Kernel Regression Utilizing External Information as Constraints

Chi-Shian Dai
Advisor: Jun Shao

1. Motivation
2. CKR for Summary Level External Data
3. CKR for Individual Level External Data
4. Simulation

Motivation

- Internal Data: $\{Y_i, \boldsymbol{U}_i\}_{i=1,\dots,n}$, $\boldsymbol{U}_i = (\boldsymbol{X}_i, \boldsymbol{Z}_i) \in \mathbb{R}^p$, $\boldsymbol{X} \in \mathbb{R}^q$.
- External Data: Sample size is $m >> n$. Provide information for $Y \sim \boldsymbol{X}$.
    - Sources: Population-based census, Past studies...
    - Summary Level Information: Least square estimator $\widehat{\beta}$ via $Y \sim \boldsymbol{X}$.
    - Individual Level information: $\{Y_i, \boldsymbol{X}_i\}_{i=n+1,\dots,n+m}$.
- Goal: Estimate $E[Y|\boldsymbol{U} = \boldsymbol{u}] := \mu(\boldsymbol{u})$

Inspiration

- Inspiring by **?**], they observe that the link between "internal" and "external" can be formulated as constraints. Hence, we consider a constrained kernel regression to estimate μ.
- Example: Let $Y = \boldsymbol{\beta}^\top \boldsymbol{X} + \boldsymbol{\gamma}^\top \boldsymbol{Z} + \epsilon$, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z}$, then there is a naive constraints

$$\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}.$$

  So, with the help of external data, we only have to focus on estimating $\boldsymbol{\gamma}$.
- We generalize this idea to kernel regression. The proposed method call constrained kernel regression (CKR).

1 Motivation

2 CKR for Summary Level External Data

3 CKR for Individual Level External Data

4 Simulation

Optimization Form of Kernel Regression

- Given a kernel κ and bandwidths l, and b.
- Kernel Regression estimate for $\mu(\boldsymbol{u})$:

$$\widehat{\mu}_K(\boldsymbol{u}) = \arg\min_{\mu} \sum_{j=1}^{n} \kappa_l(\boldsymbol{u} - \boldsymbol{U}_j)(Y_j - \mu)^2, \tag{1}$$

where $\kappa_l(\boldsymbol{u} - \boldsymbol{U}_j) = l^{-p}\kappa\left\{l^{-1}(\boldsymbol{u} - \boldsymbol{U}_j)\right\}$.

- Kernel Regression estimate for $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_n)$, $\mu_i = \mu(\boldsymbol{U}_i)$:

$$\widehat{\boldsymbol{\mu}}_K = \arg\min_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n}\sum_{j=1}^{n} \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)(Y_j - \mu_i)^2$$

Optimization Form of Kernel Regression

- There is another equivalent form.

$$\widehat{\boldsymbol{\mu}}_K = \arg\min_{\mu_1,\ldots,\mu_n} \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k)}(Y_j - \mu_i)^2 \qquad (2)$$

- We prefer (2) since

$$\sum_{j=1}^n \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k)}(Y_j - \mu_i)^2 \approx E[(Y - \mu(\boldsymbol{U}))^2 | \boldsymbol{U} = \boldsymbol{U}_i],$$

and

$$\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k)}(Y_j - \mu_i)^2 \approx E[(Y - \mu(\boldsymbol{U}))^2]$$

Constraints for Summary Level External Data

- $\widehat{\beta}$ is a consistent estimate for $\beta_0 := E[\boldsymbol{X}\boldsymbol{X}^\top]^{-1}E[\boldsymbol{X}Y]$, which satisfy

$$E\{(Y - \boldsymbol{X}^\top\beta_0)\boldsymbol{X}\} = 0.$$

Hence, the constrained optimization can be

$$\begin{aligned} \widehat{\boldsymbol{\mu}} = \arg\min_{\mu_1,\ldots,\mu_n} \sum_{i=1}^n \sum_{j=1}^n \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{k=1}^n \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k)}(Y_j - \mu_i)^2 \\ \text{subject to} \quad \sum_{i=1}^n (\mu_i - \boldsymbol{X}_i^\top\widehat{\beta})\boldsymbol{X}_i = 0. \end{aligned} \tag{3}$$

