Robust Test for Nonlinear Interaction using Cross-Validated Kernel Ensemble

Wenying Deng, Jeremiah Zhe Liu
Department of Biostatistics
Harvard University

Biostat Seminar
July 23, 2018

Contents

Background

Consider a nutrition-environment interaction study for continuous infant health outcome y_i. For observation i, we have:

1. μ: the fixed effect of background covariates, assuming the same across all observations.
2. $\mathbf{x}_{1,i}$: $2 \sim 10$ nutrients variables (e.g. Vitamin, Folate, etc)
3. $\mathbf{x}_{2,i}$: $2 \sim 10$ environmental exposure (e.g. $PM_{2.5}$, pesticides, etc)

Effect of $\mathbf{x}_{1,i}$ and $\mathbf{x}_{2,i}$ on y_i is nonlinear.

We observe $n \approx 100$ such records, and want to investigate whether mother's nutrients intake during pregnancy $\mathbf{x}_1$ modifies the effects of prenatal exposures to $\mathbf{x}_2$, i.e. nonlinear interaction between $\mathbf{x}_1$ and $\mathbf{x}_2$.

Problem Setup

- True Model

$$y_i = \mu + h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where

$$h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Big[h_1(\mathbf{x}_{1i}) + h_2(\mathbf{x}_{2i})\Big] + h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i})$$

Problem Setup

- True Model

$$y_i = \mu + h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where

$$h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Big[h_1(\mathbf{x}_{1i}) + h_2(\mathbf{x}_{2i})\Big] + h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i})$$

  - $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ are the **main effect functions**

Problem Setup

- True Model

$$y_i = \mu + h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where

$$h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Big[h_1(\mathbf{x}_{1i}) + h_2(\mathbf{x}_{2i})\Big] + h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i})$$

  - $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ are the **main effect functions**
  - $h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \in \mathcal{H}_{12}$ is the **pure interaction function**

Problem Setup

- True Model

$$y_i = \mu + h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where

$$h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Big[h_1(\mathbf{x}_{1i}) + h_2(\mathbf{x}_{2i})\Big] + h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i})$$

  - $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ are the **main effect functions**
  - $h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \in \mathcal{H}_{12}$ is the **pure interaction function**
    - $h_{12} \perp h_1$ and $h_{12} \perp h_2$.

Problem Setup

- True Model

$$y_i = \mu + h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where

$$h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Big[h_1(\mathbf{x}_{1i}) + h_2(\mathbf{x}_{2i})\Big] + h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i})$$

  - $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ are the **main effect functions**
  - $h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \in \mathcal{H}_{12}$ is the **pure interaction function**
    - $h_{12} \perp h_1$ and $h_{12} \perp h_2$.
  - $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_{12}$ are **UNKNOWN**.

Hypothesis

- Model

$$y_i = \mu + h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \epsilon_i$$

where

$$h(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Big[h_1(\mathbf{x}_{1i}) + h_2(\mathbf{x}_{2i})\Big] + h_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i})$$

- Hypothesis

$$H_0 : h \in \mathcal{H}_0 = \mathcal{H}_1 \oplus \mathcal{H}_2$$
$$H_a : h \in \mathcal{H}_a = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_{12}$$

Overview

- **Assumption**: Given a library of kernels $\{k_d(\mathbf{x}, \mathbf{x}')\}_{d=1}^{D}$, assume

$$h_0(\mathbf{x}) = \sum_{d=1}^{D} u_d h_d(\mathbf{x}), \qquad \sum_{d=1}^{D} u_d = 1, u_d > 0$$

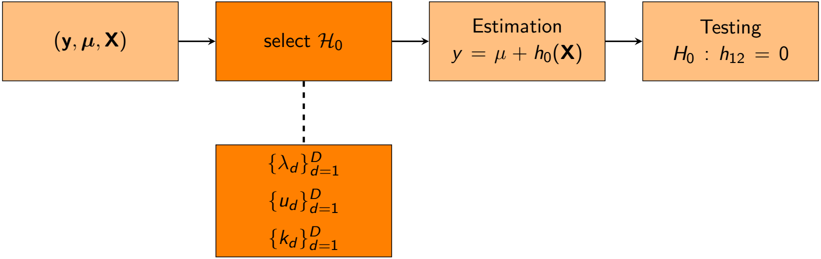
where $h_d \in \mathcal{H}_d$, the function space corresponds to $k_d(\mathbf{x}, \mathbf{x}')$.

