多重传输分布回归

Yuanying Chen，上海财经大学统计与数据科学学院　Tongyu Li，新加坡国立大学统计与数据科学系　Yang Bai，上海财经大学统计与数据科学学院　Zhenhua Lin*新加坡国立大学统计与数据科学系

arXiv:2601.03674v1 [stat.ME] 7 Jan 2026

摘要

我们研究分布对分布回归问题，其中响应分布取决于多个分布型预测变量。这类情形自然出现在结果分布由多个异质分布来源驱动的应用中，但由于Wasserstein 空间的非线性几何结构，处理起来仍颇具挑战。我们提出一种内禀回归框架，通过 Wasserstein 空间中的加权Fréchet均值聚合特定于预测变量的传输后分布。所得模型可纳入多个分布型预测变量，赋予具有可解释性的权重以量化各预测变量的相对贡献，并定义一种灵活的回归算子，使其不受参考分布选择等辅助构造方式的影响。从理论角度看，我们证明了所诱导回归算子的可识别性，并在预测Wasserstein半范数下推导出其估计的渐近保证；该半范数直接刻画复合预测映射的收敛。大量模拟研究和一项真实数据应用表明，与现有 Wasserstein回归方法相比，所提方法具有更优的预测性能和可解释性。

关键词： 分布对分布回归、多重分布预测变量、Wasser- stein空间、加权Fréchet均值

*通讯作者： linz@nus.edu.sg。

1 引言

分布数据分析作为一种统计框架，旨在分析以概率分布形式表示的数据，例如直方图、密度函数和分位数函数，近年来受到越来越多的关注。这一研究方向并非将此类对象简化为标量或向量摘要，而是聚焦于分布层面的统计推断与建模，使方法能够捕捉更丰富的结构信息。该领域的近期综述，例如 Petersen等人(2022)，凸显了其在众多领域中的广泛适用性，包括人口学（Hron et al. 2016、Bigot etal. 2017）、环境科学（Nerini & Ghattas 2007）、生物医学研究（Petersen & Müller 2016）以及金融计量经济学（Kokoszkaet al. 2019）。

在这一框架中，回归分析发挥着核心作用。在许多当代应用中，预测变量和响应变量都是概率分布，由此产生了分布对分布的回归问题（Ghodrati & Panaretos 2022、Chen等，2023、Zhu& Müller 2025、Ghosalet al. 2025）。这类情形广泛存在于不同领域，包括死亡率预测（未来的死亡年龄分布取决于多个历史人口分布）、气候科学（区域结果分布受多种分布式气候强迫驱动）以及生理信号分析（响应汇总来自异质分布来源的信息）。这类问题的一个显著特点是，响应分布同时受到多个预测变量分布的影响，因此多分布回归具有重要的实际意义，也亟需能够处理多个分布输入的回归框架。

分布对分布回归的一项核心难题，源于 Wasserstein 空间缺乏线性结构。这种非欧几里得结构使经典线性回归工具无法适用，并催生了多种替代性方法。一类研究通过适当的变换将分布嵌入希尔伯特空间（

变换（Petersen & Müller 2016，Chen 等，2019），另一类则利用 Wasserstein 流形的切空间线性化，在局部线性化空间中进行回归（Chen 等，2023，Zhang 等，2022）。尽管这些方法可能有效，但它们可能会扭曲 Wasserstein 空间的内在几何结构，或依赖于使解释和推广变得复杂的局部构造。近年来，人们提出了基于最优传输（OT）的回归模型（Ghodrati &Panaretos2022，Zhu & Müller2023，Ghodrati &Panaretos 2024），这类模型通过传输映射直接关联预测分布与响应分布，从而得到内在定义的回归算子，既保留了底层 Wasserstein 几何结构，又能在分布域中作出清晰的解释。

尽管取得了这些进展，要将内在 Wasserstein回归模型扩展到包含多个预测变量分布的情形，仍然颇具挑战。Wasserstein 空间的非线性几何使得难以定义既良定义、可解释又满足交换律的聚合运算，而简单的扩展可能会受到结构性限制。近期研究已开始着手解决这些问题。例如，测地最优传输（GOT）回归（Zhu & Müller 2025）沿 Wasserstein 测地线定义加法和数乘，但施加了严格的结构约束，且所得估计依赖于预测变量排序。基于平行传输（Chen 等，2024）的替代构造缓解了其中一些局限，但要为多变量分布对分布回归建立灵活且易于解释的框架，仍面临挑战。

为弥补这些不足，我们提出一种新的内在分布对分布回归框架，该框架基于 Wasserstein空间中的加权 Fréchet 均值。从结构上看，所提出的模型将条件回归算子表示为特定于预测变量的传输后分布的加权 Fréchet 均值。每个预测变量通过一对由传输映射及其对应的标量权重组成的元素进入

模型，这二者共同刻画预测变量分布如何变形并对响应分布作出贡献。我们引入一个固定的参考分布作为锚点，其作用类似于经典回归中的截距。重要的是，尽管选择不同的参考分布会导致不同的单独传输映射，但最终得到的回归算子和预测值保持不变，从而确保模型不依赖辅助构造。这种表述得到了一种灵活且可扩展的回归算子，能够自然地处理多个预测变量分布，同时保持可解释性并避免依赖预测变量排序。

我们的方法具有三项主要优势：

1. 灵活性。该模型能够灵活地刻画依赖于预测变量的变换。具体而言，变换映射不局限于指向弗雷歇均值的测地线，因此能够涵盖更丰富的分布形状变形。2. 可扩展性。引入参考分布后，可以自然地从单个预测变量扩展到多个预测变量，同时保持可识别性。参考分布可以自由选择，前提是其固定且已知。3. 可解释性。回归算子可分解为传输映射和受限于单位单纯形的标量系数，从而清晰且直观地量化预测变量效应。

从理论角度出发，我们证明了所诱导回归算子的可识别性，并推导了所提出的 M 估计量在预测性 Wasserstein 半范数下的渐近保证。在该半范数下，总体风险在真实回归算子附近具有全局二次展开，尽管参数空间具有无限维和非欧几里得性质，仍提供了强有力的识别结构。该框架得出的一致性和收敛速率直接对应于预测

准确性，而无需分别识别各个传输映射或权重。由此得到的收敛行为呈现两阶段结构。当回归算子基于理想化分布进行估计时，估计量达到由单调传输映射类的内在熵所决定的立方根收敛速率。当分布通过经验样本观测时，则会引入额外的近似误差；我们的分析刻画了由受试者数量与分布内样本量的相对增长速率所决定的相变。

从计算角度来看，所提出的形式可实现高效且可扩展的估计程序。利用一维Wasserstein距离的显式分位数表示，无限维优化问题可化简为一系列有限维凸规划。具体而言，更新每个传输映射相当于进行加权保序回归，而更新权重向量则对应于单纯形上的约束最小二乘问题。这种分块结构带来了稳定的数值实现，并且能够自然地适应预测变量数量的变化。

大量模拟研究展示了所提方法在有限样本下的预测表现。具体而言，我们在一系列场景中，将均方根预测误差（均方根误差）与现有最优传输回归（Ghodrati & Panaretos 2022）和 GOT 回归（Zhu &Müller 2025）方法进行比较，涵盖单预测变量和多预测变量设置。针对死亡率预测的真实数据应用进一步表明，所提出的框架不仅能够给出准确的预测值，还能对预测变量特定的分布效应进行可解释的评估。

本文其余部分安排如下。第2节简要回顾 Wasserstein空间的数学基础，随后在第3节介绍我们的回归模型，包括关键的可识别性结果和估计框架。在第4节，我们将确立估计量的理论性质。计算算法

详见第5节。随后，我们通过第6节中的模拟验证所提出的模型；第7 节则通过死亡率数据的应用展示该方法。证明和辅助结果见补充材料。

2 预备知识

2.1 最优传输

我们首先简要回顾 Wasserstein空间和最优传输映射。关于更多背景知识，参见例如 Villani (2021)、Santambrogio (2015)和Ambrosio (2008年)。令$\mathcal{S} = [s_0, s_1]$ 为 $\mathbb{R}$ 的一个紧区间，并令$\mathcal{W}_2(\mathcal{S})$表示具有有限二阶矩的$\mathcal{S}$ 上概率测度的集合，即$\mathcal{W}_2(\mathcal{S}) = \{\mu \in \mathcal{P}(\mathcal{S}) : \int_{\mathcal{S}} |x|^2 d\mu(x) < \infty\}$，其中$\mathcal{P}(\mathcal{S})$是$\mathcal{S}$上的概率测度集合。对于两个相对于勒贝格测度绝对连续的测度 $\mu, \nu \in \mathcal{W}_2(\mathcal{S})$，平方 2-Wasserstein距离可表示为

$$d^2_{\mathcal{W}}(\mu, \nu) = \int_{\mathcal{S}} \{F_\nu^{-1} \circ F_\mu(x) - x\}^2 \, \mathrm{d}\mu(x) = \int_0^1 \left| F_\mu^{-1}(p) - F_\nu^{-1}(p) \right|^2 \, \mathrm{d}p,$$

其中 F_ν^{-1}是 ν的分位数函数， F_μ 是 μ的累积分布函数。从 μ 到 ν 的相应最优传输映射为 $T_{\mu\to\nu} := F_\nu^{-1} \circ F_\mu$，这是一个将 μ 推前至 ν的单调映射。受这一显式表示的启发，我们将广泛使用传输映射。

