# Kernel Ridge Regression under Random Projection: Computational-and-Statistical Trade-off

Joint work with Prof. Shang and Prof. Cheng

BaT Group Meeting · Meimei Liu · March 23, 2016

# Outline

# Model Description and RKHS

- Suppose that data $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. copies of $(X, Y)$, consider the following nonparametric regression model

$$y_i = f^*(x_i) + \epsilon_i, \quad \text{for } i = 1, \cdots, n \tag{1.1}$$

  where $\epsilon_i$'s are iid Sub-Gaussian with $E[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$.

- Suppose $f^*$ belong to a RKHS $\mathcal{H}$ with kernel function $K(\cdot, \cdot)$, satisfying $\langle f, K(x, \cdot)\rangle_{\mathcal{H}} = f(x)$, for all $f \in \mathcal{H}$.

- By Mercer's theorem, the kernel function $K$ associated with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ admits an eigen-decomposition

$$K(x, t) = \sum_{k=1}^{\infty} \mu_k \phi_k(x) \phi_k(t), \tag{1.2}$$

# Primary assumptions on RKHS

- The kernel function $K$ is bounded on the diagonal line, i.e., there exists a finite positive constant $c_{\mathcal{H}} > 0$ such that

$$\sup_{x \in \mathcal{X}} |K(x, x)| \leq c_{\mathcal{H}}.$$

- Furthermore, the eigenfunctions $\{\phi_k\}_{k=0}^{\infty}$ are uniformly bounded on $\mathcal{X}$, i.e., there exists a finite constant $c_K > 0$ such that

$$\sup_{j \in \mathbb{N}} \|\phi_j\|_{\sup} \leq c_K. \tag{1.3}$$

# Kernel Ridge Regression

- The classic KRR estimate

$$\widehat{f}_n := \arg\min_{f\in\mathcal{H}} \ell_{n,\lambda}(f) \equiv \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_{\mathcal{H}}^2 \right\}. \tag{1.4}$$

- Let $K$ be the $n$-dimensional kernel matrix with entries $K_{ij} = K(x_i, x_j)$ for $1 \le i, j \le n$. By representer theorem, $\widehat{f}_n$ must have the form

$$\widehat{f}_n = \sum_{i=1}^{n} \omega_i^{\dagger} K(\cdot, x_i)$$

for a real vector $\omega^{\dagger} = (\omega_1^{\dagger}, \ldots, \omega_n^{\dagger})^T$.

- This reduces (1.4) to solve the following optimization problem:

$$\omega^{\dagger} = \arg\min_{\omega \in \mathbb{R}^n} \left\{ \omega^T K^2 \omega - 2 y^T K \omega^T + n\lambda \omega^T K \omega \right\}. \quad (1.5)$$

- The KRR estimator is expressed as $\widehat{f}_n(\cdot) = \sum_{i=1}^{n} \omega_i^{\dagger} K(\cdot, x_i)$.

- KRR estimator could achieve the minimax prediction error for various classes of kernels,see [2] and [3].

- KRR estimator could achieve the minimax prediction error for various classes of kernels,see [2] and [3].
- But consider the computational complexity when dealing large-scale data:
  - The $n$ dimensional quadratic program in (1.5) requires $\mathcal{O}(n^3)$ via QR decomposition.
  - The $n$ dimensional matrix $K$ is dense in general, so requires storage of order $n^2$ numbers.

To reduce the complexity involved in an $n$-dimensional kernel matrix without sacrificing statistical estimation optimality

- $s$-dimensional random projection on $n$-dimensional kernel matrix.

