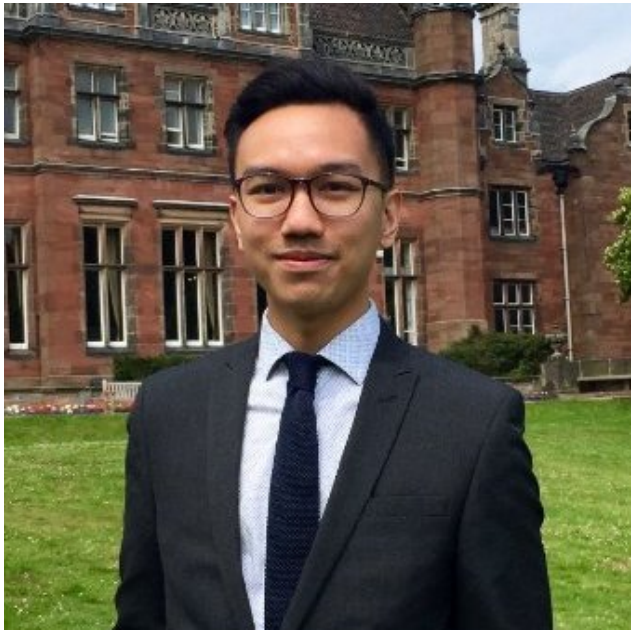
Selecting interaction effects using I-priors

Wicher Bergsma & Haziq Jamil

London School of Economics

Co-worker



Haziq Jamil (Universiti Brunei Darussalam)
github.com/haziqj

Outline

Regression with linear functions

Example (Salary explained by age, gender, ethnicity)

Saturated model:

$$y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{1i}x_{2i}\beta_{12} + \ldots + x_{1i}x_{2i}x_{3i}\beta_{123} + \varepsilon_i$$

Two-way interaction effect of gender and age: salary growth different between men and women.
Three way interaction effect of gender, age and ethnicity: effect of gender on salary growth depends on ethnicity.

- Hierarchical model: if an effect is omitted, so are the corresponding higher order effects, e.g., $y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{2i}x_{3i}\beta_{23} + \varepsilon_i$
- Which hierarchical model best fits the data?

Common selection methods:

- Stepwise selection using ML estimation
- Lasso
- Spike and slab priors on βs

Regression with nonlinear functions

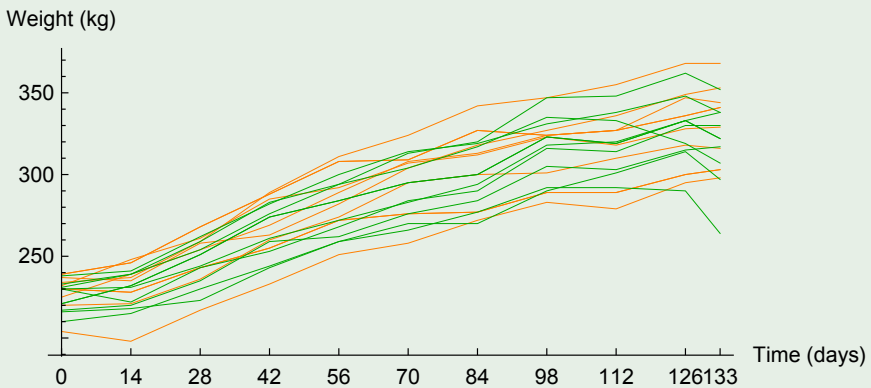
Saturated model with 3 arbitrary covariates:

$$y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_{12}(x_{1i}, x_{2i}) + \ldots + f_{123}(x_{1i}, x_{2i}, x_{3i}) + \varepsilon_i$$

Common estimation methods: Tikhonov regularization, Gaussian process regression. Model selection not straightforward.

Example (cow growth)

Sample of 8 (out of 60) growth curves of cows, two treatments:



Response: weight; covariates: time, treatment, cow index
(alternatively: response: growth curve; covariates treatment and cow index).
Does treatment affect growth? If so, does treatment affect cows differentially?

Outline

Outline definition

Model: $p(x|\theta) \propto e^{\theta(x)}$, $\theta \in \Theta$, Θ RKKS (defined below).