Overview

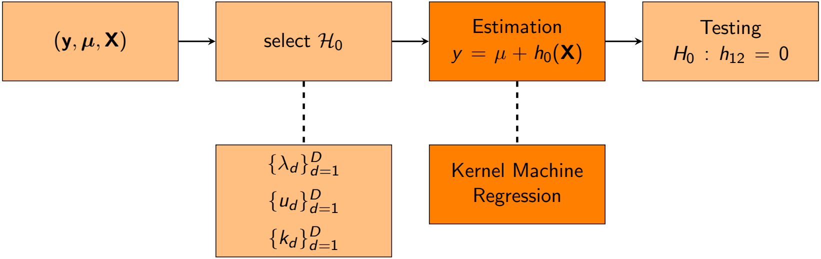
- **Model**:

$$
\begin{aligned}
\mathbf{y} &= \boldsymbol{\mu} + h_0(\mathbf{x}) \\
&= \boldsymbol{\mu} + \sum_{d=1}^{D} u_d h_d(\mathbf{x}) \\
&= \boldsymbol{\mu} + \sum_{d=1}^{D} u_d \mathbf{K}_d \boldsymbol{\alpha}_d
\end{aligned}
$$

Overview

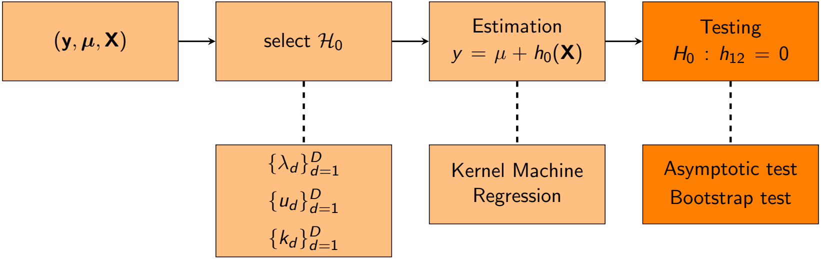

$(\mathbf{y}, \boldsymbol{\mu}, \mathbf{X})$
select $\mathcal{H}_0$
Estimation
$y = \mu + h_0(\mathbf{X})$
Testing
$H_0 : h_{12} = 0$
$\{\lambda_d\}_{d=1}^D$
$\{u_d\}_{d=1}^D$
$\{k_d\}_{d=1}^D$

Overview



Estimates

Overview


$(\mathbf{y}, \boldsymbol{\mu}, \mathbf{X})$
select $\mathcal{H}_0$
Estimation
$y = \mu + h_0(\mathbf{X})$
Testing
$H_0 : h_{12} = 0$
$\{\lambda_d\}_{d=1}^D$
$\{u_d\}_{d=1}^D$
$\{k_d\}_{d=1}^D$
Kernel Machine Regression
Asymptotic test
Bootstrap test

Denote

$$\mathbf{A}_\lambda = \mathbf{K}(\mathbf{X}, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I}]^{-1}$$

and

$$\mathbf{y}^\star = \mathbf{y} - \hat{\boldsymbol{\mu}}, \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- $tr(\mathbf{A}_\lambda)$ is the effective number of model parameters
- $\mathbf{y}^\star$ is centered

