，过度正则化会对整体重构性能产生负面影响。

图4 展示了噪声尺度为$\sigma = 0.25$时的图像去噪结果。对于 ikPCA，我们选择了最优正则化参数$\lambda = 1.3$。主成分数量 d 则针对

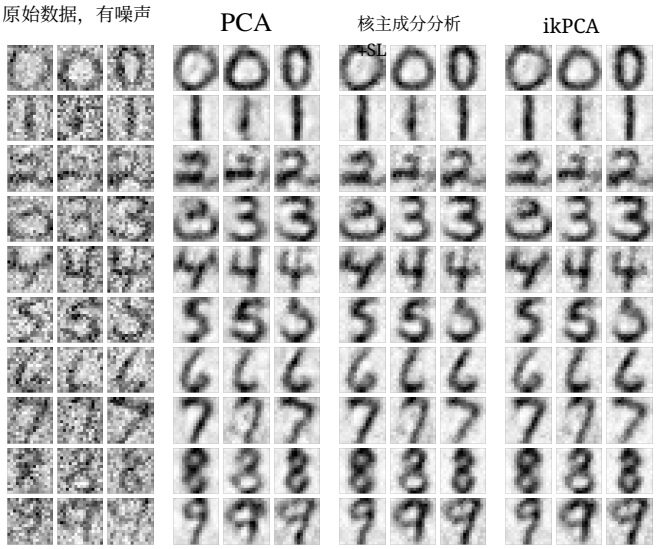


图 4：USPS 重建。

每种方法均选择使均方误差最小的参数。所有方法的图像去噪能力在视觉上都相近，这一点也得到了图3中均方误差差异的支持。

C. 心电图

心电图（ECG）是一项常规医学检查，用于记录心脏的电活动，通常用于诊断各种心脏疾病。然而，记录过程中测得的噪声可能会使诊断变得复杂。人们已提出多种方法来降低心电信号中的噪声。为进行比较，已有两种方法被提出：(1) 人为向信号中添加噪声，并与信号本身的原始版本进行比较[14]–[16], 或 (2) 对现有信号进行去噪，并将其与作为无噪声参考信号的平均心搏进行比较 [17], [18]。我们选择后一种方法，以考虑现实环境中的噪声情况。

我们使用来自中国生理信号挑战赛 2018年（CPSC）的心电图[2]，其中包含时长为 6 至 60 秒的数据 [19]。在 918 份无异常的心电图中，我们选取了时长最长的记录，并在此示例中聚焦于单个导联

[1]代码可在https://github.com/dgedon/invertible kernel PCA获取 _

[2]数据可在http://2018.icbeb.org/Challenge.html获取

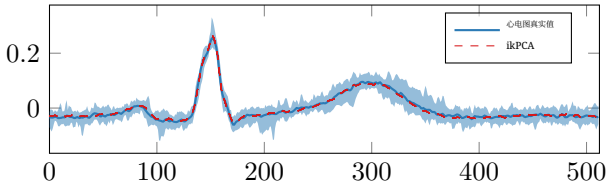


图5：对 I 导联的心电图心搏进行去噪。蓝色区域标示21个测试心搏的最小值和最大值。红色虚线显示使用 ikPCA 得到的所有测试重建结果。

。为提取心搏，我们首先使用高通滤波器去除基线漂移。然后，我们识别 R波峰[20]，，将每个峰之间的间隔重新采样为 512 个样本，最后按照 Johnstone 等人的预处理方法，将 R波峰定位在第 150 个样本处等人 [18]。

我们从所选心电图中提取了70个心搏；其中49个用于训练，21个用于测试。应用核主成分分析+SL，使用 $\gamma = 10$ 和$\lambda = 15$，以及 ikPCA，使用 $\gamma = 5 \cdot 10^{-5}$和 $\lambda = 10$，并采用至少512个随机傅里叶特征 r，仅使用第一个分量便实现了完美的信号去噪，如图5所示。在500次仿真中，ikPCA 的均方误差为 $2.6 \pm 0.8 \cdot 10^{-5}$，与核主成分分析+SL 的结果相近，而主成分分析的均方误差略高。