CKR for Summary Level External Data

- (3) is a quadratic programming. Hence, it can be solved by Lagrange multiplier.
- For arbitrary $\boldsymbol{u} \in \mathcal{U}$, we apply additional kernel regression by replacing Y with $\widehat{\boldsymbol{\mu}}$.

$$\widehat{\mu}_{CK}(\boldsymbol{u}) = \sum_{i=1}^{n} \widehat{\mu}_i \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i) \Bigg/ \sum_{i=1}^{n} \kappa_b(\boldsymbol{u} - \boldsymbol{U}_i). \tag{4}$$

Non-Constrained Methods

- Kernel Regression (KR): $\widehat{\mu}_K(\boldsymbol{u})$
- Double Kernel Regression (DKR): Consider CKR without applying any constraints. Use notation $\widehat{\mu}_{DK}(\boldsymbol{u})$.
  - Step 1: Estimate $(\mu(\boldsymbol{U}_1), \ldots, \mu(\boldsymbol{U}_n))$ by Kernel Regression
  - Step 2: Estimate $\mu(\boldsymbol{u})$ by additional kernel regression replacing Y with the results in first step.

Assumption for Normality

(A1) The response Y has a finite $E|Y|^s$ with $s > 2 + p/2$. The covariate vector $\boldsymbol{U}$ has compact support $\mathcal{U}$, and has a positive definite covariance matrix. The density of $\boldsymbol{U}$ is bounded away from infinity and zero, and has bounded second-order derivatives.

(A2) Functions $\mu(\boldsymbol{u}) = E(Y|\boldsymbol{U} = \boldsymbol{u})$ and $\sigma^2(\boldsymbol{u}) = E[\{Y - \mu(\boldsymbol{U})\}^2|\boldsymbol{U} = \boldsymbol{u}]$ are Lipschitz continuous. The function $\mu(\boldsymbol{u})$ has bounded third-order derivatives.

(A3) The kernel κ is a positive bounded density with mean zero and finite sixth moments. Furthermore, κ is Lipschitz continuous.

(A4) The bandwidths b in (1) and l in (3) are polynomial rate of n, satisfying $b \to 0$, $l \to 0$, $l/b \to \gamma \in (0, \infty)$, and $nb^{4+p} \to c \in [0, \infty)$ as the internal sample size $n \to \infty$.

(A5) The external sample size m satisfies $n = O(m)$, i.e., n/m is bounded by a fixed constant.

Theorem

Theorem 1

Assume conditions (A1)-(A5). Then, as $n \to \infty$,

$$\sqrt{nb^p}\{\widehat{\boldsymbol{\mu}}_t(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \to N\big(B_t(\boldsymbol{u}), V_t(\boldsymbol{u})\big) \quad \textit{in distribution}, \tag{5}$$

where $t = DK$ or CK,

$$\begin{aligned}
B_{DK}(\boldsymbol{u}) &= c^{1/2}(1+\gamma^2)A(\boldsymbol{u}),\\
B_{CK}(\boldsymbol{u}) &= c^{1/2}[(1+\gamma^2)A(\boldsymbol{u}) - \gamma^2\boldsymbol{x}^\top\boldsymbol{\Sigma}_X^{-1}E\{\boldsymbol{X}A(\boldsymbol{U})\}],\\
V_{DK}(\boldsymbol{u}) &= \frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})}\int\left\{\int \kappa(\boldsymbol{w}-\boldsymbol{v}\gamma)\kappa(\boldsymbol{v})d\boldsymbol{v}\right\}^2 d\boldsymbol{w},\\
V_{CK}(\boldsymbol{u}) &= V_{DK}(\boldsymbol{u}),\\
A(\boldsymbol{u}) &= \int \kappa(\boldsymbol{v})\left\{\tfrac{1}{2}\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v} + \nabla\mu(\boldsymbol{u})^T\boldsymbol{v}\boldsymbol{v}^T\nabla\log f_U(\boldsymbol{u})\right\}d\boldsymbol{v}, \qquad (6)
\end{aligned}$$

and f_U is the density of $\boldsymbol{U}$.

