Towards Inference for Kernel Machines

Yair Goldberg

Banff, August 2016

אוניברסיטת חיפה
University of Haifa
جامعة حيفا

Outline

"There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the **data are generated by a given stochastic data model**. The other uses **algorithmic models and treats the data mechanism as unknown.**"

Leo Breiman

Kernels

A function

$$k : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}, \quad \mathcal{Z} \subset \mathbb{R}^d,$$

which is symmetric and positive definite is called a kernel function

Examples

- Linear kernel:

$$k_{\text{Linear}}(z_1, z_2) = z_1^T z_2\,, \quad z_1, z_2 \in \mathcal{Z} \subset \mathbb{R}^d$$

- Gaussian RBF kernel:

$$k_\rho(z_1, z_2) = e^{-\frac{\|z_1 - z_2\|^2}{\rho}}\,, \quad z_1, z_2 \in \mathcal{Z} \subset \mathbb{R}^d$$

Reproducing Kernel Hilbert Spaces

For a kernel k, for every fixed $z_0 \in \mathcal{Z} \subset \mathbb{R}^d$ define the function $k_{z_0}(\cdot)$

$$k_{z_0}(z) = k(z_0, z)$$

A kernel function k is called **reproducing kernel for a Hilbert space** $\mathcal{H}$ if

- $k_{z_0}(\cdot) \in \mathcal{H}$ for all $z_0 \in \mathcal{Z}$.
- The reproducing property holds:

$$h(z_0) = < h, k_{z_0} >, \quad h \in \mathcal{H}, z_0 \in \mathcal{Z}\,.$$

Reproducing Kernel Hilbert Spaces

For a kernel k, for every fixed $z_0 \in \mathcal{Z} \subset \mathbb{R}^d$ define the function $k_{z_0}(\cdot)$

$$k_{z_0}(z) = k(z_0, z)$$

A kernel function k is called **reproducing kernel for a Hilbert space** $\mathcal{H}$ if

- $k_{z_0}(\cdot) \in \mathcal{H}$ for all $z_0 \in \mathcal{Z}$.
- The reproducing property holds:

$$h(z_0) =< h, k_{z_0} >, \quad h \in \mathcal{H}, z_0 \in \mathcal{Z}\,.$$

Reproducing Kernel Hilbert Spaces

The space

$$\mathcal{H}_{\text{pre}} = \left\{ \sum_{i=1}^{n} \alpha_i k_{z_i}(z) \;:\; \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n, z_1, \ldots, z_n \in \mathcal{Z} \right\}$$

with the inner product

$$\left\langle \sum_{i=1}^{n} \alpha_i k_{z_i}(z), \sum_{j=1}^{m} \beta_j k_{z_j}(z) \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(z_i, z_j)$$

is dense in the RKHS defined by the kernel k.

Clearly, the reproducing property holds for $h(z) = \sum_{i=1}^{n} \alpha_i k(z_i, z)$:

$$h(z) \equiv \sum_{i=1}^{n} \alpha_i k_{z_i}(z) = \left\langle \sum_{i=1}^{n} \alpha_i k(z_i, \cdot), k(z, \cdot) \right\rangle$$

Properties of RKHS

Let $\mathcal{H}$ be defined by the Gaussian RBF kernel

$$k_\rho(z_1, z_2) = e^{-\frac{\|z_1 - z_2\|^2}{\rho}}.$$

Assume that $\mathcal{Z} \subset \mathbb{R}^d$ is compact.
Then $\mathcal{H}$ is dense in the $C(\mathcal{Z})$, the class of continuous function on $\mathcal{Z}$.

Kernel Machines (Support Vector Machines)

- Let $D = \{(Z_1, Y_1), \dots, (Z_n, Y_n) : Z_i \in \mathcal{Z}, Y_i \in \mathbb{R}\}$ be n pairs of i.i.d. random vectors.
- The **kernel machine decision function** $h_{D,\lambda}$ is given by

$$h_{D,\lambda} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, h(Z_i)) + \lambda \|h\|_{\mathcal{H}}^2$$

where

  - $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with kernel k,
  - $\lambda > 0$ is a regularization constant
  - L is a loss function.