# Projected Kernel Ridge Regression (PKRR)

- The idea is to randomly project $K$ to a lower-dimensional $s \times n$ random matrix $SK$, then (1.5) becomes

$$\widehat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^s} \left\{ \alpha^T (SK)(KS^T)\alpha - 2\alpha^T SKy + n\lambda \alpha^T SKS^T \alpha \right\}. \tag{2.1a}$$

Heuristically, solutions to (2.1a) and (1.5) satisfy the apparent relationship $\omega^\dagger = S^T \widehat{\alpha}$. Therefore, we can replace $\omega^\dagger$ in $\widehat{f}_n$ by $\widehat{\alpha}$ to get the following PKRR estimator:

$$\widehat{f}_r(\cdot) = \sum_{i=1}^{n} (S^T \widehat{\alpha})_i K(\cdot, x_i). \tag{2.1b}$$

# Choice of Projection Matrix

- How to determine projection dimension $s$?

$$s = \arg\min_j \{\hat{\mu}_j \leq \delta_n\}. \tag{2.2}$$

  where $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n > 0$ are elements of diagonal matrix $D$ in the SVD of $K = UDU^T$.

- What kind of $S$ can we use?
  - the random Gaussion projection matrix;
  - sub-sampling projection;
  - randomized orthogonal system (ROS) projection.

# Condition on Projection Matrix $S$

- For $K = UDU^T$, let $U = (U_1, U_2)$, where $U_1 \in \mathbb{R}^{n \times s}$, and $U_2 \in \mathbb{R}^{n \times (n-s)}$. Let $D_1 = diag\{\hat{\mu}_1, \cdots, \hat{\mu}_s\}$ and $D_2 = diag\{\hat{\mu}_{s+1}, \cdots, \hat{\mu}_n\}$.
- With probability approaching one, there exist positive constants $c_1$ and $c_2$, such that

$$c_1 \leq \lambda_{\min}((SU_1)^T(SU_1)) \leq \lambda_{\max}((SU_1)^T(SU_1)) \leq c_2,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the minimum and maximum eigenvalues of a square matrix.

# Condition on Projection Matrix $S$

- There exists a univerisal constant $c$ such that, with probability approaching one,

$$\|SU_2\, D_2^{1/2}\|_{\text{op}} \leq c\delta_n,$$

  where $\|\cdot\|_{\text{op}}$ is the operator norm of a matrix.

- Remark: $U_1$ and $U_2$ represent the loading matrices corresponding to the $s$ principal eigenvalues and $n-s$ inprincipal eigenvalues, respectively.
  We assume a "good" random matrix $S$ satisfy that the matrix $(SU_1)^T(SU_1)$ is not singular and that the matrix $SU_2D_2^{1/2}$ is controlled by certain order $\delta_n$.

# Purpose of this paper:
# **Co**mputationally e**ff**icient inf**er**ence (coffee)

- How small can the projection dimension $s$ be chosen while still retaining minimax optimality in estimation and testing?
- Local CI: For any $x_0 \in \mathcal{X}$, the asymptotic normality of $\widehat{f}_r(x_0)$
- Using PKRR $\widehat{f}_r$ for Global hypothesis testing:

$$H_0 : f^* = f_0 \iff H_1 : f^* \neq f_0. \tag{2.3}$$

# Some Preliminary Definition

- Define $\langle \cdot, \cdot \rangle$ as $\langle f, g \rangle := \langle f, g \rangle_{L^2(\mathbb{P})} + \lambda \langle f, g \rangle_{\mathcal{H}}$, for any $f, g \in \mathcal{H}$. Clearly, $\langle \cdot, \cdot \rangle$ is an inner product on $\mathcal{H}$. Let $\| \cdot \|$ be the corresponding norm.
- There exists a positive definite self-adjoint operator $W_\lambda : \mathcal{H} \to \mathcal{H}$ such that $\langle W_\lambda f, g \rangle = \lambda \langle f, g \rangle_{\mathcal{H}}$.
- Define

$$\gamma(\lambda) = \sum_{j \in \mathbb{Z}} \frac{1}{1 + \lambda / \mu_j}.$$

  Let $h = h(\lambda)$ be a positive-valued function of $\lambda$ satisfying that $h(\lambda) \leq \gamma(\lambda)^{-1}$.