- LooCV: leave-one-out Cross Validation

$$\underset{\lambda \in \Lambda}{argmin} \Big\{ log\ \mathbf{y}^{\star T}[\mathbf{I} - diag(\mathbf{A}_\lambda) - \frac{1}{n}\mathbf{I}]^{-1}(\mathbf{I} - \mathbf{A}_\lambda)^2[\mathbf{I} - diag(\mathbf{A}_\lambda) - \frac{1}{n}\mathbf{I}]^{-1}\mathbf{y}^\star \Big\}$$

- AICc: small sample size version of AIC

$$\underset{\lambda \in \Lambda}{argmin}\Big\{ log\ \mathbf{y}^{\star T}(\mathbf{I} - \mathbf{A}_\lambda)^2\mathbf{y}^\star + \frac{2[tr(\mathbf{A}_\lambda) + 2]}{n - tr(\mathbf{A}_\lambda) - 3}\Big\}$$

- GCVc: small sample size version of GCV

$$\underset{\lambda \in \Lambda}{argmin}\Big\{ log\ \mathbf{y}^{\star T}(\mathbf{I} - \mathbf{A}_\lambda)^2\mathbf{y}^\star - 2log[1 - \frac{tr(\mathbf{A}_\lambda)}{n} - \frac{2}{n}]_+\Big\}$$

- GMPML: Generalized Maximum Profile Marginal Likelihood

$$\underset{\lambda \in \Lambda}{argmin}\Big\{ log\ \mathbf{y}^{\star T}(\mathbf{I} - \mathbf{A}_\lambda)\mathbf{y}^\star - \frac{1}{n-1} log\ | \ \mathbf{I} - \mathbf{A}_\lambda \ | \Big\}$$

Note: λ_{AIC} is always smaller than λ_{GCV}. Derive

- ERM: Empirical Risk Minimization

$$\hat{\mathbf{u}} = \underset{\mathbf{u} \in \Delta}{argmin} \parallel \sum_{d=1}^{D} u_d \hat{\epsilon}_d \parallel^2 \quad where\ \Delta = \{\mathbf{u} \mid \mathbf{u} \geq 0, \parallel \mathbf{u} \parallel_1 = 1\}$$

- AVE: Simple Averaging

$$u_d = 1/D \quad \text{for} \quad d = 1, 2, ...D$$

- EXP: Exponential Weighting

$$u_d(\beta) = \frac{exp(- \parallel \hat{\epsilon}_d \parallel_2^2 / \beta)}{\sum_{d=1}^{D} exp(- \parallel \hat{\epsilon}_d \parallel_2^2 / \beta)} \quad \text{for} \quad d = 1, 2, ...D$$

Then produce the final ensemble prediction:

$$\hat{\mathbf{h}}_0 = \sum_{d=1}^{D} \hat{u}_d \mathbf{h}_d = \sum_{d=1}^{D} \hat{u}_d \mathbf{A}_{d,\hat{\lambda}_d} \mathbf{y}^\star = \hat{\mathbf{A}} \mathbf{y}^\star$$

where $\hat{\mathbf{A}} = \sum_{d=1}^{D} \hat{u}_d \mathbf{A}_{d,\hat{\lambda}_d}$ is the ensemble matrix.

- Polynomial Kernel: polynomial functions

$$k(\mathbf{x}, \mathbf{x}') = (b + \mathbf{x}^T\mathbf{x}')^p$$

- Gaussian Kernel: infinitely differentiable functions

$$k(\mathbf{x}, \mathbf{x}') = exp\Big(-\frac{\mid \mathbf{x} - \mathbf{x}' \mid^2}{2l^2}\Big)$$

- Matérn 1/2 Kernel: continuous, non-differentiable functions

$$k(\mathbf{x}, \mathbf{x}') = exp(-\frac{\mid \mathbf{x} - \mathbf{x}' \mid}{l})$$

- Matérn 5/2 Kernel: twice-differentiable functions

$$k(\mathbf{x}, \mathbf{x}') = (1 + \frac{\sqrt{5} \mid \mathbf{x} - \mathbf{x}' \mid}{l} + \frac{5 \mid \mathbf{x} - \mathbf{x}' \mid^2}{3l^2})exp(-\frac{\sqrt{5} \mid \mathbf{x} - \mathbf{x}' \mid}{l})$$