V. 计算注意事项

为 ikPCA 添加重建阶段并不会增加计算复杂度。原因是，获取重建的开销小于核主成分分析分解的开销（当输入维度p 小于样本数量 n时）。然而，我们的方法需要用 r个随机傅里叶特征来近似核映射。当 $r < n$ 时，这可能会降低计算开销；但当 $r > n$ 时，与核主成分分析相比，计算开销会增加 $r/n + 1$ 倍。

a) 核主成分分析的计算开销：某些核函数有闭式表达式，可通过 $\mathcal{O}$次运算计算得出。此时，核主成分分析的开销主要来自格拉姆矩阵的求逆$\boldsymbol{K}$ ，这需要 $\mathcal{O}(n^3)$次运算。格拉姆矩阵的矩阵元素 (i, j) 等于 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$。

b) 特征空间中主成分分析的计算开销：在 ikPCA 中，我们用有限维的 r维特征 $\Phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}) \ldots, \phi_r(\boldsymbol{x}))$ 近似核函数，并对特征的协方差矩阵 $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_i^n \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_i)^\top$执行主成分分析。计算矩阵各元素及其谱分解需要 $\mathcal{O}(r^2 n + r^3)$次运算。因此，对于$r < n$，近似核函数并计算 $\boldsymbol{\Sigma}$ 的谱分解，可能比像核主成分分析那样直接处理格拉姆矩阵 $\boldsymbol{K}$ 更高效。

然而，如果 $r > n$ 这一优势减弱，那么使用格拉姆矩阵 $\boldsymbol{K} = \Phi(\boldsymbol{X})^\top \Phi(\boldsymbol{X})$ ^ ^而非 $\boldsymbol{\Sigma}$可能更有效。 $\boldsymbol{K}$ 与 $n\boldsymbol{\Sigma}$具有相同的（非零）特征值，其特征向量乘以 $\Phi(\boldsymbol{X})$后，即可得到 n ^$\boldsymbol{\Sigma}$的特征向量。此形式的计算成本为 $\mathcal{O}(rn^2 + n^3)$。因此，其成本是采用具有闭式解的核函数时所得成本的 $r/n + 1$ 倍。

根据 [13]， $r = \Omega\left(\frac{p}{\epsilon^2} \log \frac{D}{\epsilon}\right)$ 需要 个随机傅里叶特征才能确保近似误差在直径为 D 的空间上小于 ϵ。因此， r 随输入维度 p 线性增长。在 S 曲线示例中，$r \ll n$；而在 USPS 示例中， $r > n$，由于 p 较大，我们的方法失去了计算优势。对于像后者这样的高维数据，可以使用 Nyström 近似 [21]，在 r [22] 方面也可能更高效。

c) 重建成本：ikPCA 中的重建成本主要来自求解优化问题(4c)。为此，我们需要计算 $\boldsymbol{W}$的奇异值分解，仅需计算一次，成本为 $\mathcal{O}(p^3 + p^2 r)$。因此，对每个新的 $\mathcal{O}(pr + dr)$，重建的求解成本为。 ^$\boldsymbol{\alpha}$。

VI. 结论与讨论

我们提出了 kPCA 的一个可逆版本+SL。传统方法通过求解一个有监督问题，将潜在空间映射回输入空间；而我们的方法能够自然地获得这一映射。我们使用随机傅里叶特征 $\Phi(x) = \sigma(\boldsymbol{W}x + \boldsymbol{b})$近似核变换。尽管非线性函数 σ 可能不可逆，我们发现它在某个子域内可以求逆。我们可以利用这一发现，将其输入分解为可逆部分和不可逆部分，并绕过后者。我们通过三个示例展示了该方法在去噪方面的有效性：S 曲线玩具问题、美国邮政服务图像数据集和心电图。

我们将我们的方法与对称核主成分分析+SL进行比较。此处的对称性意味着，用于压缩和重建的核被定义为相同，这一设计受到常见框架中实现方式的启发 [7]，[8]。然而，文中 [6] 的方法在设计上并不局限于此。相反，由于重建过程中需要对非线性变换进行自然求逆，ikPCA 必须采用非对称设置。尽管在我们展示的数值实验中，我们的方法与核主成分分析+SL表现相当，但尚不确定其与经过良好调优的非对称核主成分分析+SL相比会如何。

尽管我们的方法很简单，但仍有广泛的可能扩展。为了扩展表征能力，我们可以以层次化方式堆叠多层 $\Phi(x) = \sigma(\boldsymbol{W}x + \boldsymbol{b})$ 变换。由此，我们得到一种更接近深度自编码器的结构。这可能有助于在理论上已得到充分确立的核方法领域与神经网络之间建立更多联系。类似地，我们可以将方法中的随机傅里叶特征视为一个未经训练的单层神经网络。将我们的方法扩展到经过训练的神经网络，将能够在无需重新训练的情况下自然地执行重建任务。最后，我们针对欠参数化数据（使用 $p/n \approx 0.25$ 的 USPS 示例）和过参数化数据（使用 $p/n \approx 10$ 的心电图示例）进行了实验。这一点结合使用大量随机傅里叶特征，引发了关于去噪模型的过参数化和良性过拟合的问题 [23]，[24]。