2.2 瓦瑟斯坦弗雷歇均值

接下来，我们回顾 Wasserstein空间（Agueh 与 Carlier， 2011）中弗雷歇均值的概念，它是通过 $\mathcal{W}_2(\mathcal{S})$的度量结构定义的。在总体层面，设$\xi$ 是 $\mathcal{W}_2(\mathcal{S})$ 中的随机概率测度，其分布为 P。Wasserstein 弗雷歇

均值定义为泛函的任意一个极小化元：

$$F(b) = \int_{\mathcal{W}_2(\mathcal{S})} d^2_{\mathcal{W}}(b, \xi) \mathrm{d}P(\xi),\ b \in \mathcal{W}_2(\mathcal{S}).$$

在有限样本情形下，令 $\mu_1, \ldots, \mu_n \in \mathcal{W}_2(\mathcal{S})$，并令 $\lambda_1, \ldots, \lambda_n$ 为满足 $\sum_{i=1}^n \lambda_i = 1$ 的非负权重。相应的（加权）Wasserstein Fréchet均值定义为下式的任一极小化元：

$$F(b) = \sum_{i=1}^n \lambda_i\, d^2_{\mathcal{W}}(b, \mu_i),\ b \in \mathcal{W}_2(\mathcal{S}).$$

可以保证此类极小化元存在；若至少一个 μ_i 相对于勒贝格测度绝对连续，则极小化元唯一。在第 2.1 节的一维最优传输设定下，Wasserstein Fréchet均值可用分位数函数简洁地刻画。具体而言，若 $\mu_1, \ldots, \mu_n$ 的分位数函数分别为 $F^{-1}_{\mu_1}, \ldots, F^{-1}_{\mu_n}$，则 Fréchet 均值 $\bar{\mu}$ 的分位数函数为

$$F^{-1}_{\bar{\mu}}(p) = \sum_{i=1}^n \lambda_i F^{-1}_{\mu_i}(p), \quad p \in (0, 1).$$

在总体层面，将有限和替换为期望，即可得到类似的刻画。

2.3 记号

本节最后介绍一些记号约定。对于函数 f 和 g，我们用 $f \circ g$ 表示它们的复合，定义为 $(f \circ g)(x) = f\{g(x)\}$。对于 $a, b \geq 0$，我们记作 $a \lesssim b$，或等价地记作 $b \gtrsim a$，如果存在一个通用常数 $C > 0$ 使得 $a \leq Cb$。概率测度 P 的支集记为 $\mathrm{supp}(P)$。对于可测函数 $f : \mathcal{S} \to \mathbb{R}$ 以及定义在 $\mathcal{S}$ 上的概率测度 μ，我们用 $\|f\|_{L^2(\mu)}$ 表示其关于 μ 的 L^2 范数。当 μ 是 $[0, 1]$, 上的勒贝格测度时，我们简记为 $\|f\|_{L^2}$。令 Δ^p 表示 $\mathbb{R}^{p+1}$中的概率单纯形，

$$\Delta^p := \left\{ \boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^\top \in [0, 1]^{p+1} : \sum_{j=0}^p \alpha_j = 1 \right\}.$$

对于可测映射 $T:\mathcal{S} \to \mathcal{S}$ 以及定义在 $\mathcal{S}$ 上的概率测度 μ，推前测度 $T\#\mu$ 定义为 $(T\#\mu)(A) = \mu(T^{-1}(A))$ $A \subset \mathcal{S}$可测。

3 方法

现在，我们正式介绍针对具有多个预测变量分布场景而设计的回归模型。设 $(\boldsymbol{\xi}, \eta) = (\xi_1, \xi_2, \ldots\ldots, \xi_p, \eta)$ 为取值于 $(\mathcal{W}_2(\mathcal{S}))^p \times \mathcal{W}_2(\mathcal{S})$ 中的随机元素，其联合分布为 P，并假设每个分布都关于 $\mathcal{S}$ 上的勒贝格测度绝对连续。类比欧几里得空间中的均值回归模型（其中回归函数为条件期望 $f(X) := \mathbb{E}(Y \mid X)$），我们将条件弗雷歇均值型回归算子 $\Gamma : (\mathcal{W}_2(\mathcal{S}))^p \to \mathcal{W}_2(\mathcal{S})$ 定义为

$$\Gamma(\boldsymbol{\xi}) := \underset{b \in \mathcal{W}_2(\mathcal{S})}{\arg\min} \int_{\mathcal{W}_2(\mathcal{S})} d^2_{\mathcal{W}}(b, \eta)\, \mathrm{d}P(\eta \mid \boldsymbol{\xi}),$$

其中，在温和的正则性条件下，极小化元是唯一的。

为获得一个易于处理且具有可解释性的回归结构，我们引入一个固定且已知的参考分布 ξ_0，其作用类似于截距。我们假设回归算子 $\Gamma(\boldsymbol{\xi})$ 可以表示为经过适当变换的预测变量分布的加权 Wasserstein Fréchet均值：

$$\Gamma(\boldsymbol{\xi}) = \underset{b \in \mathcal{W}_2(\mathcal{S})}{\arg\min} \sum_{j=0}^{p} \alpha_j d^2_{\mathcal{W}}(b, T_j\#\xi_j), \quad \boldsymbol{\xi} \in (\mathcal{W}_2(\mathcal{S}))^p, \tag{1}$$

其中，$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots\ldots, \alpha_p) \in \Delta^p$ 是一个非负权重向量，且权重之和为一，$\boldsymbol{T} = (T_0, T_1, \ldots\ldots, T_p) \in \mathcal{T}^{p+1}$，而 $\mathcal{T}$表示 $\mathcal{S}$ 上所有可容许的最优传输映射的集合，

$$\mathcal{T} := \{T \in \mathcal{S} \to \mathcal{S} : T \text{ is increasing}, T(s_0) = s_0, T(s_1) = s_1\}.$$

在所提出的模型中，每个 T_j 描述预测变量分布ξ_j的一种形状变形，而标量权重 α_j 决定其相对贡献。这两个组成部分共同构成预测变量分布 ξ_j的“回归系数”。为锚定模型，引入一个固定且已知的参考分布 ξ_0 ；只要它相对于 $\mathcal{S}$ 上的勒贝格测度绝对连续，其选择可以是任意的。尽管选择不同的 ξ_0 会改变相应的最优传输映射T_0，但诱导项 $T_0\#\xi_0$ 保持不变。因此，参考分布发挥着类似截距的作用，同时使整体回归算子保持不变。在实践中，参考分布的常见选择包括$\mathcal{S}$ 上的均匀分布或响应分布的弗雷歇均值。

接下来，我们定义回归模型

$$\eta = T_\varepsilon \# \Gamma^*(\boldsymbol{\xi}), \quad \Gamma^*(\boldsymbol{\xi}) = \underset{b \in \mathcal{W}_2(\mathcal{S})}{\arg\min} \sum_{j=0}^{p} \alpha_j^* d_{\mathcal{W}}^2(b, T_j^* \# \xi_j) \tag{2}$$

其中， $\boldsymbol{\alpha}^* \in \Delta^p$ 和 $\boldsymbol{T}^* \in \mathcal{T}^{p+1}$为真实参数。算子 T_ε 充当扰动算子，并满足逐点期望条件 $\mathbb{E}(T_\varepsilon \mid \boldsymbol{\xi}) = \mathrm{id}_{\mathcal{S}}$，其中 $\mathrm{id}_{\mathcal{S}} : \mathcal{S} \to \mathcal{S}$ 表示恒等映射。该条件类似于经典假设，即噪声误差的期望为零。

备注 1。 （2）中的结构可视为多元线性回归的类比

$$Y = \beta_0 \cdot 1 + \sum_{j=1}^{p} \beta_j X_j + \epsilon,$$

对于（X_1，…， X_p， Y） $\in \mathbb{R}^p \times \mathbb{R}$. 每个回归系数 β_j都可以分解为方向项 $\beta_j/|\beta_j|$ 和大小项 $|\beta_j|$。在我们的框架中，这两个作用分别由传输映射 T_j 和权重 α_j承担。另一个类比之处在于，在线性模型中作为截距出现的常数 1，原则上可以替换为任意固定常数，相应地调整系数，但拟合得到的回归函数保持不变。

3.1 可识别性

任何回归框架的一项基本要求是，其总体参数必须可识别：不同的参数值应产生不同的回归算子。现在我们将证明，在较弱的正则性条件下，所提出的模型是可识别的。我们首先明确刻画回归算子。

命题1。 对于任意 $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)$，回归算子 $\Gamma(\boldsymbol{\xi})$ 定义于 (1) 由下式给出

$$F^{-1}_{\Gamma(\boldsymbol{\xi})} = \sum_{j=0}^{p} \alpha_j T_j \circ F^{-1}_{\xi_j}.$$

假设两组参数 $(\boldsymbol{\alpha},\ \boldsymbol{T})$ 和 $(\boldsymbol{\alpha}',\ \boldsymbol{T}')$ 中的所有权重均严格为正，并且对于从乘积分布中抽取的所有可容许预测变量，两组参数产生相同的回归算子。预测变量分布的相互独立结构仅用于确立参数的可识别性，并非模型构建或估计所必需。也就是说，对于所有 $(\xi_1, \xi_2, \ldots, \xi_p) \sim P_{\xi_1} \otimes P_{\xi_2} \otimes \ldots \otimes P_{\xi_p}$，其中 P_{ξ_j}表示在 ξ_j 上诱导的边缘分布，

$$\sum_{j=0}^{p} (\alpha_j T_j - \alpha_j' T_j') \circ F^{-1}_{\xi_j} = 0. \tag{3}$$