Variance Component Test for Interaction

Translate Hypothesis: Under LMM representation:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{h} + \boldsymbol{\epsilon} \qquad \text{where} \qquad \mathbf{h} \sim N(\mathbf{0}, \tau \mathbf{K}) \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

- Under H_0: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, same as $k = k_1 + k_2 \quad (\Rightarrow) \quad \mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2$

Variance Component Test for Interaction

Translate Hypothesis: Under LMM representation:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{h} + \boldsymbol{\epsilon} \qquad \text{where} \qquad \mathbf{h} \sim N(\mathbf{0}, \tau\mathbf{K}) \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

- Under H_0: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, same as $k = k_1 + k_2 \quad (\Rightarrow) \quad \mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2$
- Under H_a: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_{12}$, same as $k = k_1 + k_2 + k_{12} \quad (\Rightarrow) \quad \mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_{12}$

Variance Component Test for Interaction

Translate Hypothesis: Under LMM representation:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{h} + \boldsymbol{\epsilon} \qquad \text{where} \qquad \mathbf{h} \sim N(\mathbf{0}, \tau\mathbf{K}) \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

- Under H_0: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, same as $k = k_1 + k_2 \quad (\Rightarrow) \quad \mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2$
- Under H_a: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_{12}$, same as $k = k_1 + k_2 + k_{12} \quad (\Rightarrow) \quad \mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_{12}$
- So write $\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2 + \delta * \mathbf{K}_{12}$

Variance Component Test for Interaction

Translate Hypothesis: Under LMM representation:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{h} + \boldsymbol{\epsilon} \qquad \text{where} \qquad \mathbf{h} \sim N(\mathbf{0}, \tau\mathbf{K}) \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

- Under H_0: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, same as $k = k_1 + k_2 \quad (\Rightarrow) \quad \mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2$
- Under H_a: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_{12}$, same as $k = k_1 + k_2 + k_{12} \quad (\Rightarrow) \quad \mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_{12}$
- So write $\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2 + \delta * \mathbf{K}_{12}$
- Test for $H_0 : \delta = 0$ using VCT Detail

What if our sample size is pretty small?

How to make full use of our limited sample?

What if our sample size is pretty small?

Bootstrap!

What if our sample size is pretty small?

A good approximation to the distribution of the test statistic under sampling from the true null-hypothesis model is the distribution of the test statistic under sampling from the fitted null-hypothesis model.

1. Obtain parameter estimates from the original data by fitting a null-hypothesis model

$$E(\mathbf{y}^{\star}) = \mathbf{K}_0(\mathbf{K}_0 + \lambda\mathbf{I})^{-1}\mathbf{y} = \mathbf{A}_0\mathbf{y}$$

1. Obtain parameter estimates from the original data by fitting a null-hypothesis model

$$E(\mathbf{y}^{\star}) = \mathbf{K}_0(\mathbf{K}_0 + \lambda \mathbf{I})^{-1}\mathbf{y} = \mathbf{A}_0\mathbf{y}$$

2. Sample $\mathbf{Y}^{\star}$ for each individuals with a random noise, whose variance is also estimated.

1. Obtain parameter estimates from the original data by fitting a null-hypothesis model

$$E(\mathbf{y}^{\star}) = \mathbf{K}_0(\mathbf{K}_0 + \lambda \mathbf{I})^{-1}\mathbf{y} = \mathbf{A}_0\mathbf{y}$$

2. Sample $\mathbf{Y}^{\star}$ for each individuals with a random noise, whose variance is also estimated.
3. Compute the test statistic, based on fitting the alternative-hypothesis model to the samples obtained in Step 2.