致谢

相关计算由林雪平大学国家超级计算中心提供的超级计算资源Berzelius支持完成。克努特与爱丽丝·瓦伦堡基金会。

参考文献

[1] S. Shalev-Shwartz 和 S. Ben-David，《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。[2] B. Sch¨olkopf、A. Smola 和 K.-R. M¨uller，“核主成分分析”，载于《国际人工神经网络会议》，1997年，第583–588页。[3] C. J. C. Burges，“简化的支持向量决策规则”，载于《国际机器学习会议》，1996年，第71–77页。[4] S. Mika、B. Sch¨olkopf、A. Smola、K.-R. M¨uller、M. Scholz 和 G. R¨atsch，“特征空间中的核主成分分析与去噪”，《神经信息处理系统进展》，第11卷，1998年。[5] J. T. Kwok 和 I. W. Tsang，“核方法中的原像问题”，载于《第20届国际机器学习会议论文集》，2003年，第408–415页。[6] G. H. Bakır、J. Weston 和 B. Sch¨olkopf，“学习寻找原像”，《神经信息处理系统进展》，第16卷，第449–456页，2004年。[7] F. Pedregosa、G. Varoquaux、A. Gramfort、等，“Scikit-learn：Python中的机器学习”，《机器学习研究期刊》，第12卷，第2825–2830页，2011年。[8] J. Bezanson、A. Edelman、S. Karpinski 和 V. B. Shah，“Julia：一种数值计算的新方法”，《SIAM评论》，第59卷，第1期，第65–98页，2017年。[9] H. Bourlard 和 Y. Kamp，“ 多层感知器与奇异值分解实现的自联想 ”，《生物控制论》，第59卷，第4–5期，第291–294页，1988年。[10] G. E. Hinton 和 R. Zemel，“自编码器、最小描述长度与亥姆霍兹自由能”，《神经信息处理系统进展》，第6卷，1993年。[11] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，载于《国际学习表征会议》，2014年。[12] D. J. Rezende、S. Mohamed 和 D. Wierstra，“深度生成模型中的随机反向传播与近似推断”，载于《国际机器学习会议》，PMLR，2014年，第1278–1286页。[13] A. Rahimi 和 B. Recht，“大规模核机器的随机特征”，载于《神经信息处理系统进展》第20卷，2008年，第1177–1184页。[14] R. Sameni、M. B. Shamsollahi、C. Jutten 和 G. D. Clifford，“用于心电图去噪的非线性贝叶斯滤波框架”，《IEEE生物医学工程汇刊》，第54卷，第12期，第2172–2185页，2007年。[15] P. Xiong、H. Wang、M. Liu、S. Zhou、Z. Hou 和 X. Liu，“基于改进去噪自编码器的心电信号增强”，《人工智能工程应用》，第52卷，第194–202页，2016年。[16] H.-T. Chiang、Y.-Y. Hsieh、S.-W. Fu、K.-H. Hung、Y. Tsao 和 S.-Y. Chien，“使用全卷积去噪自编码器降低心电信号中的噪声”，《IEEE Access》，第7卷，第60 806–60813页，2019年。[17] F. Castells、P. Laguna、L. S¨ornmo、A. Bollmann 和 J. M. Roig，《心电图信号处理中的主成分分析》，*EURASIP*《信号处理进展期刊》，第 2007 卷，第 1–21 页，2007年。[18] I. M. Johnstone 和 A. Y. Lu，《稀疏主成分分析》，*arXiv* 预印本 *arXiv:0901.4392*，2009年。[19] F. Liu、C. Liu、L. Zhao、等，《用于评估心电图节律和形态异常检测算法的开放访问数据库》，《医学影像与健康信息学期刊》，第 1368–1373 页，2018年。[20] C. Xie、L. McCullum、A. Johnson、T. Pollard、B. Gow 和 B. Moody，*Python* 波形数据库软件包（*wfdb*），2023年。[21] C. Williams 和 M. Seeger，《使用 Nystr¨om 方法加速核机器》，《神经信息处理系统进展》，第 13 卷，2000年。[22] T. Yang、Y.-F. Li、M. Mahdavi、R. Jin 和 Z.-H. Zhou，《Nystr¨om 方法与随机傅里叶特征：理论与实证比较》，《神经信息处理系统进展》，第 25 卷，2012年。[23] A. Radhakrishnan、K. Yang、M. Belkin 和 C. Uh- ler，《过参数化自编码器中的记忆》，*arXiv* 预印本 *arXiv:1810.10333*，2018年。[24] P. L. Bartlett、P. M. Long、G. Lugosi 和 A. Tsigler，《线性回归中的良性过拟合》，《美国国家科学院院刊》，第 117 卷，第 48 期，第 30 063–30070 页，2020年。