Asymptotic Mean Integrated Square Error

- $\mathrm{AMISE}(\widehat{\mu}_t) = E[\{B_t(\boldsymbol{U})\}^2 + V_t(\boldsymbol{U})], \quad t = CK$ or DK,
- Observe that $\boldsymbol{\xi} := \Sigma_X^{-1} E\{\boldsymbol{X} A(\boldsymbol{U})\}]$ is a linear coefficient of fitting $A(\boldsymbol{U})$ by $\boldsymbol{X}$. Hence,

$$E[\{A(\boldsymbol{U}) - \boldsymbol{X}^\top \boldsymbol{\xi}\}\boldsymbol{X}] = 0.$$

- From this we can derive
$$E\{B_{CK}(\boldsymbol{U})\}^2 = c(1+\gamma^2)^2 E\{A(\boldsymbol{U}) - \boldsymbol{X}^\top \boldsymbol{\xi}\}^2 + cE\{\boldsymbol{X}^\top \boldsymbol{\xi}\}^2,$$
$$E\{B_{DK}(\boldsymbol{U})\}^2 = c(1+\gamma^2)^2 E\{A(\boldsymbol{U})\}^2$$
$$= c(1+\gamma^2)^2 E\{A(\boldsymbol{U}) - \boldsymbol{X}^\top \boldsymbol{\xi}\}^2 + c(1+\gamma^2)^2 E\{\boldsymbol{X}^\top \boldsymbol{\xi}\}^2.$$

CKR V.S. KR

- $B_K(\boldsymbol{u}) = c^{1/2}A(\boldsymbol{u})$ and $V_K(\boldsymbol{u}) = \frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})} \int \{\kappa(\boldsymbol{v})\}^2 d\boldsymbol{v}$.
- $B_{CK}(\boldsymbol{u}) = c^{1/2}[(1+\gamma^2)A(\boldsymbol{u}) - \gamma^2 \boldsymbol{x}^\top \boldsymbol{\Sigma}_X^{-1} E\{\boldsymbol{X}A(\boldsymbol{U})\}]$,
- $V_{CK}(\boldsymbol{u}) = \frac{\sigma^2(\boldsymbol{u})}{f_U(\boldsymbol{u})} \int \left\{\int \kappa(\boldsymbol{w} - \boldsymbol{v}\gamma)\kappa(\boldsymbol{v}) d\boldsymbol{v}\right\}^2 d\boldsymbol{w}$,
- If $\gamma = 0$, $B_{CK} = B_K$ and $V_{CK} = V_K$. So, $\inf_\gamma \mathrm{AMISE}(\widehat{\mu}_{CK})(\gamma) \leq \mathrm{AMISE}(\widehat{\mu}_K)$
- If γ increases, V_{CK} decreases and B_{CK}^2 increases.

Theorem 2

Under the conditions in Theorem 1 and an additional condition that $\int \nabla^2 \kappa(\boldsymbol{u})\kappa(\boldsymbol{u}) d\boldsymbol{u}$ being strictly negative definite, $\mathrm{AMISE}(\widehat{\mu}_{CK}) < \mathrm{AMISE}(\widehat{\mu}_K)$ for c and γ in a neighborhood of 0.

1 Motivation

2 CKR for Summary Level External Data

3 CKR for Individual Level External Data

4 Simulation

CKR for Individual Level External Data

- First, estimate $h = E[Y|X]$ via kernel regression.
- Second, observe that for all real function g

$$E\{Y - h(\boldsymbol{X})\}g(\boldsymbol{X}) = 0.$$

- Consider the coresponding constrained optimization.

$$\begin{aligned} \widehat{\boldsymbol{\mu}} = \arg \min_{\mu_1,\ldots,\mu_n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_j)}{\sum_{k=1}^{n} \kappa_l(\boldsymbol{U}_i - \boldsymbol{U}_k)} (Y_j - \mu_i)^2 \\ \text{subject to} \quad \sum_{i}^{n} \{\mu_i - \widehat{h}(\boldsymbol{X}_i)\} g(\boldsymbol{X}_i) = 0. \end{aligned} \tag{7}$$

- Question: How to choose g?

Theorem 3

Assume (A1)-(A5) in Theorem 1 and

(A1′) *The matrix* $\boldsymbol{\Sigma}_g = E\{g(\boldsymbol{X})g(\boldsymbol{X})^\top\}$ *is positive definite.*

(A2′) *The functions* $h(\boldsymbol{x})$ *and* $g(\boldsymbol{x})$ *are Lipschitz continuous. The function* $h(\boldsymbol{x})$ *has bounded third-order derivatives.*

(A3′) *The kernel function used in the kernel regression based on the external dataset satisfies condition (A3).*

(A4′) *The bandwidth used in the kernel regression based on the external dataset is of the order* $m^{-1/(4+q)}$ *as* $m \to \infty$.