1. Obtain parameter estimates from the original data by fitting a null-hypothesis model

$$E(\mathbf{y}^{\star}) = \mathbf{K}_0(\mathbf{K}_0 + \lambda \mathbf{I})^{-1}\mathbf{y} = \mathbf{A}_0\mathbf{y}$$

2. Sample $\mathbf{Y}^{\star}$ for each individuals with a random noise, whose variance is also estimated.
3. Compute the test statistic, based on fitting the alternative-hypothesis model to the samples obtained in Step 2.
4. Repeat Steps 2 and 3 for B times, to obtain an approximate distribution of the test statistic.

1. Obtain parameter estimates from the original data by fitting a null-hypothesis model

$$E(\mathbf{y}^\star) = \mathbf{K}_0(\mathbf{K}_0 + \lambda\mathbf{I})^{-1}\mathbf{y} = \mathbf{A}_0\mathbf{y}$$

2. Sample $\mathbf{Y}^\star$ for each individuals with a random noise, whose variance is also estimated.
3. Compute the test statistic, based on fitting the alternative-hypothesis model to the samples obtained in Step 2.
4. Repeat Steps 2 and 3 for B times, to obtain an approximate distribution of the test statistic.
5. Compute the test statistic for the original data, based on fitting the alternative- hypothesis model.

1. Obtain parameter estimates from the original data by fitting a null-hypothesis model

$$E(\mathbf{y}^{\star}) = \mathbf{K}_0(\mathbf{K}_0 + \lambda\mathbf{I})^{-1}\mathbf{y} = \mathbf{A}_0\mathbf{y}$$

2. Sample $\mathbf{Y}^{\star}$ for each individuals with a random noise, whose variance is also estimated.
3. Compute the test statistic, based on fitting the alternative-hypothesis model to the samples obtained in Step 2.
4. Repeat Steps 2 and 3 for B times, to obtain an approximate distribution of the test statistic.
5. Compute the test statistic for the original data, based on fitting the alternative- hypothesis model.
6. Compute the p-value, by comparing the test statistic in Step 5 to the distribution in Step 4.

Generate data under

$$y = \mu + h_1(\mathbf{x}_1) + h_2(\mathbf{x}_2) + \delta * h_1(\mathbf{x}_1) * h_2(\mathbf{x}_2) + \epsilon$$

- Vary $\delta \in \{0, 0.1, 0.2, 0.3\}$;
- 3 polynomial kernels ($p = 1, 2, 3$) representing finite-dimensional, parametric functions of different degree of nonlinearity;
- 3 Gaussian RBF kernels ($l = 0.5, 1, 1.5$) representing smooth functions of different frequency;
- 6 Matern kernels, with $l \in \{0.5, 1, 1.5\}$ and $\nu \in \{\frac{3}{2}, \frac{5}{2}\}$, representing functions with different frequency and differentiability.

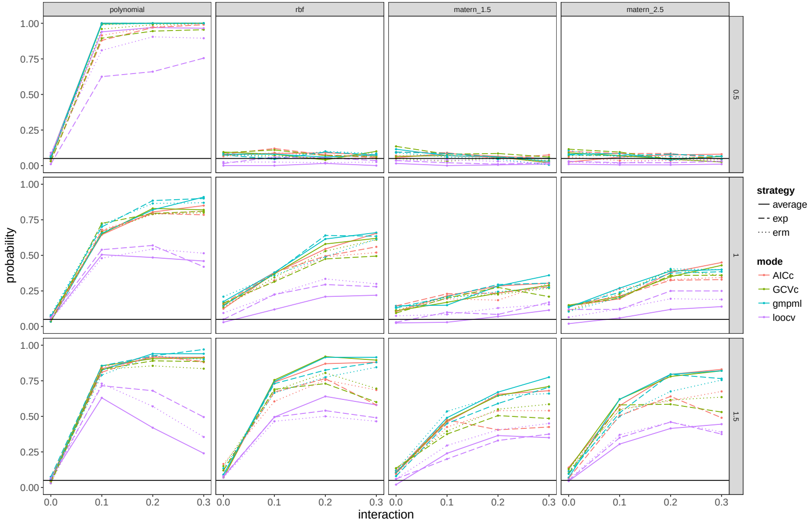
Fit null model:

$$y = \mu + h_1(\mathbf{x}_1) + h_2(\mathbf{x}_2)$$

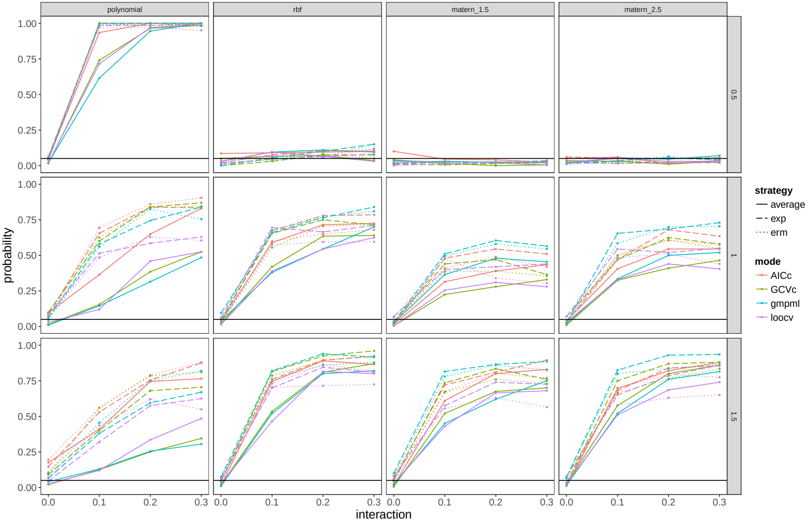
using below kernels:

1. 3 polynomial kernels with degree ($p = 1, 2, 3$); Fig1
2. 3 RBF kernels with wavelength ($l = 0.6, 1, 2$); Fig2
3. 3 polynomial kernels ($p = 1, 2, 3$) and 3 RBF kernels ($l = 0.6, 1, 2$); Fig3
4. 3 $l = 1$ Matern kernels ($\nu = 1/2, 3/2, 5/2$) and 3 RBF kernels ($l = 0.6, 1, 2$). Fig4

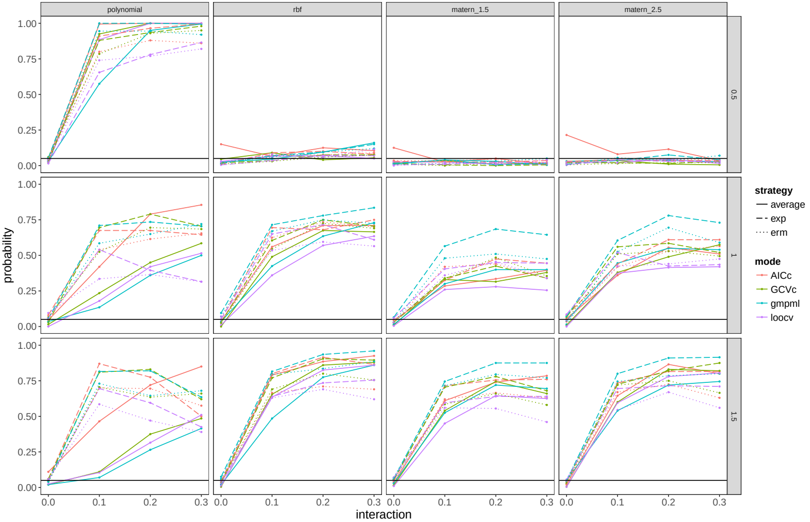
Result


polynomial
rbf
matern_1.5
matern_2.5
0.5
1
1.5
probability
interaction
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
strategy
average
exp
erm
mode
AICc
GCVc
gmpml
loocv