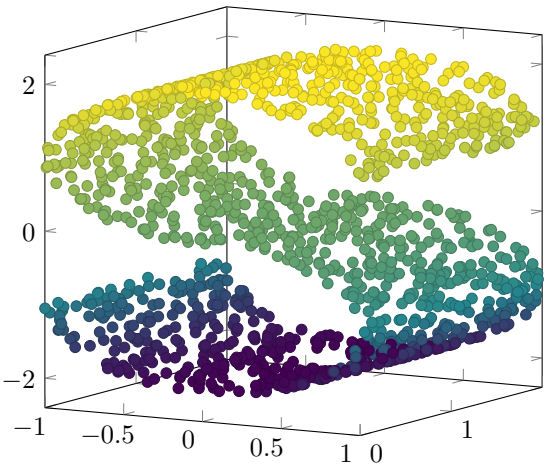
附录

A. S 曲线数据集的补充结果

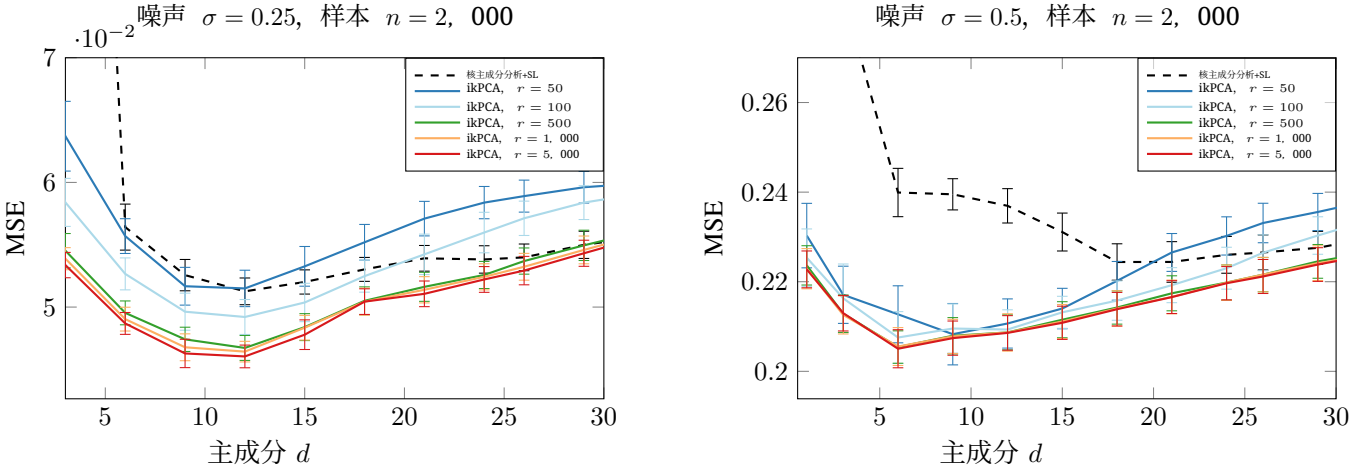
S 曲线由以下方程组生成，其中变量 t 通常用作标签[3]。对于我们的用途，我们不需要标签，只关注输入 $\boldsymbol{x}$。向 $\boldsymbol{v} \sim \mathcal{N}(0,\ \sigma \boldsymbol{I}_3)$ 添加高斯噪声 $\boldsymbol{x}$。

$$
\begin{aligned}
t &\sim \mathcal{U}\left(-\frac{3}{2}\pi, \frac{3}{2}\pi\right), \\
x_1 &= \sin t, \\
x_2 &\sim \mathcal{U}(0, 2), \\
x_3 &= \operatorname{sign}(t)\,(\cos t - 1).
\end{aligned}
$$

图A-1 展示了 S曲线数据集的可视化结果。图 A-2 是图 2 的扩展，使用了更多的随机傅里叶特征 r，并且加性噪声的水平提高了一级。对于这个数据集，所需的随机傅里叶特征少于训练数据点。因此，我们的方法在数值计算上更快。仅需 $r = 500$ 个特征（少于样本数的 $1/4$ $n = 2,\ 000$）即可达到最佳性能曲线。我们发现，在噪声值较大时，ikPCA 的表现甚至优于 kPCA+SL 和 PCA，且 r 的影响较不明显。



图A-1：S 曲线数据集的可视化。颜色表示回归标签 t.