Then, as $n \to \infty$, *CKR with constraints* (7) *have following properties.*

$$\sqrt{nb^p}\{\widehat{\mu}_{CK}(\boldsymbol{u}) - \mu(\boldsymbol{u})\} \to N\big(B_{CK}(\boldsymbol{u}), V_{CK}(\boldsymbol{u})\big) \quad \textit{in distribution,}$$

where

$$B_{CK}(\boldsymbol{u}) = c^{1/2}[(1+\gamma^2)A(\boldsymbol{u}) - \gamma^2 g(\boldsymbol{x})^\top \boldsymbol{\Sigma}_g^{-1} E\{g(\boldsymbol{X})A(\boldsymbol{U})\}]$$

and $V_{CK}(\boldsymbol{u})$ *and* $A(\boldsymbol{u})$ *are the same as those in Theorem 1.*

Choose of g

- $E\{B_{CK}(\boldsymbol{U})\}^2 = c\{(1+\gamma^2)^2-1\}E\{A(\boldsymbol{U})-g(\boldsymbol{X})^\top\boldsymbol{\xi}_g\}^2+cE\{A(\boldsymbol{U})\}^2$, where $\boldsymbol{\xi}_g = \boldsymbol{\Sigma}_g^{-1}E\{g(\boldsymbol{X})A(\boldsymbol{U})\}]$ is a linear coefficient of $A(\boldsymbol{U}) \sim g(\boldsymbol{X})$.
- Find g to minimize $E\{A(\boldsymbol{U}) - g(\boldsymbol{X})^\top\boldsymbol{\xi}_g\}^2$.
- The best one is $g^* = E[A(\boldsymbol{U})|\boldsymbol{X}]$.
- $A(\boldsymbol{u}) = \int \kappa(\boldsymbol{v}) \left\{\frac{1}{2}\boldsymbol{v}^\top\nabla^2\mu(\boldsymbol{u})\boldsymbol{v} + \nabla\mu(\boldsymbol{u})^T\boldsymbol{v}\boldsymbol{v}^T\nabla\log f_U(\boldsymbol{u})\right\} d\boldsymbol{v}$, is estimable.
- g^* is also estimable.

Theorem 4

Assume the conditions in Theorem 3 and the following additional conditions.

(C1) *The kernel κ in (A3) satisfies $\int u_k^2 \kappa(\boldsymbol{u}) d\boldsymbol{u} = 1$ and $\int u_k u_j \kappa(\boldsymbol{u}) d\boldsymbol{u} = 0$ when $k \neq j$. The kernel $\widetilde{\kappa}$ in the estimators $\widehat{\nu}_k$ and $\nabla^2_{kk} \widehat{f}_U$, $k = 1, ..., p$, has finite second-order moments, bounded $\nabla^2_{kk} \widetilde{\kappa}$, finite $\int |\nabla^2_{kk} \widetilde{\kappa}(\boldsymbol{u})| d\boldsymbol{u}$, and bounded $\sup_{\boldsymbol{u}} \lambda^{-2} |\widetilde{\kappa}(\boldsymbol{u}/\lambda)|$ and $\sup_{\boldsymbol{u}} \lambda^{-3} |\nabla_k \widetilde{\kappa}(\boldsymbol{u}/\lambda)|$ as $\lambda \to 0$, $k = 1, ..., p$.*

(C2) *The bandwidth λ_1 for $\widehat{\nu}_0$ and $\widehat{f}_U$ has order $n^{-1/(p+4)}$, the bandwidth λ_2 for $\widehat{\nu}_k$ and $\nabla^2_{kk} \widehat{f}_U$ has order $n^{-1/(p+8)}$, and the bandwidth δ in estimating g^* has order $n^{-1/(q+4)}$.*

Then, the result in Theorem 3 with $g = g^*$ holds for $\widehat{\mu}_{CK}$ using the estimated constraints $\widehat{g}^*$.