# Entropy

- Recall the $\epsilon$-packing number of $\mathcal{H}$ in metric $d$ is defined as

$$\mathcal{N}(\mathcal{H}; d, \varepsilon) = \max\{|N| : N \text{ is an } \varepsilon\text{-packing of } \mathcal{H}\}.$$

- For $\mathcal{H}_1 := \{f \in \mathcal{H} : \|f\|_{\sup} \leq 1, \|f\|_{\mathcal{H}} \leq 1\}$, the Orlicz entropy integral is defined as :

$$\omega(\delta) = \int_0^{\delta} \sqrt{\log(1 + \mathcal{N}(\mathcal{H}_1; \|\cdot\|_{\sup}, \epsilon))} d\epsilon. \tag{2.4}$$

- Also define

$$\xi_\lambda := (c_K^{-2} h \lambda^{-1})^{1/2} \omega((c_K^{-2} h \lambda^{-1})^{-1/2}). \tag{2.5}$$

# Proposition for Estimation Error

**Proposition**

*Assume* $n^{-1}(\sqrt{n}\xi_\lambda + 1)c_K^2 h^{-1} \log\log(n) \leq 1/2$, *and* $\lambda = o(1)$, *if* $\delta_n^{1/2} = o(n^{1/2}\lambda)$ *holds, then we have*

$$\|\widehat{f}_r - f^*\| = \mathcal{O}_p(r_n + \delta_n^{1/2} n^{-1/2}\lambda^{-1}),$$

*where* $r_n = (nh)^{-1/2} + \lambda^{1/2}$.

# Asymptotic Normality for PKRR $\widehat{f}_r$

## Theorem

*Define* $f_0^* = f^* + W_\lambda f^*$, *assume* $\lambda = o(1)$, $h = o(1)$, $a(h, r_n, \xi_\lambda, \delta_n) = o(n^{-1/2})$, *and* $h$ *satisfying*

$$\lim_{n\to\infty} \sigma^2 h \sum_{k=0}^{\infty} \left( \frac{\phi_k(x_0)}{1+\lambda/\mu_k} \right)^2 = \sigma_{x_0}^2 < \infty.$$

*Then for any* $x_0 \in \mathcal{X}$, *as* $n \to \infty$,

$$\sqrt{nh}\left(\widehat{f}_r(x_0) - f_0^*(x_0)\right) \xrightarrow{d} N(0, \sigma_{x_0}^2).$$

*If* $\sqrt{nh}\lambda = o(1)$, *then* $\sqrt{nh}\left(\widehat{f}_r(x_0) - f^*(x_0)\right) \xrightarrow{d} N(0, \sigma_{x_0}^2)$.

- Here $a(h, r_n, \xi_\lambda, \delta_n)$ is defined as $(nh)^{-1} c_K^2 r_n(\sqrt{n}\xi_\lambda + 1) \log\log(n) + \delta_n^{1/2} n^{-1/2} \lambda^{-1}$.
- Given any $x_0$, the $100(1-\alpha)\%$ local CI for $f(x_0)$ is

$$\widehat{f}_r(x_0) \pm z_{\alpha/2} \hat{\sigma}_{x_0}$$

Consider the following simple hypothesis:

$$H_0 : f^* = f_0 \text{ versus } H_1 : f^* \neq f_0. \tag{3.1}$$

The PLRT statistic involves the PKRR estimator, is defined as follows:

$$\mathrm{PLRT}_{r,\lambda} := \ell_{n,\lambda}(\widehat{f}_r) - \ell_{n,\lambda}(f_0). \tag{3.2}$$

# Global Hypothesis Testing

## Theorem

*Under $H_0$ in (3.1), assume $E[\epsilon^4|X] < C$, a.s., for some constant $C$. Suppose the following conditions hold:* $n\lambda h = \mathcal{O}(1)$; $(nh^2)^{-1} = o(1)$; $\lambda = o(1), h = o(1)$; $a(h, r_n, \xi_\lambda, \delta_n) = o(n^{-1/2})$; *and* $(nh)^{-1}(r_n^2 + n^{-1})(\sqrt{n}\xi_\lambda + 1) = o(h^{-1/2})$.
*Then under $H_0$, we have*

$$(2u_n)^{-1/2}(-2nr_K \cdot \mathrm{PLRT}_{r,\lambda} - nr_K\|W_\lambda f^*\|^2 - u_n) \xrightarrow{d} N(0,1) \quad (3.3)$$

*where* $u_n = h^{-1}\sigma_K^4/\rho_K^2$, $r_K = \sigma_K^2/\rho_K^2$ *for* $\sigma_K^2 = \sigma^2 h\sum_k (1+\lambda/\mu_k)^{-1}$ *and* $\rho_K^2 = \sigma^4 h\sum_k (1+\lambda/\mu_k)^{-2}$.