图A-2：随机特征分量数量对重构均方误差的影响。图2 是此处左图的修改版本。

[3]另请参阅https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_s_curve.html

B. 美国邮政服务数据集的更多结果

对于以下图表，选择 kPCA+SL 的超参数（即核宽度 γ 和正则化强度 λ），使重构均方误差达到最低。我们采用了网格搜索。所有结果均展示了 20 次随机运行的均值（误差图中还展示了标准差）。

图 A-3 探讨了随机傅里叶特征数量 r 对该数据集的影响。我们观察到，通常随机傅里叶特征越多，渐近结果越好。此外，我们注意到，我们的方法 ikPCA 会逐渐接近核主成分分析+SL $r \to \infty$，这与核近似的结果一致。

图 A-4 探讨了正则化参数 λ 对我们 ikPCA 方法重建的影响。我们观察到，需要找到一个最优权衡。对于 $\lambda \to 0^+$，ikPCA 的性能会接近主成分分析。相反，当 λ取值较大时，问题会变得过度正则化，且不再具有泛化能力。

图A-5 展示了加性噪声水平与为潜在空间选取的主成分数量 d 的综合影响。每种方法中均方误差最低时对应的成分数量显示在图 A-6左侧的图中。我们观察到，噪声值越大，最优主成分的数量越少 d，这是因为噪声水平会影响更大一部分奇异值。随后，图A-6 展示了选择最优主成分数量 d时三种方法的均方误差值。我们观察到，在该数据集中，各方法经过最优调参后的均方误差在不同噪声水平下都相近。

图 A-7展示了在为所有方法选择最优超参数时，两种不同噪声水平下 USPS 图像的重建结果。正如图 A-6 右侧图中的定量比较所示，这些重建结果在定性上也相似。

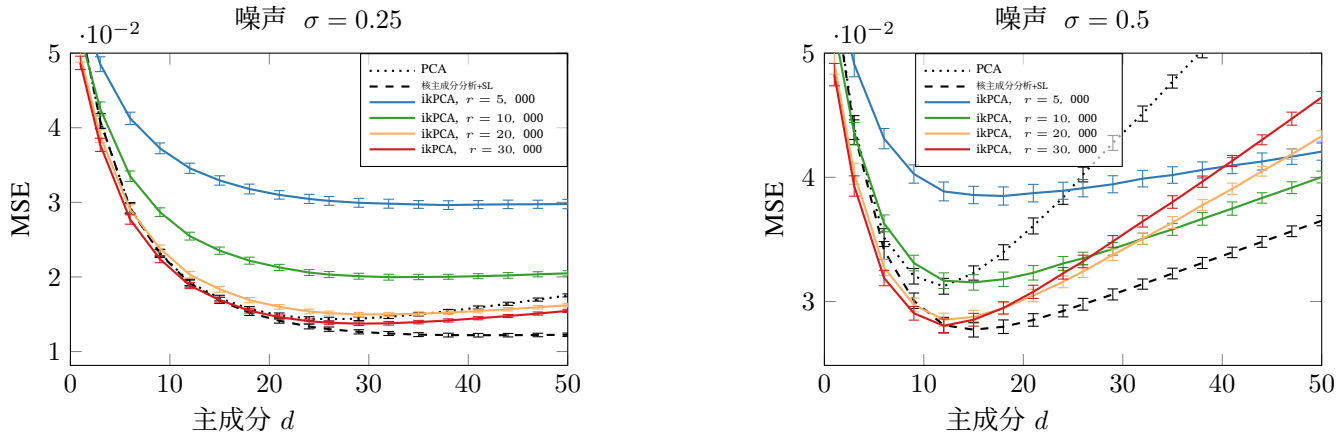


图 A-3：随机特征分量数量不同对重建均方误差的影响 E.