为保证唯一性，我们施加以下温和假设。

假设1 (非退化性)。 对于几乎所有的 $q \in (0, 1)$，集合 $R_j^q := \{F^{-1}_{\xi_j}(q) :\ \xi_j \sim P_{\xi_j}\}$ 对于每个 $j = 1, 2, \ldots, p$ 都包含多个点。

假设2 （正密度）。 参考分布 ξ_0 和均值测度$\mathbb{L}_{P_{\xi_j}} \xi_j$ 均具有在$\mathcal{S}$上几乎处处严格为正的密度函数。

假设3 (传输映射的连续性)。 每个传输映射 $T_j \in \mathcal{T}$ 在$\mathcal{S}$上连续， $j = 0,\ 1,\ \ldots\ldots,\ p$。

假设1 确保每个预测变量分布在不同样本间具有足够的变异性，因此其分位数函数在任何固定水平上都不会退化。假设2保证相关测度具有绝对连续性和正性，这意味着分位数集的像覆盖定义域$\mathcal{S}$。假设3 排除了不连续传输映射；在一维情形下，绝对连续测度之间的最优传输映射是单调的，因此几乎处处连续（在温和的正则性条件下，在$\mathcal{S}$ 上连续）。

在这些条件下，两个回归算子相等意味着每个差值$\alpha_j T_j - \alpha_j^{'} T_j^{'}$ 在$\mathcal{S}$上都是常数。由于$\mathcal{T}$中的所有传输映射都具有固定的边界条件，这样的常数只能为零，因此得到：

$$\alpha_j = \alpha_j^{'}, T_j - T_j' = 0, j = 1, 2, \ldots, p.$$

因此，该模型是可识别的。

定理1 （可识别性）。 在假设 *1*、*2*和 *3*下，参数对 $(\boldsymbol{\alpha}, \boldsymbol{T})$ 在(2) 中可由数据分布唯一确定。

3.2 估计

如前所述，回归算子中 $\boldsymbol{\xi}$ 的变化由权重 $\boldsymbol{\alpha}$ 和传输映射 $\boldsymbol{T}$共同决定。为便于记号并简化渐近性质的分析，我们引入以下参数空间：

$$\Theta = \{\boldsymbol{\alpha} \odot \boldsymbol{T} : \boldsymbol{\alpha} \in \Delta^p, \boldsymbol{T} \in \mathcal{T}^{p+1}\} \subset (L^\infty(\mathcal{S}))^{p+1},$$

其中， $\odot$ 表示逐点乘积。每个 $\boldsymbol{\theta} \in \Theta$ 都唯一确定回归算子 $\Gamma_{\boldsymbol{\theta}}$。

为估计 $\boldsymbol{\theta}$，我们采用基于最小二乘型 Wasser-

stein 损失的 M估计框架。对于任意 $(\boldsymbol{\xi},\ \eta) \in \text{supp}(P)$，定义损失

$$m_{\boldsymbol{\theta}}(\boldsymbol{\xi},\eta) = d_{\mathcal{W}}^2(\eta, \Gamma_{\boldsymbol{\theta}}(\boldsymbol{\xi})) = \left\| F_{\eta}^{-1} - \sum_{j=0}^{p} \alpha_j T_j \circ F_{\xi_j}^{-1} \right\|_{L^2}^2. \tag{4}$$

令 $M(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{\xi},\eta)\sim P}\, m_{\boldsymbol{\theta}}(\boldsymbol{\xi},\ \eta)$ 表示总体风险，并令

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_{\boldsymbol{\theta}}(\boldsymbol{\xi}_i, \eta_i) \tag{5}$$

d表示基于 n 个独立同分布样本 $\{(\boldsymbol{\xi}_i,\ \eta_i)\}_{i=1}^n$ 得到的经验对应量。

在实践中，通常无法直接观测到真实分布 $(\boldsymbol{\xi},\ \eta)$。相反，通常会从每个分布中观测独立同分布样本，并基于大小为 m 的样本构造经验估计量 $\boldsymbol{\xi}^m$ 和 η^m。正如 Ghodrati & Panaretos (2022) 所述，若每个传输映射 $T \in \mathcal{T}$均为 B-Lipschitz连续，则 Wasserstein误差在推前算子下保持稳定，即

$$d_{\mathcal{W}}^2(T\#\xi_j, T\#\xi_j^m) \lesssim d_{\mathcal{W}}^2(\xi_j, \xi_j^m),$$

因此，代入损失函数 $m_{\boldsymbol{\theta}}(\boldsymbol{\xi}^m,\ \eta^m)$ 继承了经验分布的收敛速率。定义 $M_n^m(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n m_{\boldsymbol{\theta}}(\boldsymbol{\xi}_i^m,\ \eta_i^m)$，我们有以下命题。

P命题 2. 设 $(\xi_1^m, \ldots, \xi_p^m, \eta^m)$ 为满足以下条件的经验估计量

$$d_{\mathcal{W}}(\xi_j^m, \xi_j) = \mathcal{O}_{\mathbb{P}}(r_m), d_{\mathcal{W}}(\eta^m, \eta) = \mathcal{O}_{\mathbb{P}}(r_m),$$

对某个趋于0的序列 $r_m \searrow$。若每个 $T_j \in \boldsymbol{T}$ 都是 *B-Lipschitz*连续的，则

$$\sup_{\boldsymbol{\theta}\in\Theta} |M_n^m(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta})| = \mathcal{O}_{\mathbb{P}}(r_m).$$

备注 2。 速率的确切形式 r_m 取决于经验估计量的选择以及对底层分布所施加的正则性假设。在适当条件下，可以利用文献中的现有方法，为结构化分布类建立极小极大最优速率（Petersen &Müller 2016，Panaretos &Zemel2016）。

ˆ 我们考虑一个估计量 $\boldsymbol{\theta}$，其对于 $\boldsymbol{\theta}^*$ 满足

$$M_n(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) = \mathcal{O}_{\mathbb{P}}(e_n) \tag{6}$$

对于某个确定的 $e_n \searrow 0$。命题 2 表明，M_n^m与M_n 之间的差异渐近可忽略。因此，当估计量由经验分布得到时，同一近似最优性条件（6）仍然成立，即

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} M_n^m(\boldsymbol{\theta}),$$

并且 $e_n = r_{m_n} \searrow 0$，当 $n \to \infty$ 时。因此，在理想化情形下确立的渐近性质可直接推广到具有实际意义的情形，即根据样本估计分布。

4 统计分析

本节研究通过（6）构造的估计量的渐近性质。我们的分析首先引入一种预测半范数，它能够自然地刻画回归算子的几何结构，并使总体风险得以进行二次展开。对于 $\boldsymbol{f}$，$\boldsymbol{g} \in (L^\infty(\mathcal{S}))^{p+1}$，定义

$$\langle \boldsymbol{f}, \boldsymbol{g}\rangle_P = \mathbb{E}_{\boldsymbol{\xi}\sim P_{\boldsymbol{\xi}}} \left\{ \left\langle \sum_{j=0}^{p} f_j \circ F_{\xi_j}^{-1}, \sum_{j=0}^{p} g_j \circ F_{\xi_j}^{-1} \right\rangle_{L^2} \right\} \quad \text{and} \quad \|\boldsymbol{f}\|_P = \langle \boldsymbol{f}, \boldsymbol{f}\rangle_P^{1/2},$$

其中，$\langle\cdot,\cdot\rangle_{L^2}$表示 L^2(0,1) 上的通常内积。上述构造在 $(L^\infty(\mathcal{S}))^{p+1}$ 上定义了一个半内积和一个半范数；形式化验证见补充材料 S3.4。直观而言，预测半范数 $\|\boldsymbol{f}\|_P$ 衡量由 $\boldsymbol{f}$生成的预测映射之间的 L^2 差异，并对预测变量的分布取平均。

预测半范数在参数空间 Θ 上诱导出一种自然度量。定义参数空间的直径为 $\operatorname{diam}\Theta = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_P$。由于所有输出

回归算子都属于 $\mathcal{W}_2(\mathcal{S})$，且其支集包含于 $[s_0, s_1]$，因此可得全局界 $(\operatorname{diam}\Theta)^2 \le (s_1 - s_0)^2$. 。该全局界刻画了参数空间 Θ 的整体尺度。

下一个引理表明，在预测半范数下，总体风险在 $\boldsymbol{\theta}^*$ 周围具有全局二次形式，这起到了强识别条件的作用。

引理 1（二次增长）。 对于每个 $\boldsymbol{\theta} \in \Theta$,

$$M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}^*) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_P^2.$$

这一恒等式揭示，预测半范数与总体风险的曲率一致，从而为分析估计量的收敛行为提供了一种内在度量。基于这一确定性结构，接下来我们施加一个连续性条件，以控制经验过程在 Θ 的局部邻域上的随机波动。

假设4（连续性）。 存在一个常数 $\Lambda > 0$，使得对于所有 $\boldsymbol{f} \in$（$\Theta - \Theta$）以及所有（$\boldsymbol{\xi}$, η）$\in$ supp（P），

$$\Big\| \sum_{j=0}^{p} f_j \circ F_{\xi_j}^{-1} \Big\|_{L^2} \le \Lambda \|\boldsymbol{f}\|_P,$$

其中（$\Theta - \Theta$）：$=\{\boldsymbol{f} - \boldsymbol{g}:\ \boldsymbol{f},\ \boldsymbol{g} \in \Theta\} \subset$（BV（$\mathcal{S}$））$^{p+1}$，而 BV（$\mathcal{S}$）表示 $\mathcal{S}$ 上有界变差的实值函数空间，即 $\int_{s_0}^{s_1} |\mathrm{d}f\ (s)\ | < \infty$。