- Remark:
  Since $n\|W_\lambda f^*\|^2 = o(n\lambda) = o(h)$, we have $-2nr_K \cdot \mathrm{PLRT}_{r,\lambda}$ is asymptotically $N(u_n, 2u_n)$, which is nearly $\chi^2_{u_n}$.

In the end, we consider the minimax optimality of the proposed testing in the sense of [1]. Define

$$\mathcal{F}_\zeta \equiv \{f \in \mathcal{H} | Var(f(X)^2) \leq \zeta \mathbb{E}^2[f(X)^2], \|f\|_{\mathcal{H}}^2 \leq \zeta\},$$

for some constant $\zeta > 0$. Let $f_n \in \mathcal{F}_\zeta$ and consider a sequence of local alternatives $H_{1n} : f = f_n$.

## Theorem

*Let Assumption 1-3 be satisfied and, conditions in Theorem 2 holds. Then under $H_0$ in (3.1), for any $\varepsilon > 0$ there exist positive constants $C$ and $N$ such that*

$$\inf_{n \geq N} \inf_{f_n \in \mathcal{F}_\zeta, \|f_n\| \geq c\eta_n} P\left( \left| \frac{-2nr_K \cdot \mathrm{PLRT}_{r,\lambda} - u_n}{\sqrt{2u_n}} \right| > z_\alpha | H_{1n} \ is \ true \right) \geq 1 - \varepsilon,$$

*where $\eta_n \geq \sqrt{\lambda + (nh^{1/2})^{-1}}$.*

# Example 1: Polynomial decaying kernel with $\mu_k \asymp k^{-2m}$

Polynomial decaying kernel

| | minimax estimation | local CI | minimax testing |
|---|---|---|---|
| $\lambda$ | $n^{-\frac{2m}{2m+1}}$ | $n^{-\frac{2m}{2m+1}}$ | $n^{-\frac{4m}{4m+1}}$ |
| $h$ | $n^{-\frac{1}{2m+1}}$ | $n^{-\frac{1}{2m+1}}$ | $n^{-\frac{2}{4m+1}}$ |
| $\delta_n$ | $n^{-\frac{4m-1}{2m+1}}$ | $n^{-\frac{12m^2-6m+1}{(2m+1)2m}}$ | $n^{-\frac{12m^2-7m+1}{m(4m+1)}}$ |
| $s$ | $n^{\frac{4m-1}{2m(2m+1)}}$ | $n^{\frac{12m^2-6m+1}{4m^2(2m+1)}}$ | $n^{\frac{12m^2-7m+1}{2m^2(4m+1)}}$ |

Table: Critical values of $(\lambda, h, \delta_n, s)$ to obtain optimality in estimation and inferences for polynomial decaying kernel.

# Concrete Example: Cubic Spline

Consider an $m$-order periodic Soblev space with eigenvalues $\mu_{2k-1} = \mu_{2k} = (2\pi k)^{-2m}$ for $k \geq 1$.

- when $m = 2$ (cubic spline), Table 1 suggests lower bounds for $s$: $n^{1/3}$ for estimation and $n^{1/2}$ for testing.
- Define $I_l = \int_0^\infty \frac{dx}{(1+x^{2m})^l}$ for $l = 1, 2$, we have
  - For CI, $\sigma_{x_0}^2 \sim \frac{\sigma^2 I_2}{\pi}$;
  - For PLRT, $\sigma_K^2 \sim \frac{\sigma^2 I_1}{\pi}$ and $\rho_K^2 \sim \frac{\sigma^4 I_2}{\pi}$.