1 Motivation

2 CKR for Summary Level External Data

3 CKR for Individual Level External Data

4 Simulation

Simulation

- Internal sample size : 200
- External sample size : 1000
- X, Z are normal distribution with variance 1, covariance 0.5.
- $Y = \mu(X, Z) + \epsilon$; $\epsilon \sim N(0, \sigma^2)$
  1. Additive Models:
     $\mu = X^3 + Z^2$
     $\mu = 2^{-1}cos(2X) + cos(Z)$
     $\mu = cos(X) + cos(Z)$
  2. Non-Additive Models:
     $\mu = X^3 + XZ + Z^2$

Simulation

- Let $R = 200$ be the number of independently replication.
- Let $L = 121$ be the sample size of test data.
  1. Fixed grid points on $[-1, 1] \times [-1, 1]$
  2. Random sample without replacement from the covariate $\boldsymbol{U}'s$ of the internal data set.
- Use estimated MISE to evaluate performance.

$$\mathrm{MISE} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{L} \sum_{l=1}^{L} \{\widehat{\mu}_r(\boldsymbol{T}_{r,l}) - \mu(\boldsymbol{T}_{r,l})\}^2,$$

Simulation

- Best Bandwidth: Evaluate MISE in a pool of bandwidths and display the one have the best performance.
- 10 folds cross-validation
- $$\text{Imp\%} = 1 - \frac{\min\{\mathrm{MISE}(\widehat{\mu}_{CKR}) \text{ over all CKR methods}\}}{\min\{\mathrm{MISE}(\widehat{\mu}_{K}), \mathrm{MISE}(\widehat{\mu}_{DK})\}}.$$

Test Data: Sample

model	σ	test data	b, l	estimator (CKR) with $g =$ 1	$(1, X)$	$(1, \widehat{h})$	$\widehat{g}^*$	g^*	estimator (CKR-s)	(KR)	(DKR)	Imp%
1	3	sample	best	1.045	0.924	0.912	1.055	0.843	1.152	1.081	1.056	20.17
			CV	1.165	1.148	1.073	1.19	1.176	1.409	1.239	1.181	9.14
2	3	sample	best	0.220	0.225	0.222	0.259	0.210	0.201	0.266	0.260	22.69
			CV	0.297	0.290	0.323	0.341	0.282	0.274	0.335	0.343	18.20
3	3	sample	best	0.338	0.347	0.333	0.437	0.365	0.298	0.443	0.439	32.13
			CV	0.537	0.512	0.581	0.643	0.539	0.47	0.620	0.640	24.19
4	3	sample	best	1.102	1.066	1.018	1.102	1.020	1.437	1.117	1.099	7.37
			CV	1.276	1.294	1.329	1.262	1.410	1.624	1.208	1.270	-4.47

Test Data: Fixed Grid Points on $[-1,1]^2$

model	σ	test data	b, l	estimator (CKR) with $g =$ 1	$(1, X)$	$(1, \widehat{h})$	$\widehat{g}^*$	g^*	estimator (CKR-s)	(KR)	(DKR)	Imp%
1	3	grid	best	0.175	0.171	0.181	0.205	0.206	0.243	0.210	0.208	17.78
			CV	0.382	0.365	0.341	0.388	0.355	0.432	0.412	0.384	11.19
2	3	grid	best	0.111	0.108	0.076	0.148	0.132	0.100	0.153	0.153	50.32
			CV	0.141	0.138	0.117	0.164	0.151	0.134	0.160	0.162	26.87
3	3	grid	best	0.102	0.100	0.070	0.129	0.131	0.089	0.135	0.132	46.96
			CV	0.123	0.122	0.099	0.145	0.151	0.121	0.142	0.142	30.28
4	3	grid	best	0.231	0.244	0.229	0.250	0.251	0.338	0.251	0.251	8.76
			CV	0.414	0.426	0.359	0.400	0.369	0.486	0.367	0.397	2.17

Future Works

1. We will study more when and why the CKR improves the standard KR, both empirically and theoretically.
2. With individual-level external data, we only considered kernel estimators of $h(\boldsymbol{x})$. Different types of estimators of h will be studied—for example, linear models or generalized linear models.
3. We mainly considered the AMISE as a criterion in evaluating kernel estimators. We will study some other performance measures.
4. We focused on the situation where the underlying distribution of $\boldsymbol{U}$ is the same in both internal and external data. We will consider extensions in situations where the distributions of $\boldsymbol{U}$ are different in two datasets.

References