注释 3。 假设 4 用于控制 Θ 的局部熵数。在单分量情形下，若要求随机测度ξ 相对于 $\mathbb{E}\xi$ 具有有界密度，则可推出这一连续性条件。此要求与Ghodrati& Panaretos(2022)中考虑的正则性条件密切相关。

定理 2. 在假设 *4*下，估计量 $\hat{}\boldsymbol{\theta}$ 在 (6) 满足

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_P = \mathcal{O}_{\mathbb{P}}(e_n^{1/2} + n^{-1/3}).$$

证明结合了引理 1 中的二次增长性质与补充材料 S3.7 中建立的局部经验过程界。

备注 4。 定理 2 意味着在预测半范数下具有预测一致性，即ˆ在以下意义下 $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_P \to 0$ 依概率成立，前提是 $e_n \to 0$。值得注意的是，该收敛速率是相对于预测半范数建立的，它直接量化了预测准确性，而非逐分量参数估计。权重向量 $\boldsymbol{\alpha}$ 和传输映射 $\boldsymbol{T}$ 的逐分量一致性可在参数化满足额外正则性条件时获得，但这并非预测一致性所必需，因此本文不再对此进行研究。

$n^{-1/3}$ 项反映了Wasserstein空间中由传输映射诱导的内在非欧几里得几何。具体而言，立方根收敛速率由单调传输映射类的熵所驱动，而权重的单纯形约束仅贡献有限维复杂度，并不影响主导阶。当使用基于大小为 $m = m_n$的样本所得到的经验分布构造回归算子时，额外的近似误差通过 $e_n = r_{m_n}$引入，从而由 n 与 m的相对增长速率决定相变。这揭示了两阶段误差结构：由 $n^{-1/3}$ 控制的统计误差与估计输入分布所产生的抽样引起的近似误差相互竞争。具体而言，当 m_n 增长得足够快，使得 $e_n^{1/2} = o(n^{-1/3})$时，统计误差占主导，估计量达到内在的 $n^{-1/3}$ 速率；而当 m_n 增长较慢时，抽样引起的误差将主导整体性能。

5 实现与计算细节

在建立理论框架后，接下来我们将介绍实现所提出方法所需的计算过程。具体而言，我们使用以下方法计算 $\boldsymbol{\theta}$ 的估计量

采用块坐标下降方案，在更新权重向量 $\boldsymbol{\alpha}$ 和传输映射集合 $\boldsymbol{T}$ 之间交替进行。

在估计 $\boldsymbol{T}$时，我们采用回归拟合式方案：每次更新 T_k 时，都将其余分量 $\{T_j, j \neq k\}$ 视为固定值。一个核心难点在于，如何在 Wasserstein空间中近似并优化弗雷歇平方和。为便于计算，我们用用户定义的划分上的黎曼和来近似 L^2 损失（5），其中该划分定义在 $\mathcal{S}$上。通过变量替换，目标函数可改写为关于最优传输映射的形式，随后离散化为

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{t}\left|T_{\xi_{ki}\to\eta_i}(x_r)-\alpha_k T_k(x_r)-\sum_{j\neq k}\alpha_j T_j\circ T_{\xi_{ji}\to\xi_{ki}}(x_r)\right|^2\xi_{ki}(h_r),$$

其中， t 表示区间划分$\{x_r\}_{r=1}^{t}$ 中的用户定义节点数量，该划分定义在$\{I_r\}_{r=1}^{t}$上，且$\mathcal{S}$，$h_r=|I_r|$。这些离散近似将无限维优化问题转化为节点上的有限和，从而实现高效的数值实现。随后，我们采用迭代回归拟合过程。在第 ℓ次迭代中，当前估计值 $\boldsymbol{\alpha}^{(\ell-1)}$ 位于单纯形 Δ^p 内部，且其余所有分量$\{T_j^{(\ell)}:j<k\}$ 和$\{T_j^{(\ell-1)}:\ j>k\}$ 均固定，此时通过求解以下问题来更新第 k个传输映射

$$\begin{aligned}T_k^{(\ell)}=\underset{T_k\in\mathcal{T}}{\arg\min}\,\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{t}\Bigg|&T_{\xi_{ki}\to\eta_i}(x_r)-\sum_{j<k}\alpha_j^{(\ell-1)}T_j^{(\ell)}\circ T_{\xi_{ji}\to\xi_{ki}}(x_r)\\&-\alpha_k^{(\ell-1)}T_k(x_r)-\sum_{j>k}\alpha_j^{(\ell-1)}T_j^{(\ell-1)}\circ T_{\xi_{ji}\to\xi_{ki}}(x_r)\Bigg|^2\xi_{ki}(h_r).\end{aligned}\tag{7}$$

通过定义 $w_{ir}=\xi_{ki}(h_r)$、 $z_r=T_k(x_r)$ 和

$$y_{ir}=\left(T_{\xi_{ki}\to\eta_i}(x_r)-\sum_{j<k}\alpha_j^{(\ell-1)}T_j^{(\ell)}\circ T_{\xi_{ji}\to\xi_{ki}}(x_r)-\sum_{j>k}\alpha_j^{(\ell-1)}T_j^{(\ell-1)}\circ T_{\xi_{ji}\to\xi_{ki}}(x_r)\right)/\alpha_k^{(\ell-1)},$$

近似最小化（7）可归结为以下问题:

$$\min f(z)=\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{t}\left(w_{ir}\left|y_{ir}-z_r\right|^2\right)$$

$$\text{subject to } s_0=z_1<z_2<\cdots<z_t=s_1.$$

令 $\hat{y}_r = \frac{\sum_{i=1}^{n} w_{ir} y_{ir}}{\sum_{i=1}^{n} w_{ir}}$ 和 $\tilde{w}_r = \frac{1}{n}\sum_{i=1}^{n} w_{ir}$，我们可以改写为

$$\frac{1}{n}\sum_{i=1}^{n} w_{ir}\left(y_{ir} - z_r\right)^2 = \frac{1}{n}\sum_{i=1}^{n} w_{ir}\left(y_{ir} - \hat{y}_r\right)^2 + \tilde{w}_r\left(z_r - \hat{y}_r\right)^2.$$

因此，该问题等价于加权保序回归：

$$\begin{aligned} &\min f(z) = \sum_{r=1}^{t} \tilde{w}_r\left(z_r - \hat{y}_r\right)^2 \\ &\text{subject to } s_0 = z_1 < z_2 < \cdots < z_t = s_1. \end{aligned} \tag{8}$$

因此，所有参数均通过交替迭代过程进行估计，该过程总结于算法 1 中，并在以下步骤间交替进行：**(i)** 通过回代步骤更新每个传输映射 T_k，该步骤求解一个加权保序回归问题；**(ii)** 通过在单纯形上求解约束最小二乘问题来更新权重向量 $\boldsymbol{\alpha}$ 。

用矩阵形式表示，令 $\mathbf{z} = (z_1, \ldots, z_t)^\top$、$\mathbf{y} = (\hat{y}_1, \ldots, \hat{y}_t)^\top$，以及 $W = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_t)$。问题（8）转化为带线性不等式约束的标准二次规划：

$$\min_{\mathbf{z}\in\mathbb{R}^t} (\mathbf{z} - \mathbf{y})^\top W (\mathbf{z} - \mathbf{y}), \text{ s.t. } s_0 = z_1 < z_2 < \cdots < z_t = s_1.$$

该过程将原始函数优化问题转化为一系列有限维二次规划，可使用标准凸优化工具高效求解。

6 模拟

我们开展了一系列广泛的模拟研究，以评估所提出模型在单个及多个预测变量设置下的表现。此外，我们还将其表现与最优传输模型（Ghodrati & Panaretos 2022）和GOT模型（Zhu &Müller 2025）进行比较。

算法1 所提出模型的交替迭代过程 输入：$\{\boldsymbol{\xi}_i, \eta_i\}$ $i = 1, \ldots, n$, 参考分布 ξ_0, 网格大小 t, 初始映射 $T_j^{(0)}$ =id$_{\mathcal{S}}$, $j = 0, 1, \ldots, p$.1: $\boldsymbol{\alpha}^{(0)} =$ arg min$_{\boldsymbol{\alpha}\in\Delta^p}$ $\frac{1}{n}\sum_{i=1}^n \left\| F_{\eta_i}^{-1} - \sum_{j=0}^p \alpha_j T_j^{(0)} \circ F_{\xi_{ji}}^{-1} \right\|_{L^2}^2$.2: 对于 $\ell = 1, 2, \ldots$ do 3: 对于 $k = 0, 1, \ldots, p$ do 4: 通过求解子问题（7）更新 $T(\ell)_k$，并固定 $\{T(\ell)_j : j < k\}$ 和 $\{T_j^{(\ell-1)} : j > k\}$。 5: end for 6: $\boldsymbol{\alpha}^{(\ell)} =$ arg min $_{\boldsymbol{\alpha}\in\Delta^p}$ $\frac{1}{n}\sum_{i=1}^n \left\| F_{\eta_i}^{-1} - \sum_{j=0}^p \alpha_j T_j^{(\ell)} \circ F_{\xi_{ji}}^{-1} \right\|_{L^2}^2$.7: 如果收敛，则 8: 停止，并记 $\{\widehat{\alpha}_j, \widehat{T}_j\}_{j=0}$, 1,...,$p$ $= \{\alpha_j^{(\ell)}, T(\ell)_j\}_{j=0}$,1,...,$p$; 9: end if 10: end for$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$输出： $\widehat{\boldsymbol{\alpha}}, \boldsymbol{T}, \widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\alpha}} \odot \boldsymbol{T}$.