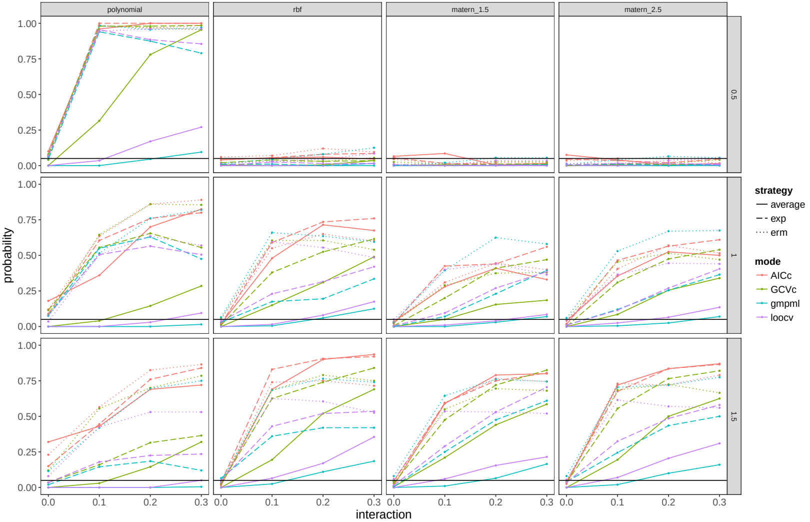
Result


polynomial
rbf
matern_1.5
matern_2.5
0.5
1
1.5
probability
interaction
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
strategy
average
exp
erm
mode
AICc
GCVc
gmpml
loocv

Result


polynomial
rbf
matern_1.5
matern_2.5
0.5
1
1.5
probability
interaction
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
strategy
average
exp
erm
mode
AICc
GCVc
gmpml
loocv

Result


polynomial
rbf
matern_1.5
matern_2.5
0.5
1
1.5
probability
interaction
0.00
0.25
0.50
0.75
1.00
0.0
0.1
0.2
0.3
strategy
average
exp
erm
mode
AICc
GCVc
gmpml
loocv

Main Message: Tuning Parameter Selection

In general,

- LooCV, guarantees correct Type I error, but potentially leads to low power; ▸ Fig4
- AICc, powerful if base kernels are as or more complex than the true one; ▸ Fig4
- GMPML, powerful if base kernels are smoother than the true one. ▸ Fig1

Main Message: Ensemble Strategy

- AVE is better if base kernels are simple and finite-dimensional; ▸ Fig1
- ERM is better if base kernels are flexible and infinite-dimensional; ▸ Fig2 ▸ Fig4
- In the more complex scenario... ▸ Fig3
- EXP, fairly greater power except when the true kernel is strictly simpler than base kernels. ▸ Fig2

Questions?

KMR as a Linear Mixed Model

- KMR Estimates:

$$\hat{\mu} = (\mathbf{1}^T(\mathbf{K_0} + \lambda * \mathbf{I})^{-1}\mathbf{1})^{-1}\mathbf{1}^T(\mathbf{K_0} + \lambda * \mathbf{I})^{-1}$$
$$\hat{\boldsymbol{\alpha}} = (\mathbf{K}_0 + \lambda * \mathbf{I})^{-1}(\mathbf{y} - \mathbf{1} * \hat{\mu})$$
$$\hat{\mathbf{h}}_0 = \mathbf{K}_0\hat{\boldsymbol{\alpha}} = \mathbf{K}_0(\mathbf{K}_0 + \lambda * \mathbf{I})^{-1}(\mathbf{y} - \mathbf{1} * \hat{\mu})$$

- Same solution as a LMM with random intercept **h**:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{h} + \boldsymbol{\epsilon} \qquad \text{where} \qquad \mathbf{h} \sim N(\mathbf{0}, \tau\mathbf{K}) \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

  where $\sigma^2/\tau = \lambda$
- Estimate variance component parameters (τ, σ^2) using REML