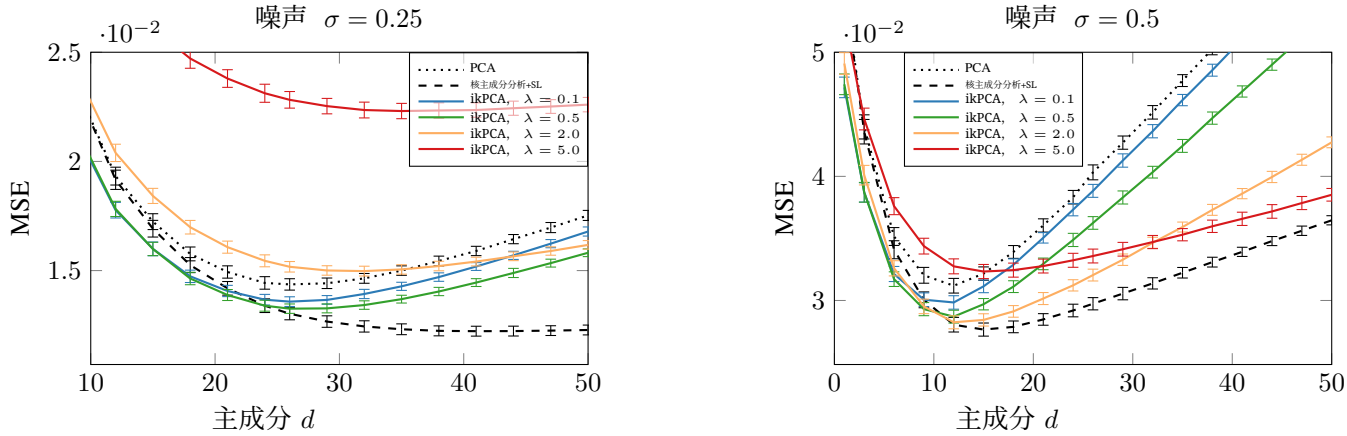


图 A-4：正则化参数 λ 对重建均方误差的影响。需要选择一个最优值。右图重复展示了图 3。

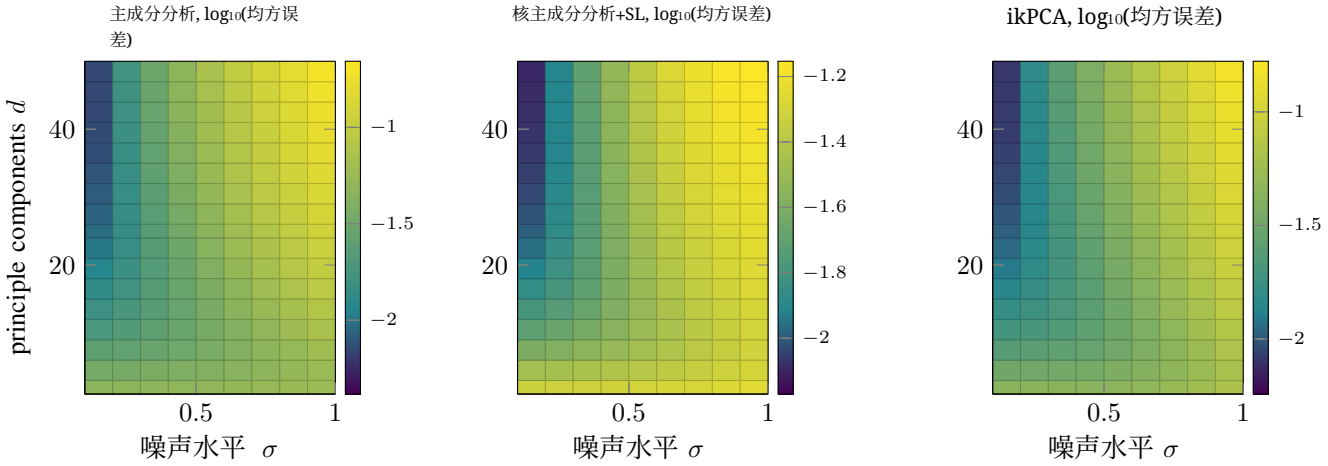