–

6.1 单预测变量

尽管我们的模型是针对多个预测变量（$p \geq 1$）开发的，但情形 $p = 1$ 是一个值得关注的特殊情况。因此，我们首先展示该单预测变量设定下的结果，以此开启模拟研究。

我们选择 $\mathcal{S} = [0, 1]$ 上的均匀分布作为参考概率测度ξ_0。对于随机预测变量分布 $\{\xi_{1i}\}_{i=1}^n$，我们考虑参数为区间 $[1, 5]$，上均匀分布随机变量的贝塔分布，即

$$f_{\xi_{1i}}(x) = b_{a_i,b_i}(x), \quad a_i \sim \mathrm{U}[1, 5], \quad b_i \sim \mathrm{U}[1, 5], \quad i = 1, \cdots, n.$$

对于噪声映射 T_{ϵ_i}，我们采用由Panaretos

& Zemel（2016）提出的随机最优映射类。设 k 为一个整数；我们定义映射 g_k： $[0,1] \to [0,1]$ 如下：

$$g_0(x) = x, \quad g_k(x) = x - \frac{\sin(\pi k x)}{|k|\pi}, \quad k \in \mathbb{Z} \setminus \{0\}. \tag{9}$$

这些函数严格递增且光滑，满足 $g_k(0) = 0$ 且 $g_k(1) = 1$ ，对任意k均成立。将 k 替换为关于零对称、服从某一分布的整数值随机变量 K 后，可得对所有 $x \in [0,1]$， $E[g_K(x)] = x$，这符合模型的要求。

关于 构成回归算子的最优映射 T_0^*和 T_1^*，我们设定

$$T_0^* = g_4, \quad T_1^* = g_3.$$

生成随机的 ξ_{1i} 和 T_{ϵ_i}后，可根据 $\alpha_1^* \in$的不同取值$\{0, 0.25, 0.5, 0.75, 1\}$得到响应分布，即 $\eta_i = T_{\epsilon_i} \# \Gamma(\xi_{1i})$，

$$F_{\eta_i}^{-1}(\cdot) = g_{K_i} \circ \left\{ (1-\alpha_1^*) T_0^* \circ F_{\xi_0}^{-1}(\cdot) + \alpha_1^* T_1^* \circ F_{\xi_{1i}}^{-1}(\cdot) \right\}. \tag{10}$$

由于实践中无法直接观测到这些分位数函数 $F_{\eta_i}^{-1}, F_{\xi_{1i}}^{-1}$，我们假设有以下随机样本观测值 $\{F_{\xi_{1i}}^{-1}(u_{i1}) \quad F_{\xi_{1i}}^{-1}(u_{i2}) \ \ldots, F_{\xi_{1i}}^{-1}(u_{im})\}$和 $\{F_{\eta_i}^{-1}(v_{i1}) \quad F_{\eta_i}^{-1}(v_{i2}) \ \ldots, \ F_{\eta_i}^{-1}(v_{im})\}$，其中 u_i, v_i是从分布 $\mathsf{U}(0,1)$ 中独立生成的。根据观测值，在 $q \in [0,1]$的网格上，利用经验分位数估计分位数函数 $F_{\xi_{1i}}^{-1}(q)$ 和 $F_{\eta_i}^{-1}(q)$。

下表展示了基于 m随机样本观测点和 n 组随机概率分布在不同参数设定下的估计结果。我们根据 $\|\widehat{\boldsymbol{\theta}}^j - \boldsymbol{\theta}^*\|_P$、 $\left|\widehat{\alpha}_1^j - \alpha_1^*\right|$、 $\|\widehat{T}_0^j - T_0^*\|_{L_2}$和 $\|\widehat{T}_1^j - T_1^*\|_{L_2}$来评估性能。在此， $\widehat{\boldsymbol{\theta}}^j$、 $\widehat{\alpha}_1^j$、 $\widehat{T}_0^j$、 $\widehat{T}_1^j$是第 j个重复数据集的估计值。表1 报告了^均值和标准差结果，其中 $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_P$是在样本量为 $0.3n$的测试集上计算得到的。

备注5：需要注意的是，当 $\alpha_1^* = 0$ 或为 1 时，相应的目标分布完全由其中一个预测变量分布决定。因此， T_1 不可识别

表 1：单预测变量设定下基于 100次重复 的蒙特卡洛均值（标准差）。

α_1^*	n	m	$\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_P$	$\lvert\widehat{\alpha}_1-\alpha_1^*\rvert$	$\left\|\widehat{T}_0-T_0^*\right\|_{L^2}$	$\left\|\widehat{T}_1-T_1^*\right\|_{L^2}$
0	50	200	0.017(0.005)	0.036(0.023)	0.016(0.006)	＼
		400	0.014(0.005)	0.035(0.020)	0.012(0.006)	＼
	200	200	0.012(0.002)	0.030(0.013)	0.012(0.003)	＼
		400	0.008(0.002)	0.018(0.010)	0.008(0.003)	＼
	400	200	0.011(0.001)	0.025(0.009)	0.011(0.002)	＼
		400	0.007(0.001)	0.017(0.008)	0.007(0.002)	＼
0.25	50	200	0.018(0.005)	0.028(0.020)	0.021(0.009)	0.056(0.025)
		400	0.015(0.005)	0.024(0.021)	0.016(0.007)	0.043(0.017)
	200	200	0.011(0.002)	0.017(0.012)	0.016(0.005)	0.035(0.011)
		400	0.008(0.002)	0.012(0.009)	0.011(0.005)	0.029(0.011)
	400	200	0.009(0.001)	0.012(0.008)	0.015(0.004)	0.033(0.011)
		400	0.006(0.001)	0.010(0.008)	0.009(0.003)	0.023(0.009)
0.50	50	200	0.017(0.005)	0.027(0.020)	0.023(0.008)	0.026(0.010)
		400	0.015(0.005)	0.023(0.022)	0.021(0.010)	0.024(0.009)
	200	200	0.009(0.002)	0.015(0.011)	0.014(0.003)	0.014(0.004)
		400	0.008(0.002)	0.014(0.011)	0.012(0.004)	0.012(0.004)
	400	200	0.007(0.001)	0.010(0.007)	0.012(0.003)	0.010(0.003)
		400	0.006(0.001)	0.010(0.009)	0.009(0.002)	0.009(0.003)
0.75	50	200	0.016(0.005)	0.038(0.026)	0.060(0.025)	0.025(0.011)
		400	0.014(0.005)	0.027(0.022)	0.049(0.024)	0.020(0.008)
	200	200	0.009(0.002)	0.029(0.018)	0.062(0.033)	0.024(0.013)
		400	0.008(0.002)	0.023(0.014)	0.042(0.019)	0.016(0.008)
	400	200	0.007(0.001)	0.026(0.018)	0.058(0.030)	0.021(0.011)
		400	0.006(0.002)	0.019(0.012)	0.041(0.019)	0.015(0.007)
1	50	200	0.015(0.005)	0.062(0.041)	＼	0.028(0.015)
		400	0.013(0.005)	0.034(0.032)	＼	0.017(0.010)
	200	200	0.009(0.003)	0.051(0.033)	＼	0.025(0.015)
		400	0.008(0.002)	0.036(0.025)	＼	0.016(0.009)
	400	200	0.007(0.002)	0.048(0.031)	＼	0.023(0.014)
		400	0.006(0.002)	0.029(0.023)	＼	0.014(0.009)

当 $\alpha_1^* = 0$ （同样地，当 $\alpha_1^* = 1$ 时， T_0 不可识别）。因此，表中未报告相应的估计误差。

根据上述讨论，我们的模型可视为 OT 模型（Ghodrati& Panaretos 2022）的扩展。具体而言，在单预测变量设定中，我们的模型纳入了一个额外的参考分布，由此形成我们称为多传输分布回归（MTDR）模型的模型。为评估其性能，我们在相同的数据生成设定（10）下，比较了 MTDR 模型和 OT 模型的均方根误差（$\sqrt{\frac{1}{0.3n}\sum_{k=n+1}^{1.3n} d_{\mathcal{W}}^2(\eta_k, \widehat{\eta}_k)}$，该误差在样本量为$0.3n$的测试集上计算），结果汇总于表2的前两列中。当 $\alpha_1^* = 1$时，MTDR 模型简化为 OT 模型

$$\eta_i = T_{\epsilon_i} \# \left[T_1 \# \xi_{1i} \right], \quad \{\xi_{1i}, \eta_i\}_{i=1}^n.$$

如表2的前两列所示，当 $\alpha_1^* = 1$时，两者得到相同的结果，这与预期一致。随着 α_1^* 减小（这意味着赋予参考分布更大的权重）， ξ_1 对 η 的回归效应会减弱。在这种情况下，缺乏捕捉此类回归结构能力的最优传输模型，其均方根误差值明显高于MTDR模型，这凸显了纳入参考分布所带来的显著性能提升。

此外，我们还将结果与Zhu 和 Müller （2025）提出的GOT模型进行了比较分析。具体而言，单预测变量设定下的GOT模型可表示为

$$\eta_i = \left[T_{\epsilon_i} \circ (\beta \odot T_{\bar{\xi} \to \xi_i}) \right] \# \bar{\eta}, \quad \{\xi_i, \eta_i\}_{i=1}^n,$$

或者，等价地，从分位数函数的角度来看，

$$F_{\eta_i}^{-1}(\cdot) = \left[T_{\epsilon_i} \circ (\beta \odot T_{\bar{\xi} \to \xi_i}) \right] \circ F_{\bar{\eta}}^{-1}(\cdot),$$