Kernel machine decision function is the minimizer of a penalized empirical risk problem.

Kernel Machines (Support Vector Machines)

- Let $D = \{(Z_1, Y_1), \ldots, (Z_n, Y_n) : Z_i \in \mathcal{Z}, Y_i \in \mathbb{R}\}$ be n pairs of i.i.d. random vectors.
- The **kernel machine decision function** $h_{D,\lambda}$ is given by

$$h_{D,\lambda} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, h(Z_i)) + \lambda \|h\|_{\mathcal{H}}^2$$

  where
  - $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with kernel k,
  - $\lambda > 0$ is a regularization constant
  - L is a loss function.

Kernel machine decision function is the minimizer of a penalized empirical risk problem.

Examples of Loss Functions

- The hinge loss:

$$L(y, h(z)) = \max\{1 - y \cdot h(z), 0\}, \quad y \in \{-1, 1\}.$$

- The quadratic loss:

$$L(y, h(z)) = (y - h(z))^2.$$

Examples of Loss Functions

- The hinge loss:

$$L(y, h(z)) = \max\{1 - y \cdot h(z), 0\}, \quad y \in \{-1, 1\}.$$

- The quadratic loss:

$$L(y, h(z)) = (y - h(z))^2.$$

The Kernel Trick

- The minimizer

$$h_{D,\lambda} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, h(Z_i)) + \lambda \|h\|_{\mathcal{H}}^2$$

can be written as

$$h_{D,\lambda}(z) = \sum_{i=1}^{n} \alpha_i k_{Z_i}(z) \,.$$

- This representation is referred to as **"the kernel trick"**.
- If the loss L is differentiable,

$$\alpha_i = \frac{\frac{\partial}{\partial_2} L(y_i, h_{D,\lambda}(Z_i))}{n\lambda}$$

Theoretical Results: Universal Consistency

Theorem:
Let

1. $\mathcal{H}$ be a 'large' RKHS.
2. L be a convex Lipschitz continuous loss function.

Choose $0 < \lambda_n < 1$ such that $\lambda_n \to 0$, and $\lambda_n^2 n \to \infty$.
Then the kernel machine method is **universally consistent**:
For every probability measure P,

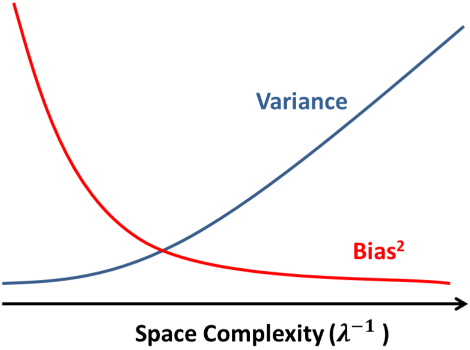
$$E\left[L(Y, h_{D,\lambda_n}(Z))\right] \xrightarrow{\text{P}} \inf_{h \in \mathcal{H}} E\left[L(Y, h(Z))\right] .$$

Theoretical Results: Universal Consistency

An equivalent representation to the kernel machine decision function:

$$h_{D,\lambda} = \underset{h \in \mathcal{H}, \|h\|_{\mathcal{H}}^2 \leq a(\lambda^{-1})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, h(Z_i))$$

where $a(\cdot)$ is some monotonic increasing function.



Least Square Kernel Machines

The **kernel machine decision function**

$$h_{D,\lambda} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - h(Z_i))^2 + \lambda \|h\|_{\mathcal{H}}^2$$

can be derived explicitly

$$\hat{\alpha}_{n \times 1} = (K_{n \times n} + \lambda I_{n \times n})^{-1} Y_{n \times 1}$$

where $K_{ij} = k(Z_i, Z_j) = e^{-\frac{\|Z_i - Z_j\|^2}{\rho}}$.