# Ex 2: Exponential dacaying kernel with $\mu_k \asymp e^{-\alpha k^p}$

Exponential decaying kernel

| | minimax estimation | local CI | minimax testing |
|---|---|---|---|
| $\lambda$ | $(\log n)^{\frac{1}{p}} n^{-1}$ | $(\log n)^{\frac{1}{p}} n^{-1}$ | $(\log n)^{\frac{1}{2p}} n^{-1}$ |
| $h$ | $(\log n)^{-\frac{1}{p}}$ | $(\log n)^{-\frac{1}{p}}$ | $(\log n)^{-\frac{1}{p}}$ |
| $\delta_n$ | $n^{-2}(\log n)^{\frac{1}{3p}}$ | $n^{-3}(\log n)^{\frac{6+p}{p}}$ | $n^{-3}(\log n)^{\frac{p+5}{p}}$ |
| $s$ | $(2\log n)^{\frac{1}{p}}$ | $(3\log n)^{\frac{1}{p}}$ | $(3\log n)^{\frac{1}{p}}$ |

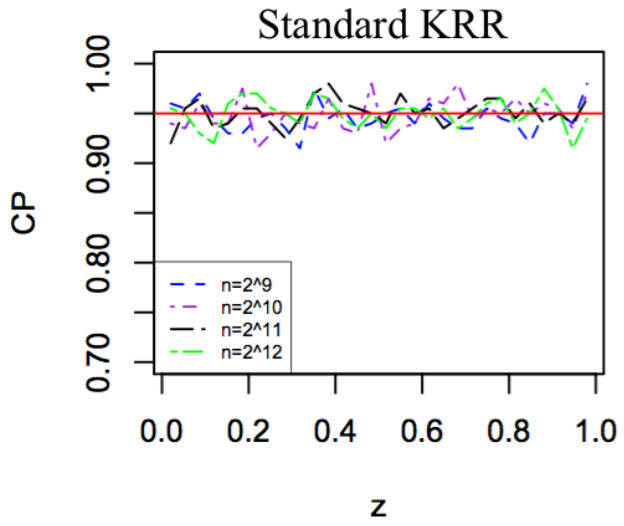
Table: Critical values of $(\lambda, h, \delta_n, s)$ to obtain optimality in estimation and inferences for exponential decaying kernel.

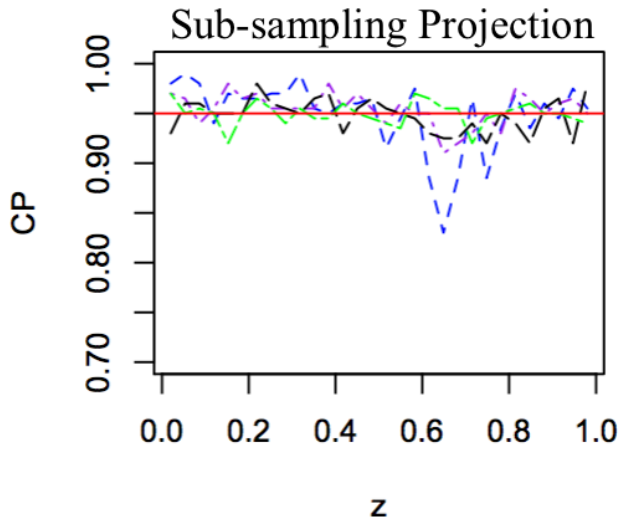
- when $p = 2$ (Gaussian kernel), the lower bound of $s$ is $\sqrt{2\log n}$ for estimation, and $\sqrt{3\log n}$ for both local CI and testing.