其中 $\bar{\xi}$ 和 $\bar{\eta}$d 分别表示 ξ 和 η 的弗雷歇均值； $\beta \in \mathbb{R}$ 是真实模型

表2：不同 α_1^* 值下 MTDR模型设定（10）（$n = m = 200$）的均方根误差值比较。每行中最小的均方根误差值以粗体突出显示。

α_1^*	$RMSE_{MTDR}$	$RMSE_{OT}$	$RMSE_{GOT}$
1	**0.076(0.011)**	**0.076(0.011)**	0.103(0.009)
0.7	**0.074(0.011)**	0.096(0.010)	0.089(0.009)
0.5	**0.073(0.010)**	0.125(0.009)	0.082(0.009)
0.3	**0.072(0.010)**	0.158(0.010)	0.076(0.009)
0	**0.070(0.009)**	0.213(0.011)	**0.070(0.009)**

¯参数； $T_{\bar{\xi}\to\xi_i}$ 是从 ξ 到 ξ_i的最优传输映射；并且$\{T_{\epsilon_i}\}_{i=1}^N$ 是随机扰动映射。操作 $\beta \odot T_{\bar{\xi}\to\xi_i}$的定义以及模型设定的更多细节，见Zhu 和 Müller (2025)。

在与（10）相同的数据生成设定下，GOT 的结果列于表2的第三列。值得注意的是，当 $\alpha_1^* = 0$时，MTDR模型退化为这样一种情形：响应分布独立于预测变量分布，仅通过扰动一个固定的参考分布生成。

$$\eta_i = T_{\epsilon_i} \# \left[T_0 \# \xi_0\right], \quad \{\xi_i, \eta_i\}_{i=1}^n.$$

这个特殊情形与纯截距模型一致，该模型在Ghodrati&Panaretos（2022）的注3中讨论，即

$$T_0 = T_{\xi_0 \to \bar{\eta}} = F_{\bar{\eta}}^{-1} \circ F_{\xi_0},$$

这也对应于 $\beta = 0$ 的 GOT 模型。因此，如表 2所示，当 $\alpha_1^* = 0$ 时，MTDR模型和GOT模型的均方根误差值相同。对于 α_1^* 和 β 的其他取值，在我们的模拟设定中，不同的数据生成机制均导致GOT模型的均方根误差值 consistently 高于MTDR模型。

补充材料S1提供了在GOT模型的数据生成设定下对三种方法的比较。

6.2 多预测变量

在下一节中，我们将推广到多预测变量设定，首先考虑 $p = 2$的情况。 ξ_0 和 T_{ϵ_i} 的选择与第6.1节相同，ξ_{1i}、 ξ_{2i}、 T_0^*、 T_1^*和 T_2^*的设定如下：

$$f_{\xi_{1i}}(x) = b_{a_{1i},b_{1i}}(x), a_{1i} \sim \mathrm{U}[1,5], b_{1i} \sim \mathrm{U}[1,5],$$

$$f_{\xi_{2i}}(x) = b_{a_{2i},b_{2i}}(x), a_{2i} \sim \mathrm{U}[2,6], b_{2i} \sim \mathrm{U}[2,6],$$

$$T_0^* = g_4, \quad T_1^* = g_3, \quad T_2^* = g_{-5}.$$

然后，可以根据 $\boldsymbol{\alpha}^* = (\alpha_0^*, \alpha_1^*, \alpha_2^*)^\top \in \{(0.3, 0.35, 0.35)^\top, (0.2, 0.2, 0.6)^\top, (0, 0.5, 0.5)^\top, (0, 1, 0)^\top\}$的不同取值获得响应分布，即 $\eta_i = T_{\epsilon_i} \# \Gamma(\xi_{1i}, \xi_{2i})$，

$$F_{\eta_i}^{-1}(\cdot) = g_{K_i} \circ \left\{ \alpha_0^* T_0^* \circ F_{\xi_0}^{-1}(\cdot) + \alpha_1^* T_1^* \circ F_{\xi_{1i}}^{-1}(\cdot) + \alpha_2^* T_2^* \circ F_{\xi_{2i}}^{-1}(\cdot) \right\}. \tag{11}$$

同样地，我们假设有以下随机样本观测值：$\{F_{\xi_{1i}}^{-1}(u_{i1}) \ F_{\xi_{1i}}^{-1}(u_{i2}) \ldots, F_{\xi_{1i}}^{-1}(u_{in})\}$、$\{F_{\xi_{2i}}^{-1}(v_{i1}) \ F_{\xi_{2i}}^{-1}(v_{i2}) \ldots, F_{\xi_{2i}}^{-1}(v_{in})\}$和$\{F_{\eta_i}^{-1}(w_{i1}) , F_{\eta_i}^{-1}(w_{i2}) \ldots, F_{\eta_i}^{-1}(w_{in})\}$，其中 u_i、 v_i、 w_i是从分布U(0, 1)中独立生成的。根据这些观测值，在由 q 个值 $\in [0,1]$构成的网格上，利用经验分位数估计分位数函数 $F_{\xi_{1i}}^{-1}(q)$、 $F_{\xi_{2i}}^{-1}(q)$和 $F_{\eta_i}^{-1}(q)$。

我们还呈现了基于 m 随机样本观测点以及 n 多组随机概率分布，在不同参数设定下的估计结果。我们评估了$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$以下指标的性能： $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_P$， $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_{\ell^2}$， $\|T_0 - T_0^*\|_{L_2}$， $\|T_1 - T_1^*\|_{L_2}$和 $\|T_2 - T_2^*\|_{L_2}$，如表3所示。

表3：多预测变量设置下基于100次重复的蒙特卡洛均值（标准差）。

$\boldsymbol{\alpha}^*$	n	m	$\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_P$	$\|\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*\|_{\ell^2}$	$\left\|\widehat{T}_0-T_0^*\right\|_{L^2}$	$\left\|\widehat{T}_1-T_1^*\right\|_{L^2}$	$\left\|\widehat{T}_2-T_2^*\right\|_{L^2}$
$\begin{pmatrix}0.3\\0.35\\0.35\end{pmatrix}$	50	200	0.020(0.006)	0.094(0.056)	0.057(0.032)	0.050(0.024)	0.060(0.038)
		400	0.019(0.005)	0.082(0.049)	0.047(0.023)	0.045(0.021)	0.049(0.026)
	200	200	0.010(0.002)	0.043(0.022)	0.028(0.011)	0.024(0.009)	0.026(0.009)
		400	0.009(0.002)	0.041(0.021)	0.026(0.010)	0.022(0.009)	0.021(0.008)
	400	200	0.008(0.002)	0.034(0.018)	0.022(0.010)	0.019(0.008)	0.021(0.010)
		400	0.007(0.002)	0.032(0.018)	0.018(0.008)	0.015(0.006)	0.015(0.005)
$\begin{pmatrix}0.2\\0.2\\0.6\end{pmatrix}$	50	200	0.020(0.005)	0.096(0.051)	0.084(0.047)	0.077(0.037)	0.038(0.013)
		400	0.019(0.005)	0.085(0.052)	0.069(0.035)	0.078(0.040)	0.029(0.013)
	200	200	0.011(0.002)	0.054(0.022)	0.053(0.026)	0.057(0.028)	0.031(0.009)
		400	0.009(0.002)	0.044(0.024)	0.041(0.016)	0.043(0.020)	0.020(0.007)
	400	200	0.008(0.002)	0.046(0.018)	0.051(0.022)	0.052(0.024)	0.032(0.009)
		400	0.007(0.002)	0.035(0.017)	0.033(0.016)	0.038(0.018)	0.019(0.007)
$\begin{pmatrix}0\\0.5\\0.5\end{pmatrix}$	50	200	0.020(0.005)	0.113(0.061)	\	0.050(0.027)	0.048(0.027)
		400	0.018(0.006)	0.089(0.055)	\	0.039(0.025)	0.035(0.019)
	200	200	0.010(0.002)	0.069(0.034)	\	0.034(0.017)	0.024(0.010)
		400	0.009(0.002)	0.052(0.027)	\	0.021(0.010)	0.016(0.007)
	400	200	0.008(0.002)	0.062(0.029)	\	0.031(0.017)	0.020(0.009)
		400	0.006(0.002)	0.040(0.023)	\	0.016(0.008)	0.012(0.006)
$\begin{pmatrix}0\\1\\0\end{pmatrix}$	50	200	0.017(0.005)	0.123(0.056)	\	0.040(0.019)	\
		400	0.015(0.005)	0.085(0.048)	\	0.026(0.014)	\
	200	200	0.009(0.002)	0.105(0.042)	\	0.040(0.017)	\
		400	0.008(0.002)	0.069(0.036)	\	0.024(0.013)	\
	400	200	0.008(0.002)	0.098(0.039)	\	0.039(0.016)	\
		400	0.006(0.002)	0.065(0.036)	\	0.023(0.014)	\

备注6。与备注5一致，当 $\alpha_k^* = 0$， $k \in \{0, 1, 2\}$时，相应的 T_k 无法识别，因为目标分布取决于另一个预测变量分布。因此，表中未报告 T_k 的估计误差。

在前述比较的基础上，我们现在转向多预测变量设定。由于最优传输模型无法直接应用于包含多个预测变量分布的情形，我们将比较范围限定为GOT模型。在这种情况下，假设存在一个未知的真实顺序 j_1^*， j_2^*，用于确定两个预测变量在回归模型中的顺序，即含有两个预测变量的GOT模型为

$$\eta_i = \left[T_{\epsilon_i} \circ (\beta_1 \odot T_{\bar{\xi}_{j_1^* \to \xi_{j_1^* i}}}) \circ (\beta_2 \odot T_{\bar{\xi}_{j_2^* \to \xi_{j_2^* i}}})\right] \# \bar{\eta}, \quad \{\xi_{1i}, \xi_{2i}, \eta_i\}_{i=1}^n,$$