Question: How to choose

- the kernel bandwidth parameter ρ
- the regularization parameter λ

Least Square Kernel Machines

The **kernel machine decision function**

$$h_{D,\lambda} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - h(Z_i))^2 + \lambda \|h\|_{\mathcal{H}}^2$$

can be derived explicitly

$$\hat{\alpha}_{n\times 1} = (K_{n\times n} + \lambda I_{n\times n})^{-1} Y_{n\times 1}$$

where $K_{ij} = k(Z_i, Z_j) = e^{-\frac{\|Z_i - Z_j\|^2}{\rho}}$.

Question: How to choose

- the kernel bandwidth parameter ρ
- the regularization parameter λ

Semiparametric Least Square Kernel Machines

- Let $D = \{(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n) : X_i \in \mathcal{X} \subset \mathbb{R}^p, Z_i \in \mathcal{Z}, Y_i \in \mathbb{R}\}$ be n triples of i.i.d. random vectors.
- The minimizer of

$$h_{D,\lambda} = \underset{\beta \in \mathbb{R}^p, h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^T X_i - h(Z_i))^2 + \lambda \|h\|_{\mathcal{H}}^2$$

is given by

$$\hat{\beta} = \left\{X^T V^{-1} X\right\}^{-1} X^T V^{-1} Y$$
$$\hat{\alpha} = \lambda^{-1} V^{-1} \left(Y - X\hat{\beta}\right)$$

where $V = (\lambda^{-1} K + I)^{-1}$.

Mixed Effect Model Representation

In this part I follow Liu, Lin, and Ghosh (2007).

Assume the following linear mixed model

$$Y_{n\times1} = X_{n\times p}\beta_{p\times1} + h_{n\times1} + \varepsilon_{n\times1},$$

where

- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$,
- h is random effect with distribution $\mathcal{N}(0, \tau K)$, $\tau = \sigma^2/\lambda$,
- and h and ε are independent.

Note that Z appears implicitly in the variance of h.

Mixed Effect Model Representation

In this part I follow Liu, Lin, and Ghosh (2007).

Assume the following linear mixed model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + h_{n\times 1} + \varepsilon_{n\times 1},$$

where

- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$,
- h is random effect with distribution $\mathcal{N}(0, \tau K)$, $\tau = \sigma^2/\lambda$,
- and h and ε are independent.

Note that Z appears implicitly in the variance of h.

Mixed Effect Model Representation

In this part I follow Liu, Lin, and Ghosh (2007).

Assume the following linear mixed model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + h_{n\times 1} + \varepsilon_{n\times 1},$$

where

- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$,
- h is random effect with distribution $\mathcal{N}(0, \tau K)$, $\tau = \sigma^2/\lambda$,
- and h and ε are independent.

Note that Z appears implicitly in the variance of h.

Bayesian Point of View

Assume the model

$$Y = X\beta + h + \varepsilon,$$

such that

- $y \mid (\beta, h(z)) \sim N\{x^T\beta + h(z), \sigma^2\}$
- $h(\cdot) \sim \mathrm{GP}\{0, \tau k(\cdot,\cdot)\}$
- $\beta \propto 1,$

Minimization Problem

The log posterior density for β and h is (up to a constant)

$$-(Y - X\beta - h)^T(\sigma^2 I)^{-1}(Y - X\beta - h) - h^T(\tau K)^{-1}h\,.$$

Writing $h = K\alpha$, and maximizing the log posterior density is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i + K\alpha)^2 + \alpha^T K\alpha$$

which by the representation theorem is the same as minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i + h(Z_i))^2 + \lambda\|h\|_{\mathcal{H}}^2$$

over all $\beta \in \mathbb{R}^p$ and $h \in \mathcal{H}$

Minimization Problem

The log posterior density for β and h is (up to a constant)

$$-(Y - X\beta - h)^T(\sigma^2 I)^{-1}(Y - X\beta - h) - h^T(\tau K)^{-1}h\,.$$

Writing $h = K\alpha$, and maximizing the log posterior density is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i + K\alpha)^2 + \alpha^T K\alpha$$

which by the representation theorem is the same as minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i + h(Z_i))^2 + \lambda\|h\|_{\mathcal{H}}^2$$

over all $\beta \in \mathbb{R}^p$ and $h \in \mathcal{H}$

LSKM vs LMM

Finding least square kernel machine decision function
is equivalent to
estimation in linear mixed effect model

Question: What do we gain from the mixed model representation?