Back

If we denote $\mathbf{U}_K$ and $\{\eta_{K,j}\}_{j=1}^n$ as the eigenvector and eigenvalues of $\mathbf{K}$, then $\mathbf{A}_\lambda$ adopts the form:

$$\mathbf{A}_\lambda = \mathbf{U}_K diag(\frac{\eta_{K,j}}{\eta_{K,j}+\lambda})\mathbf{U}_K^T = \mathbf{U}_K \mathbf{D}_{K,\lambda} \mathbf{U}_K^T$$

Calculate the derivatives of the objective functions with respect to λ respectively,

$$\frac{\partial f_{AIC}}{\partial \lambda} = \frac{2tr\left[\mathbf{U}_K^T \mathbf{y}^\star \mathbf{y}^{\star T} \mathbf{U}_K (\mathbf{D}_{K,\lambda} - 1)\frac{\partial \mathbf{D}_{K,\lambda}}{\partial \lambda}\right]}{tr[\mathbf{U}_K^T \mathbf{y}^\star \mathbf{y}^{\star T} \mathbf{U}_K (\mathbf{I} - \mathbf{D}_{K,\lambda})^2]} + \frac{2tr\left(\frac{\partial \mathbf{D}_{K,\lambda}}{\partial \lambda}\right)}{n} \quad (1)$$

$$\frac{\partial f_{GCV}}{\partial \lambda} = \frac{2tr\left[\mathbf{U}_K^T \mathbf{y}^\star \mathbf{y}^{\star T} \mathbf{U}_K (\mathbf{D}_{K,\lambda} - 1)\frac{\partial \mathbf{D}_{K,\lambda}}{\partial \lambda}\right]}{tr[\mathbf{U}_K^T \mathbf{y}^\star \mathbf{y}^{\star T} \mathbf{U}_K (\mathbf{I} - \mathbf{D}_{K,\lambda})^2]} + \frac{2tr\left(\frac{\partial \mathbf{D}_{K,\lambda}}{\partial \lambda}\right)}{n - tr(\mathbf{A}_\lambda) - 1} \quad (2)$$

Notice for j^{th} element of diagonal vector of $\mathbf{D}_{K,\lambda}$, its derivative with respect to λ is negative,

$$\frac{\partial}{\partial \lambda}\left[\frac{\eta_{K,j}}{\eta_{K,j}+\lambda}\right] = -\frac{\eta_{K,j}}{(\eta_{K,j}+\lambda)^2} < 0, \quad for\ j = 1,\ 2, ...,\ n$$

Further notice that the difference between (1) and (2) focuses on the second terms, both of which are increasing function of λ.

$$\frac{2tr\left(\frac{\partial \mathbf{D}_{K,\lambda}}{\partial \lambda}\right)}{n - tr(\mathbf{A}_\lambda) - 1} < \frac{2tr\left(\frac{\partial \mathbf{D}_{K,\lambda}}{\partial \lambda}\right)}{n}$$

Therefore, when $\frac{\partial f_{AIC}}{\partial \lambda} = 0$, $\frac{\partial f_{GCV}}{\partial \lambda} < 0$, which means $\lambda_{AIC} < \lambda_{GCV}$.

Back

Variance Component Test Detail

To perform Variance Component (Score) Test for $H_0 : \delta = 0$:

1. **Obtain test statistic** $\hat{T}_0$:
   1. by taking REML derivative with respect to δ
   2. Final Expression of the test statistic:

$$\hat{T}_0 = \hat{\tau} * (\mathbf{y} - \hat{\boldsymbol{\mu}})^T \mathbf{V}_0^{-1} \, \mathbf{K}_{12} \, \mathbf{V}_0^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}})$$

2. **Obtain null distribution**:
   - Since $\hat{T}_0$ is of quadratic form, null distribution is a mixture of χ^2_1
   - Approximate null distribution F_0 using Satterthwaite
3. **Calculate p value**: $P = 1 - F_0(\hat{T}_0)$ Back