或者等价地，从分位数函数的角度来看，

$$F_{\eta_i}^{-1}(\cdot) = \left[T_{\epsilon_i} \circ (\beta_1 \odot T_{\bar{\xi}_{j_1^* \to \xi_{j_1^* i}}}) \circ (\beta_2 \odot T_{\bar{\xi}_{j_2^* \to \xi_{j_2^* i}}})\right] \circ F_{\bar{\eta}}^{-1}(\cdot),$$

其中， $\bar{\xi}_1$、 $\bar{\xi}_2$ 和 $\bar{\eta}$ 分别表示 ξ_1、 ξ_2 和 η的弗雷歇均值；$(\beta_1, \beta_2)^\top \in \mathbb{R}^{2^-}$为模型参数；$T_{\bar{\xi}_j \to \xi_{ji}}$、 $j = 1, 2$为从 ξ_j 到 ξ_{ji}的最优传输映射；而$\{T_{\epsilon_i}\}_{i=1}^N$ 为随机扰动映射。

在与（11）相同的数据生成设定下，我们比较了不同 $\boldsymbol{\alpha}^*$取值时 MTDR 模型和 GOT 模型的均方根误差，结果汇总于表4。对于不同的 $\boldsymbol{\alpha}^*$取值，在我们的模拟设定中，由于模型结构各不相同，所提出的 MTDR 模型始终比 GOT 模型具有更小的 RMSE 值。这里不考虑 $\boldsymbol{\alpha}^* = (1,0,0)^\top$ 这一情形，因为它与表 $\alpha_1^* = 0$2所报告的情形完全相同。在 GOT 模型的数据生成设定下，两种方法的结果见补充材料S2。

表4：在 MTDR模型设置下（ $\boldsymbol{\alpha}^*$ 的不同取值的均方根误差值比较（11）（$n = m = 200$）。每行中较小的均方根误差值以粗体突出显示。

$\boldsymbol{\alpha}^*$	$RMSE_{MTDR}$	$RMSE_{GOT}$
(0。3，0。35，0。35)$^\top$	**0.074(0.013)**	0.080(0.012)
(0。2，0。2，0。6)$^\top$	**0.075(0.013)**	0.081(0.012)
(0，0。5，0。5)$^\top$	**0.076(0.014)**	0.086(0.012)
$(0,1,0)^\top$	**0.076(0.012)**	0.103(0.009)

7 数据应用

现在，我们转向一个真实数据示例，以展示模型的性能和实际优势。人类死亡率长期以来一直是人口学和老龄化研究领域的核心议题。本节使用人类死亡率数据库（HMD, www.mortality.org）的数据，该数据库提供了34个国家男性和女性的死亡年龄分布。我们的目标是利用同一国家2005年的女性和男性死亡年龄分布这两个预测变量，预测2010年男性的死亡年龄分布。采用相隔五年的数据也出于实际考虑：例如，在人寿保险经验研究中，死亡率通常按五年为间隔报告（Yue &Huang2011），研究人员也据此开发了相应模型。基于这一考虑，我们尝试采用以下形式对目标分布建模并进行预测，

$$\eta^m_{2010,i} = T_{\epsilon_i}\#\Gamma(\xi^f_{2005,i}, \xi^m_{2005,i}), \quad \{\xi^f_{2005,i}, \xi^m_{2005,i}, \eta^m_{2010,i}\}_{i=1}^{34},$$

其中

$$\Gamma(\xi^f_{2005,i}, \xi^m_{2005,i})) = \arg\min_{\omega} \Big\{ \alpha_0 d^2_{\mathcal{W}}(\omega, T_0\#\xi_0) + \alpha_1 d^2_{\mathcal{W}}(\omega, T_1\#\xi^f_{2005,i}) + \alpha_2 d^2_{\mathcal{W}}(\omega, T_2\#\xi^m_{2005,i}) \Big\}.$$

我们将 $\xi_0 = \bar{\eta}^m_{2010}$ （$\{\eta^m_{2010,i}\}$ 的弗雷歇均值）设为参考分布。不过，我们也测试了其他参考分布，结果相似。

作为比较，我们还考虑了GOT模型（Zhu& Müller2025），其形式如下：

$$\eta^m_{2010,i} = \left[T_{\epsilon_i} \circ (\beta_1 \odot T_{\bar{\xi}_{j_1^*} \to \xi_{j_1^*,i}}) \circ (\beta_2 \odot T_{\bar{\xi}_{j_2^*} \to \xi_{j_2^*,i}})\right] \# \bar{\eta}^m_{2010},$$

其中，(j_1^*, j_2^*) 是两个预测变量 ξ^m_{2005}和 ξ^f_{2005}的真实顺序； $\bar{\xi}^m_{2005}$、 $\bar{\xi}^f_{2005}$和 $\bar{\eta}^f_{2010}$分别表示 $\xi^m_{2005、i}$、$\xi^f_{2005、i}$ 和$\eta^m_{2005、i}$的弗雷歇均值。

我们进一步比较所提出的MTDR模型与GOT模型和最优传输模型（Ghodrati& Panaretos 2022），其中最优传输模型仅使用2005年男性分布作为预测变量。我们采用留一交叉验证方案：对于每个国家 i，使用其余33个国家的样本拟合各模型，然后预测 $\eta^m_{2010、i}$。使用平均Wasserstein距离（AWD）评估预测准确性：$= \frac{1}{34}\sum_{k=1}^{34} d_{\mathcal{W}}(\nu_k, \widehat{\nu}_k)$，结果见表5。

表5：使用MTDR、OT和GOT模型预测34个国家2010年男性死亡年龄分布的留一交叉验证误差（平均Wasserstein距离）。

MTDR模型		OT	GOT
AWD	0.530	0.618	0.559

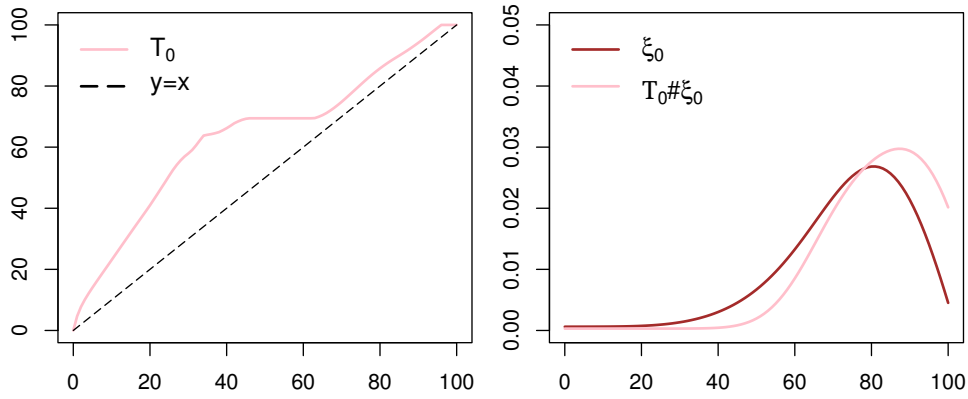
表 5 显示，在这一真实数据情境下，所提出的MTDR模型在三种方法中预测误差最低，优于最优传输和GOT。与OT模型相比，MTDR模型纳入了更多预测变量，因此其预测误差更低完全在意料之中。事实上，即使MTDR模型仅使用一个预测变量，即2005年的男性分布，其预测误差仍显著更低。这一改进很大程度上归因于参考

分布 ξ_0 在提升预测准确性方面发挥了不可忽视的作用。这一发现表明，纳入参考分布能够显著提升预测性能。

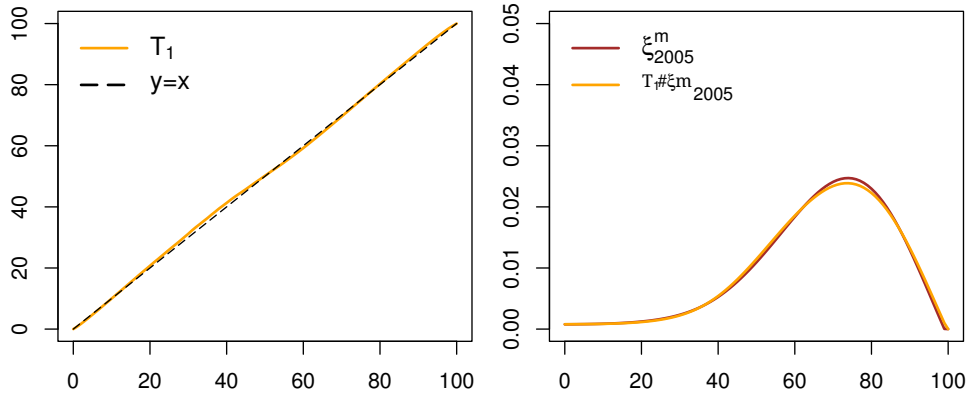
与GOT模型相比，我们的方法在预测误差方面仅取得了边际性的改进。然而，所估计的标量系数和映射系数仍然为理解潜在的回归结构提供了清晰且易于解释的洞见。遵循 Ghodrati & Panaretos （2022）提出的解释框架，这些映射系数量化了预测分布形状的变化；更形式化地说，它们刻画了预测分布的概率质量如何重新排列，而标量系数则可以视为量化每个预测变量分布对目标分布的相对贡献。