LSKM vs LMM

Finding least square kernel machine decision function
is equivalent to
estimation in linear mixed effect model

Question: What do we gain from the mixed model representation?

Topic 1: Estimation

We would like to estimate the following parameters:

1. the coefficient vector β,
2. the function $h_{n\times 1} \equiv K_{n\times n}\alpha_{n\times 1}$
3. the noise variance σ^2,
4. the regularization constant λ or equivalently $\tau = \lambda^{-1}\sigma^2$,
5. the kernel bandwidth parameter ρ.

We have $n+p+3$ parameters to estimate and only n observations.

Topic 1: Estimation

We would like to estimate the following parameters:

1. the coefficient vector β,
2. the function $h_{n\times1} \equiv K_{n\times n}\alpha_{n\times1}$
3. the noise variance σ^2,
4. the regularization constant λ or equivalently $\tau = \lambda^{-1}\sigma^2$,
5. the kernel bandwidth parameter ρ.

We have $n + p + 3$ parameters to estimate and only n observations.

Topic 1: Estimation

- Given σ^2, τ, and ρ:
  - Estimation of β and h is done using the log posterior maximization
  - Same estimators as standard kernel machine estimation
- The parameters σ^2, τ, and ρ can be estimated using REML.

Questions:

1. Are these estimators reasonable?
   - Normality was only assumed for mathematical convenience.
   - All the random effects are dependent.
2. Can it replace cross-validation?

Topic 1: Estimation

- Given σ^2, τ, and ρ:
  - Estimation of β and h is done using the log posterior maximization
  - Same estimators as standard kernel machine estimation
- The parameters σ^2, τ, and ρ can be estimated using REML.

Questions:

1. Are these estimators reasonable?
   - Normality was only assumed for mathematical convenience.
   - All the random effects are dependent.
2. Can it replace cross-validation?

Topic 1: Estimation

- Given σ^2, τ, and ρ:
  - Estimation of β and h is done using the log posterior maximization
  - Same estimators as standard kernel machine estimation
- The parameters σ^2, τ, and ρ can be estimated using REML.

Questions:

1. Are these estimators reasonable?
   - Normality was only assumed for mathematical convenience.
   - All the random effects are dependent.
2. Can it replace cross-validation?

Topic 1: Estimation

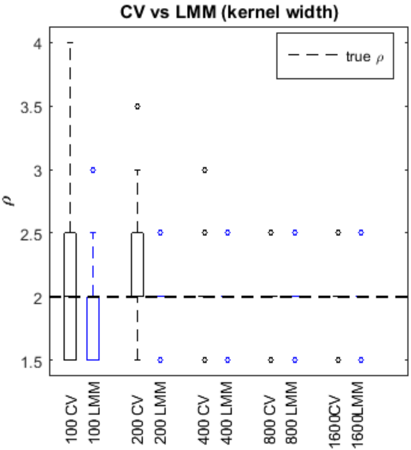
- Given σ^2, τ, and ρ:
  - Estimation of β and h is done using the log posterior maximization
  - Same estimators as standard kernel machine estimation
- The parameters σ^2, τ, and ρ can be estimated using REML.

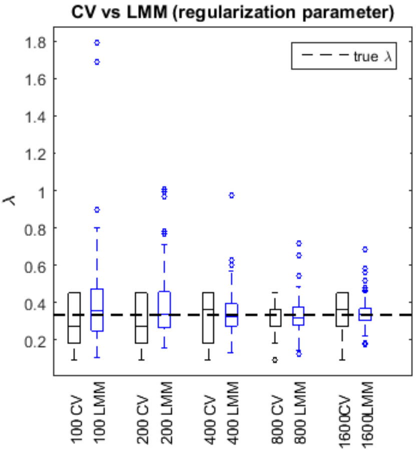
Questions:

1. Are these estimators reasonable?
   - Normality was only assumed for mathematical convenience.
   - All the random effects are dependent.
2. Can it replace cross-validation?