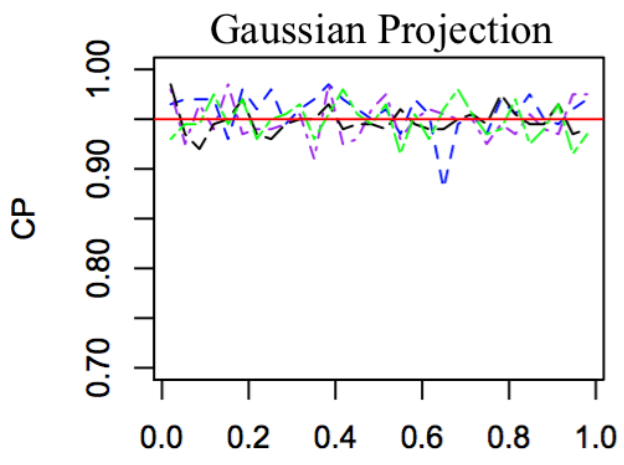
# Simulation study 1: local CI for polynomial decaying kernel

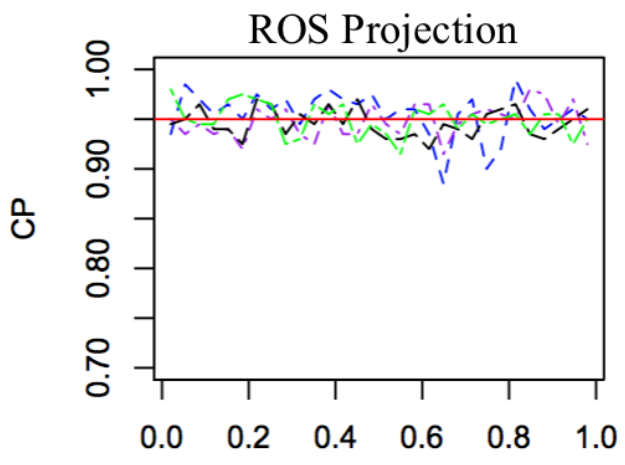
- The true function $f^*(x) = 3\beta_{30,17}(x) + 2\beta_{3,11}(x)$, where $\beta_{a,b}$ is the Beta density function.
- $X_i \sim \text{Unif}[0,1]$, $\sigma = 1$ and sample sizes $n = 2^9, 2^{10}, 2^{11}, 2^{12}$.
- The projection dimension $s$ was chosen as $s = n^{1/2}$.

# Simulation study 1: local CI for polynomial decaying kernel



# Simulation study 2: PLRT for $H_0$ : f is linear

- $f^*(z) = -0.5 + 3z + c\sin(2\pi z)$, for $c = 0, 0.5, 1, 1.5$.
- For polynomial kernel, projection dimension was chosen as $s = n^{1/2}$.
- For Gaussian kernel, the projection dimension was chosen as $s = \sqrt{3 \log n}$