图 A-5：噪声和潜在空间维度对重构均方误差的综合影响。

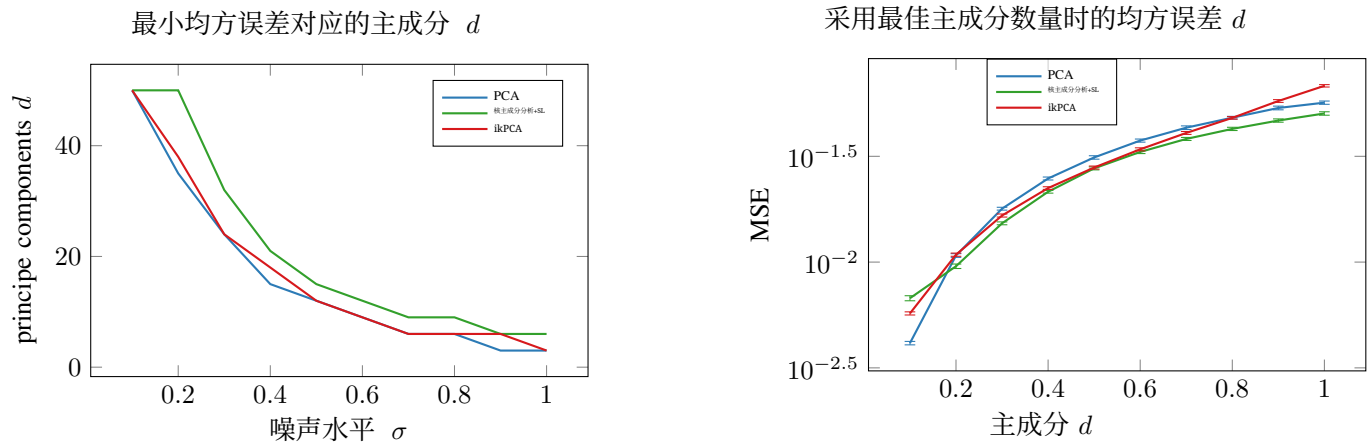
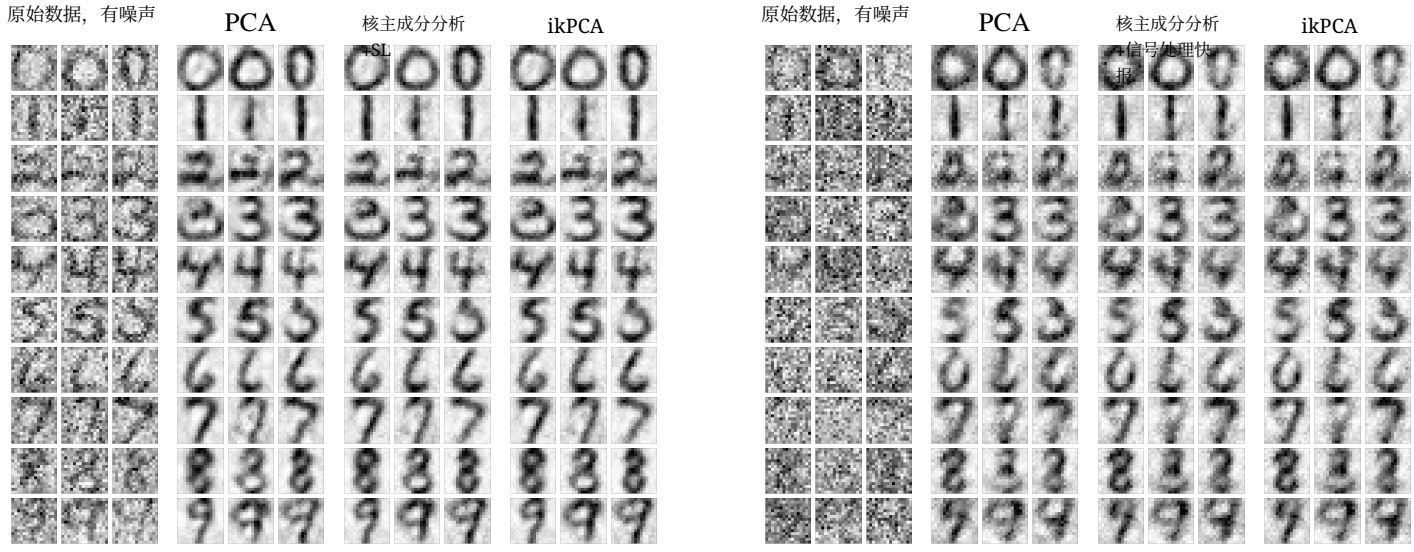