Topic 1: Estimation - Some Simulations

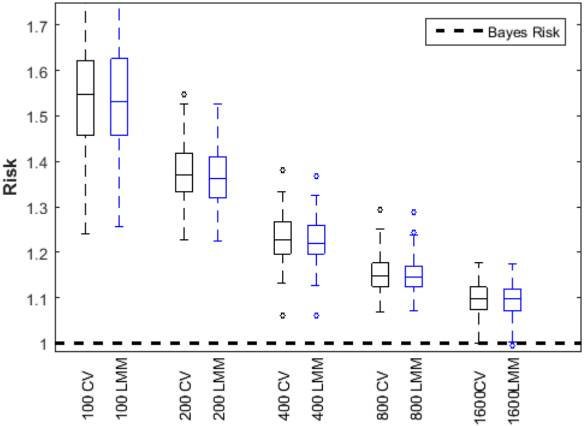
Setting A (Model holds): $h \sim GP\{0, k(\cdot,\cdot)\}$.





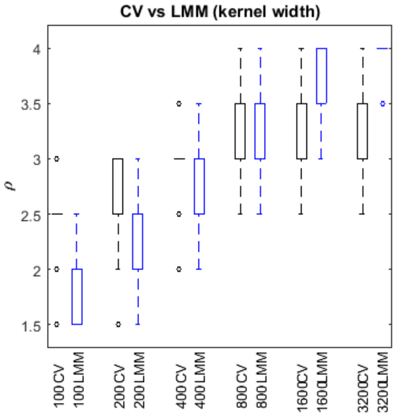
Topic 1: Estimation - Some Simulations

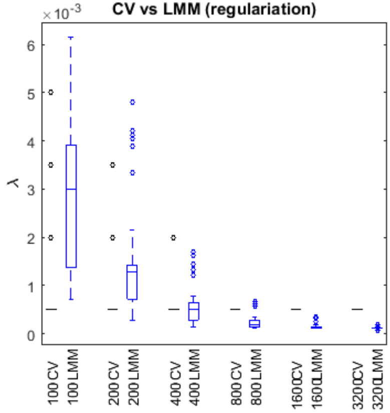
Setting A (Model holds): $h \sim GP\{0, k(\cdot,\cdot)\}$.



Topic 1: Estimation - Some Simulations

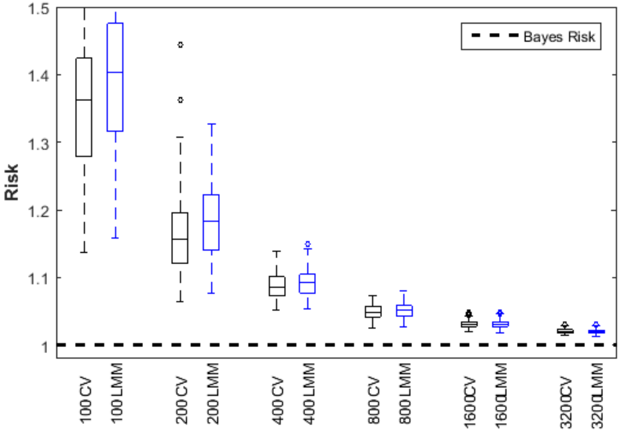
Setting B (h fixed):
$h(Z) = 10\cos(Z_1) - 15Z_2^2 + 10e^{-Z_3 Z_4} - 8\sin(Z_5)\cos(Z_3) + 20Z_1Z_5.$





Topic 1: Estimation - Some Simulations

Setting B (h fixed):
$h(z) = 10\cos(z_1) - 15z_2^2 + 10e^{-z_3 z_4} - 8\sin(z_5)\cos(z_3) + 20z_1 z_5\,, \quad z \in \mathbb{R}^5.$