# Simulation study 2: PLRT for $H_0$ : f is linear

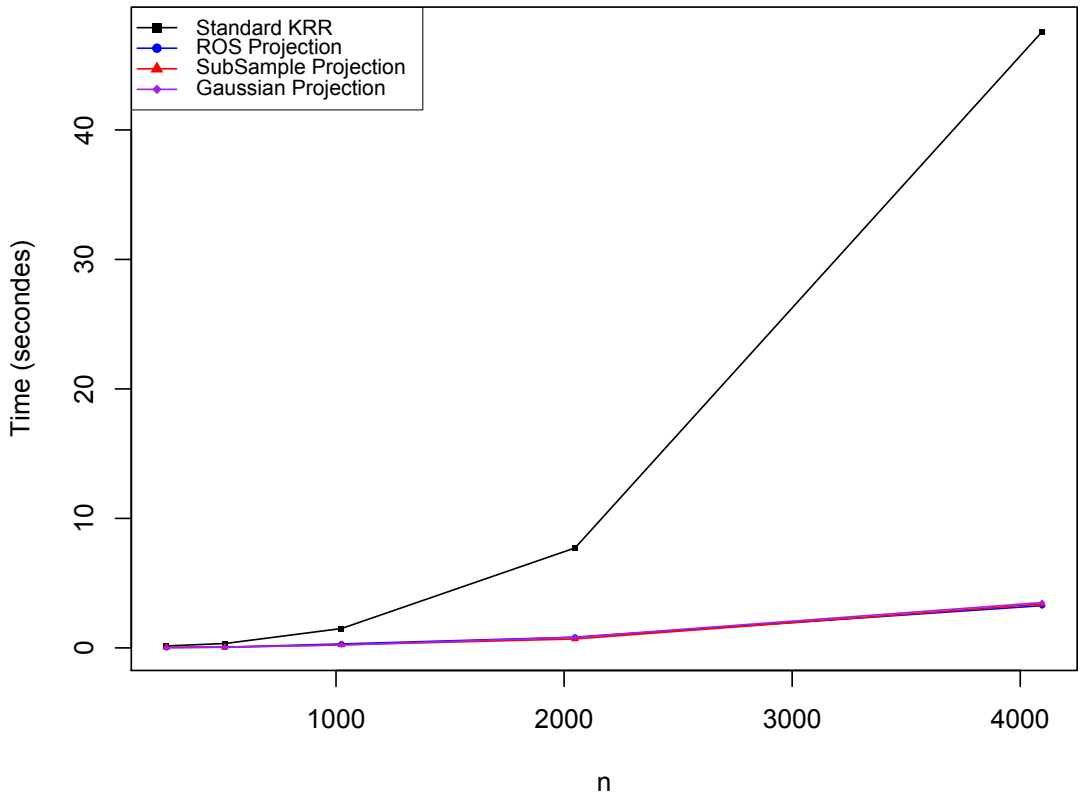



Figure: *Time for calculating PLRT in testing $H_0$ : f is linear when the true function is $f_0(z) = -0.5 + 3z + \sin(2\pi z)$.*

| Projection Matrix | Polynomial decaying kernel | | | | | Exponential decaying kernel | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | $c=0$ | $c=0.5$ | $c=1$ | $c=2$ | $n$ | $c=0$ | $c=0.5$ | $c=1$ | $c=2$ |
| Sub Gaussian | 256 | 4.6 | 81.0 | 100.0 | 100.0 | 256 | 5.0 | 8.0 | 100.0 | 100.0 |
| | 512 | 5.0 | 95.0 | 100.0 | 100.0 | 512 | 5.2 | 48.6 | 100.0 | 100.0 |
| | 1024 | 5.0 | 100.0 | 100.0 | 100.0 | 1024 | 4.8 | 98.4 | 100.0 | 100.0 |
| | 2048 | 5.1 | 100.0 | 100.0 | 100.0 | 2048 | 4.7 | 100.0 | 100.0 | 100.0 |
| Sub Sampling | 256 | 3.9 | 82.0 | 100.0 | 100.0 | 256 | 5.2 | 8.6 | 100.0 | 100.0 |
| | 512 | 4.7 | 97.0 | 100.0 | 100.0 | 512 | 5.0 | 49.0 | 100.0 | 100.0 |
| | 1024 | 5.0 | 100.0 | 100.0 | 100.0 | 1024 | 4.8 | 98.6 | 100.0 | 100.0 |
| | 2048 | 5.0 | 100.0 | 100.0 | 100.0 | 2048 | 4.9 | 100.0 | 100.0 | 100.0 |
| ROS | 256 | 4.0 | 80.0 | 100.0 | 100.0 | 256 | 5.5 | 8.0 | 100.0 | 100.0 |
| | 512 | 4.9 | 98.0 | 100.0 | 100.0 | 512 | 5.4 | 49.8 | 100.0 | 100.0 |
| | 1024 | 4.7 | 100.0 | 100.0 | 100.0 | 1024 | 4.8 | 99.8 | 100.0 | 100.0 |
| | 2048 | 4.8 | 100.0 | 100.0 | 100.0 | 2048 | 5.1 | 100.0 | 100.0 | 100.0 |

Table 3: Sizes and Powers of PLRT in testing $H_0$ : g is linear when the true function is $g_0(z) = -0.5 + z + c(\sin(\pi z) - 0.5)$ for $c = 0, 0.5, 1, 1.5$, respectively. Significance level is 95%.

Yuri I Ingster.

Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii.

*Math. Methods Statist*, 2(2):85–114, 1993.

Ingo Steinwart, Don R Hush, Clint Scovel, et al.

Optimal rates for regularized least squares regression.

In *COLT*, 2009.

Tong Zhang.

Learning bounds for kernel regression using effective data dimensionality.

*Neural Computation*, 17(9):2077–2098, 2005.