作为一个说明性示例，我们考虑保加利亚的情况。在留一法下，所估计的标量系数为 $\widehat{\alpha}_0 = 0.08$、 $\widehat{\alpha}_1 = 0.83$和 $\widehat{\alpha}_2 = 0.09$。相应的估计传输映射 T_0、 T_1 和 T_2 显示在图1的第一列中，而第二列展示了这些映射如何改变相应预测变量分布的形状。由于 ξ_0 是人为选择的参考分布，对 ξ_0 作出不同选择将得到不同的 T_0，因此我们直接关注复合映射 $T_0 \# \xi_0$。对于 T_1，如图1（b）的左侧面板所示，该映射非常接近恒等线 $y = x$，这意味着 $T_1 \# \xi^m_{2005}$ 几乎与 ξ^m_{2005}相同。

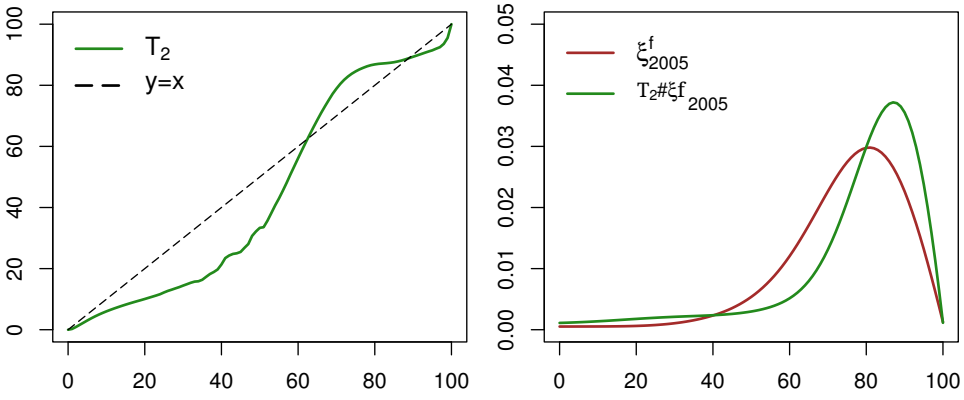
基于上述对图1中保加利亚的解读，我们进一步说明变换后的预测变量分布与拟合目标分布之间的关系。图2展示了根据估计系数得到的拟合分布 $\widehat{\eta}^m_{2010}$ 、观测目标分布 η^m_{2010}，以及三个变换后的预测变量分布 $T_0 \# \xi_0$、$T_1 \# \xi^m_{2005}$ 和 $T_2 \# \xi^f_{2005}$。拟合分布 $\widehat{\eta}^m_{2010}$ 对应于 $T_0 \# \xi_0$、 $T_1 \# \xi^m_{2005}$ 和 $T_2 \# \xi^f_{2005}$的加权Fréchet均值，其权重分别为对应的估计标量系数 $\widehat{\alpha}_0 = 0.08$、 $\widehat{\alpha}_1 = 0.83$和 $\widehat{\alpha}_2 = 0.09$。鉴于权重相对较大



(a) T_0



(b) T_1



(c) T_2

图1：在留一法下，以保加利亚为例，估计的传输映射（左列）及其对相应预测变量分布（右列）的影响。左列中的实线表示估计的映射 T_k,($k = 0$,1, 2)，虚线表示恒等映射 $y = x$。

在 $T_1\#\xi^m_{2005}$ 上，因此拟合分布更接近这个变换后的预测变量。

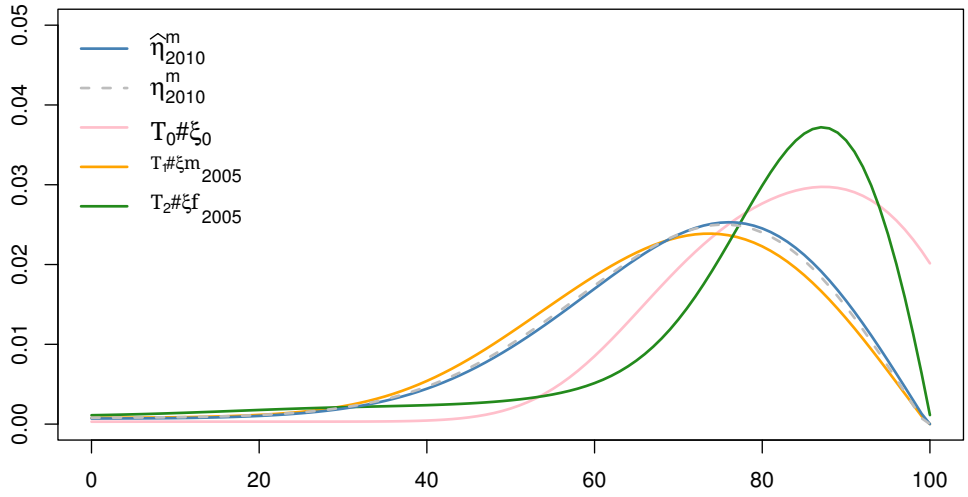


图2：对于保加利亚，拟合的目标分布 $\hat{\eta}^m_{2010}$ （蓝色）、观测目标分布 η^m_{2010} （灰色虚线）以及变换后的预测变量分布 $T_0\#\xi_0$ （粉色）、$T_1\#\xi^m_{2005}$ （橙色）和 $T_2\#\xi^f_{2005}$ （绿色）。

这个示例突显了所提出的模型不仅能够实现准确预测，还能提供可解释的洞见，揭示每个预测变量分布在塑造目标分布中的作用。

补充材料

单个预测变量下的GOT模型设定比较。多个预测变量下的GOT模型设定比较。证明。

参考文献

Agueh, M. & Carlier, G. (2011年), ‘Wasserstein空间中的重心’,《*SIAM*数学分析杂志》**43**(2), 904–924.

Ambrosio, L. (2008年), ‘度量空间及概率测度空间中的梯度流，以及对数凹测度下Fokker-Planck方程的应用’,意大利数学联盟公报**1**(1), 223–240.

Bigot, J., Gouet, R., Klein, T. & López, A. (2017), ‘通过凸主成分分析进行Wasserstein空间中的测地线主成分分析’, 亨利·庞加莱研究所年鉴：概率与统计**53**(1), 1–26.

网址： *https://doi.org/10.1214/15-AIHP706*

Chen, X., Fu, M., Huang, Y. & Deng, X. (2024), ‘分布输入—分布输出回归’,*arXiv* 预印本 *arXiv:2405.11626*.

Chen, Y., Lin, Z. & Müller, H.-G. (2023), ‘Wasserstein 回归’, 《美国统计学会会刊》**118**(542), 869–882.**网址：** *https://doi.org/10.1080/01621459.2021.1956937*

Chen, Z., Bao, Y., Li, H. & Spencer Jr, B. F. (2019), ‘基于 Lqd-rkhs 的分布对分布回归方法，用于恢复缺失的结构健康监测数据的概率分布’, 《机械系统与信号处理》**121**, 655–674.

Ghodrati, L. & Panaretos, V. M. (2022), ‘通过最优传输映射进行分布对分布回归’, 《生物统计学》109(4), 957–974.

Ghodrati, L. & Panaretos, V. M. (2024), ‘关于分布自回归与迭代传输’, 《时间序列分析杂志》45(5),739–770.

Ghosal, R., Ghosh, S. K., Schrack, J. A. & Zipunnikov, V. (2025), ‘基于分位数函数的分布结果回归及其在连续监测心率与身体活动建模中的应用’, 《美国统计学会会刊》第 1–13 页。

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K. &Filzmoser, P.(2016), ‘贝叶斯空间中密度函数的单纯形主成分分析’, 《计算统计与数据分析》94, 330–350.

Kokoszka, P., Miao, H., Petersen, A. &Shang, H. L.(2019), ‘密度函数预测及其在横截面收益和日内收益中的应用’, 《国际预测杂志》35(4), 1304–1317.

Nerini, D. & Ghattas, B. (2007), ‘使用函数型回归树对密度函数进行分类：在海洋学中的应用’, 《计算统计与数据分析》51(10), 4984–4993.

Panaretos, V. M. & Zemel, Y. (2016), ‘点过程的振幅变化与相位变化’,《统计年鉴》44(2), 771 – 812.**网址：** *https://doi.org/10.1214/15-AOS1387*

Petersen, A. & Müller, H.-G. (2016), ‘通过变换到希尔伯特空间对密度函数进行函数型数据分析’, 《统计年鉴》44(1),183 –218.**网址：** *https://doi.org/10.1214/15-AOS1363*

Petersen,A., Zhang, C. & Kokoszka, P. (2022), ‘将概率密度函数建模为数据对象’, 《计量经济学与统计学》21, 159 – 178.**网址：**
https://www.sciencedirect.com/science/article/pii/S245230622100054X

Santambrogio, F. (2015), ‘应用数学家的最优传输’. Villani, C. (2021), 最优传输专题, 第58卷, 美国数学学会. Yue, C.-S. J. & Huang, H.-C. (2011), ‘台湾人寿保险赔付发生经验研究’,《日内瓦风险与保险论文集：问题与实践》**36**(4),718– 733.

Zhang, C., Kokoszka, P. & Petersen, A. (2022), ‘Wasserstein自回归模型用于密度时间序列’, 《时间序列分析杂志》**43**(1), 30 – 52.**网址：**
https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsa.12590

Zhu, C. & Müller, H.-G. (2023), ‘自回归最优传输模型’, 《皇家统计学会杂志*B*辑：统计方法学》**85**(3), 1012–1033.

Zhu, C. & Müller, H.-G. (2025), ‘测地线最优传输回归’, 《生物统计学》第asaf086页.

URL： *https://doi.org/10.1093/biomet/asaf086*