Topic 1: Estimation

Summary

Simulations seem to work when

- LMM holds (h is random)
- h is fixed but unknown

Problems

1. Does estimation using Linear Mixed Model work for
   - Heteroscedastic noise?
   - Higher dimensions?
2. What about asymptotic convergence for
   - β and h
   - σ^2, λ, and the kernel bandwidth ρ

Topic 1: Estimation

Summary

Simulations seem to work when

- LMM holds (h is random)
- h is fixed but unknown

Problems

1. Does estimation using Linear Mixed Model work for
   - Heteroscedastic noise?
   - Higher dimensions?
2. What about asymptotic convergence for
   - β and h
   - σ^2, λ, and the kernel bandwidth ρ

Topic 2: Variance Estimation

Assume the Bayesian Model $Y = X\beta + h + \varepsilon$, such that

- $y \mid (\beta, h(z)) \sim N\{x^T\beta + h(z), \sigma^2\}$
- $h(\cdot) \sim \mathrm{GP}\{0, \tau k(\cdot,\cdot)\}$
- $\beta \propto 1$,

The variance can be written as

$$\mathrm{Cov}(\hat{\beta}) = (X^T V^{-1} X)^{-1}$$
$$\mathrm{Cov}(\hat{h} - h) = \tau K - (\tau K) P (\tau K).$$

where

$$P = V^{-1} - V^{-1} X \left(X^T V^{-1} X\right)^{-1} X^T V^{-1}, \quad V = \sigma^2 I + \tau K .$$

Topic 2: Variance Estimation

Assume the Frequentist model

$$Y = X\beta + h + \varepsilon,$$

such that

- $y \mid (\beta, h(z)) \sim N\{x^T\beta + h(z), \sigma^2\}$
- h is fixed

The variance can be written as

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2(X^TV^{-1}X)^{-1}X^TV^{-1}V^{-1}X(X^TV^{-1}X)^{-1}$$
$$\mathrm{Cov}(\hat{h}) = \sigma^2(\tau K)P^2(\tau K).$$

where

$$P = V^{-1} - V^{-1}X\left(X^TV^{-1}X\right)^{-1}X^TV^{-1}, \quad V = \sigma^2 I + \tau K\,.$$

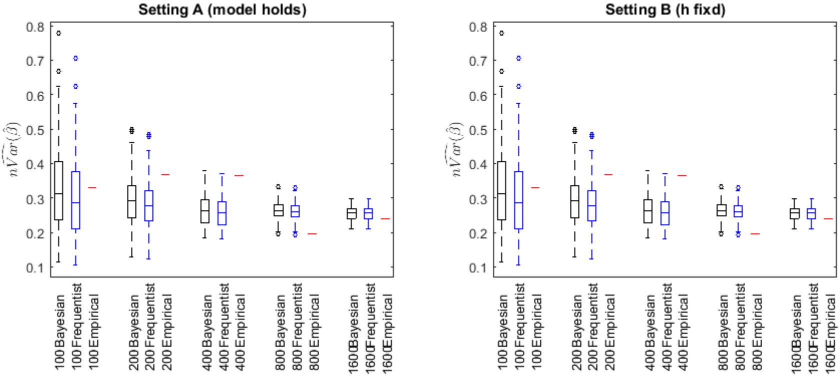
Topic 2: Variance Estimation

Questions

1. Under the Bayesian model, all observations are dependent
   - Does $\text{Var}(\hat{\beta})$ go to zero?
   - Does $\text{Var}(\hat{h})$ go to zero?
2. Which one of the estimators (frequentist vs Bayesian) is better?