图 A-6：噪声的影响。（左）在给定噪声水平下，使均方误差达到最低的最佳成分数。（右）采用最佳成分数量进行重建时，三种方法的均方误差。



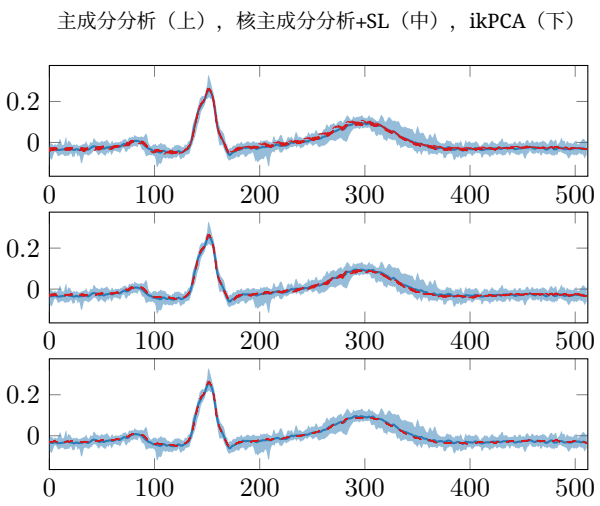
（a）噪声水平 $\sigma = 0.25$ (b) 噪声水平 $\sigma = 0.5$

图 A-7：不同方法的重建。针对每种方法选择最优超参数，以获得最低均方误差。左图重复展示了图4。

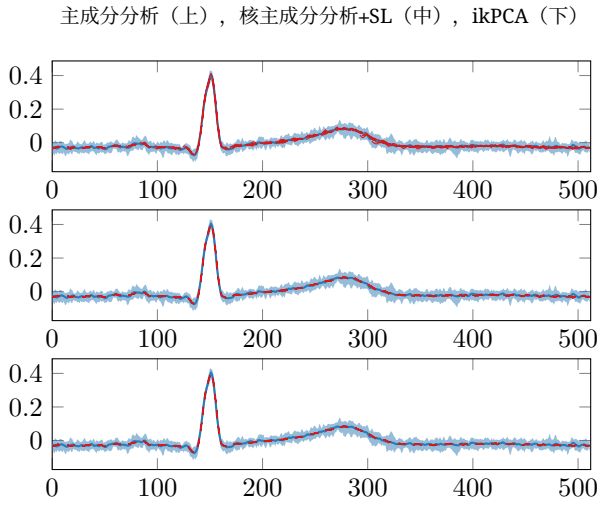
C. 心电图数据的其他结果

图 A-8 展示了另外两个心电图信号重建示例，与图 5相互补充。我们采用与正文相同的超参数。在右侧图中可以看到，对于主成分分析，有些重建结果（红色虚线）并不理想，即与真实值曲线相距较远。这导致均方误差显著增大。核主成分分析+SL 和 ikPCA 的表现相近。

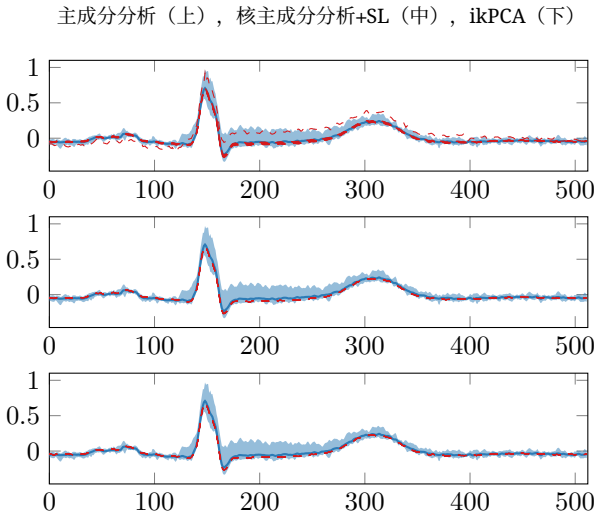
表 A-1 比较了三条心电图轨迹（见图 A-8）在不同训练/测试划分下进行 500 次仿真所得的均方误差。我们观察到，ikPCA 和核主成分分析+SL 的均方误差表现相近，而主成分分析的均方误差略高。因此，纯线性模型的心电图去噪效果不够理想。



（a）由 70 个心搏组成的心电图 I导联（训练/测试为 49/21）。



（b）由 62 个心搏组成的心电图 I导联（训练/测试为 43/19）。



（c）由 40 个心搏组成的心电图 II导联（训练/测试为 28/12）。

图 A-8：更多心电图重建结果。图A-8a与图 5中的示例相同。

	PCA	核主成分分析+SL	ikPCA	[单位]
心电图 (a)	4.00 ± 1.47	2.78 ± 0.74	2.57 ± 0.79	$[10^{-5}]$
心电图 (b)	3.20 ± 0.35	2.38 ± 0.29	2.32 ± 0.31	$[10^{-5}]$
心电图 (c)	8.37 ± 5.93	2.43 ± 1.63	2.27 ± 1.49	$[10^{-4}]$

表 A-1：不同心电图的重构均方误差。