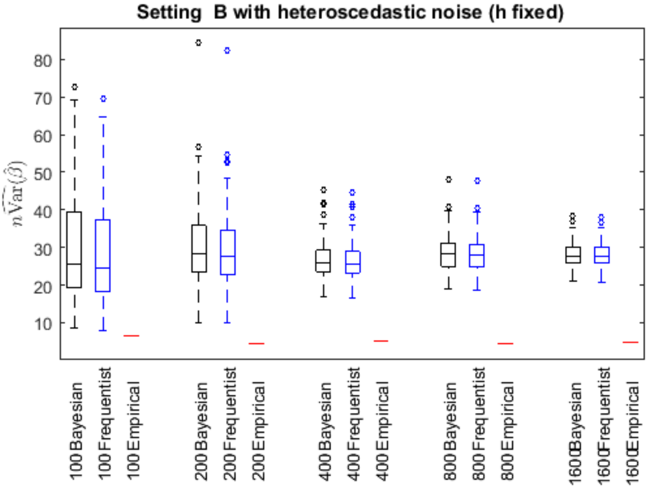
Topic 2: Variance Estimation- Some Simulations

Setting A (model holds).
Setting B (h fixed):
$h(Z) = 10\cos(Z_1) - 15Z_2^2 + 10e^{-Z_3Z_4} - 8\sin(Z_5)\cos(Z_3) + 20Z_1Z_5.$



Topic 2: Variance Estimation- Some Simulations

Setting B (h fixed) with heteroscedastic noise:
$h(Z) = 10\cos(Z_1) - 15Z_2^2 + 10e^{-Z_3Z_4} - 8\sin(Z_5)\cos(Z_3) + 20Z_1Z_5.$



Topic 2: Variance Estimation - Bayesian Model

- Consider the variance of $\hat{h}$
- For simplicity assume Random Effect Model

$$Y = h + \varepsilon$$

- Variance under Bayesian model

$$\mathrm{Cov}(\hat{h} - h) = \tau K - (\tau K) V^{-1} (\tau K) \, .$$

where $V = \tau K + \sigma^2 I$

- Using matrix identities and assuming $\sigma^2 = 1$,

$$\mathrm{Cov}(\hat{h} - h) = I - V^{-1} = I - (I + \lambda^{-1} K)^{-1}.$$

Topic 2: Variance Estimation - Frequentist Model

- Consider the variance of $\hat{h}$
- For simplicity assume random effect model

$$Y = h + \varepsilon$$

- Variance under frequentist model

$$\mathrm{Cov}(\hat{h}) = \sigma^2(\tau K)V^{-2}(\tau K)\,.$$

- Using matrix identities and assuming $\sigma^2 = 1$,

$$\mathrm{Cov}(\hat{h}) = (I + \lambda K^{-1})^{-2}.$$

Topic 3: Confidence Intervals for $h(z)$

- For simplicity assume random effect model

$$Y = h + \varepsilon$$

- Under the Bayesian model

$$\mathrm{Var}\Big(\hat{h}(z) - h(z)\Big) = \tau(1 - \tau K_z V^{-1} K_z)\,,$$

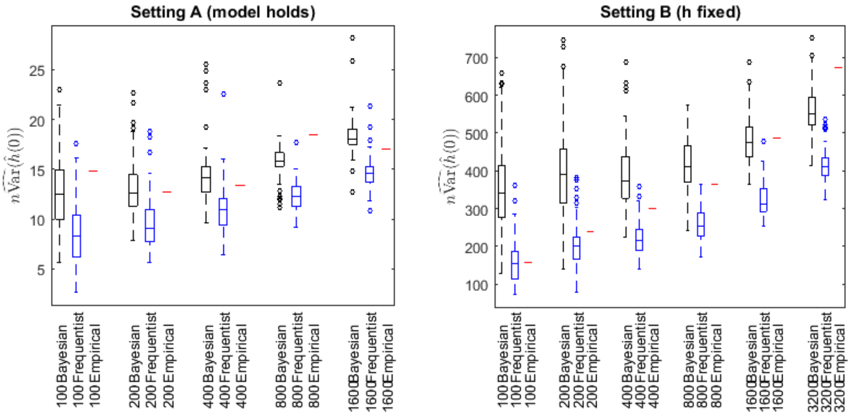
where $K_z = (k_z(Z_1), \dots, k_z(Z_n))^T$.

- Under the frequentist model

$$\mathrm{Var}\Big(\hat{h}(z)\Big) = \sigma^2(\tau K_z) V^{-2}(\tau K_z)\,,$$

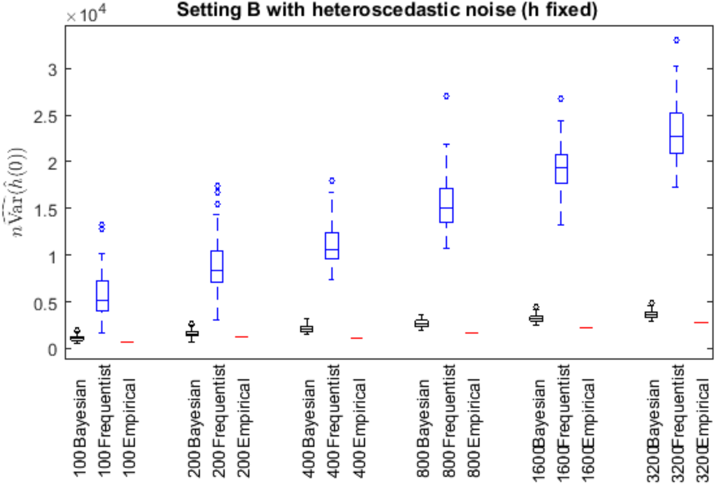
Topic 3: Confidence Intervals for $h(z)$

Setting A (model holds).
Setting B (h fixed):
$h(Z) = 10\cos(Z_1) - 15Z_2^2 + 10e^{-Z_3Z_4} - 8\sin(Z_5)\cos(Z_3) + 20Z_1Z_5$.



Topic 3: Confidence Intervals for $h(z)$

Setting B (h fixed) with heteroscedastic noise:
$h(Z) = 10\cos(Z_1) - 15Z_2^2 + 10e^{-Z_3Z_4} - 8\sin(Z_5)\cos(Z_3) + 20Z_1Z_5$.



Summary

- There is a mathematical connection between Kernel Machines and Mixed Effect Models
- We discussed only least square kernel machines but similar connections were established using Generalized Mixed Effect Models

Questions

- **Estimation:** Can the LMM posterior maximization replace cross validation?
- **Inference for β:** Under which assumption is reliable?
- **Confidence Intervals:** Under which assumptions can they be used?

Comment

Testing for $h \equiv 0$: Shown to work under the null.

Summary

- There is a mathematical connection between Kernel Machines and Mixed Effect Models
- We discussed only least square kernel machines but similar connections were established using Generalized Mixed Effect Models

Questions

- **Estimation:** Can the LMM posterior maximization replace cross validation?
- **Inference for** β: Under which assumption is reliable?
- **Confidence Intervals:** Under which assumptions can they be used?

Comment

Testing for $h \equiv 0$**:** Shown to work under the null.

Summary

- There is a mathematical connection between Kernel Machines and Mixed Effect Models
- We discussed only least square kernel machines but similar connections were established using Generalized Mixed Effect Models

Questions

- **Estimation:** Can the LMM posterior maximization replace cross validation?
- **Inference for** β: Under which assumption is reliable?
- **Confidence Intervals:** Under which assumptions can they be used?

Comment

Testing for $h \equiv 0$**:** Shown to work under the null.

"**Notions of significance tests, confidence intervals, posterior intervals** and all the formal apparatus of inference are valuable tools to be used as guides, but not in a mechanical way; t**hey indicate the uncertainty that would apply under somewhat idealized, maybe very idealized, conditions** and as such are often lower bounds to real uncertainty."

D. R. Cox

Towards Inference for Kernel Machines
Magic or Illusion?

Special thanks to

- Yael Travis-Lumer (University of Haifa)
- Malka Gorfine (Tel-Aviv University)
- Yanyuan Ma (Pennsylvania State University)

Thank you all for listening.