Observations $x_1, \ldots, x_n \sim_{\text{iid}} p(\cdot|\theta_0)$.

Fisher information for θ evaluated at a fixed θ^* is positive definite and hence defines RKHS $(\Theta_n, \|\cdot\|_{\Theta_n})$ on (at most n-dimensional) subspace of Θ.

I-prior for θ: prior maximizing entropy subject to $\|\theta\|_{\Theta_n} =$ constant.

The I-prior can be shown to be *proper* (with probability 1, realizations are in Θ).

I-prior for regression function

With $\mathcal{F}$ the RKKS (defined below) with reproducing kernel h on a set $\mathcal{X}$, suppose

$$y_i = f_0(x_i) + \varepsilon_i = \langle f_0, h(x_i, \cdot)\rangle_{\mathcal{F}} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $(\varepsilon_1, \ldots, \varepsilon_n) \sim \text{MVN}(0, \Psi^{-1})$.

The I-prior for f_0 is then given by:

$$f(x) = \sum h(x, x_i) w_i, \quad (w_1, \ldots, w_n) \sim \text{MVN}(0, \Psi)$$

Loosely: the more difficult an f would be to estimate, the lower its prior mass. Or, the more information on a linear functional of f, the larger its prior variance, and the smaller the influence of the prior mean on the posterior mean (and vice versa).

Why I-priors?

- I-prior is proper, hence posterior mean *admissible* by Wald's complete class theorem.
- I-prior function of likelihood, no further user choices needed.
- EM algorithm with simple E and M steps available, because normalizing constant cancels in complete data likelihood.

Main competitors:

- Tikhonov regularizer: *inadmissible* in infinite dimensions (Chakraborty & Panaretos, 2019). E.g., cubic spline smoother,

$$\hat{f}_n = \arg\min \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \int \ddot{f}(x)^2 dx,$$

  has suboptimal convergence rate to every true f with $\int \ddot{f}(x)^2 dx < \infty$.
- Gaussian process regression:
  - Requires user to choose (i) space in which regression function lives, and (ii) a prior over that space.
  - Complex M step in EM algorithm.

Outline

Reproducing kernel Krein spaces

Definition (Krein space)

A vector space $\mathcal{F}$ equipped with the inner product $\langle \cdot, \cdot, \rangle_{\mathcal{F}}$ is called a *Krein space* if there are two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$ spanning $\mathcal{F}$ such that

- All $f \in \mathcal{F}$ can be decomposed as $f = f_+ + f_-$ where $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$.
- For all $f, f' \in \mathcal{F}$, $\langle f, f' \rangle_{\mathcal{F}} = \langle f_+, f'_+ \rangle_{\mathcal{H}_+} - \langle f_-, f'_- \rangle_{\mathcal{H}_-}$.

Definition (RKKS)

A Krein space of functions is called a *reproducing kernel Krein space* (RKKS) if the point evaluator is continuous.

Theorem (reproducing kernel)

Every RKKS $\mathcal{F}$ of real-valued functions on a set $\mathcal{X}$ possesses a unique symmetric reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $f \in \mathcal{F}$, $x \in \mathcal{X}$,

- *$h(x, \cdot) \in \mathcal{F}$ (basis functions)*
- *$f(x) = \langle h(x, \cdot), f \rangle_{\mathcal{F}}$ (reproducing property)*

Illustration hierarchical interaction spaces

$$y_i = \underbrace{\alpha + f_1(x_{1i}) + f_2(x_{2i}) + f_{12}(x_{1i}, x_{2i})}_{f(x_{1i}, x_{2i})} + \varepsilon_i$$

x_{1i}, x_{2i} in arbitrary sets $\mathcal{X}_1$, $\mathcal{X}_2$, resp.
$\mathcal{C}_1$, $\mathcal{C}_2$ set of constant functions on $\mathcal{X}_1$, $\mathcal{X}_2$, resp.
$\mathcal{F}_1$, $\mathcal{F}_2$ vector spaces of functions on $\mathcal{X}_1$, $\mathcal{X}_2$, e.g.,

$$\mathcal{F}_1 = \{f : \mathbb{R} \to \mathbb{R} \,|\, f(x) = x\beta\}$$

$$\mathcal{F}_2 = \{f : \mathbb{R} \to \mathbb{R} \,|\, \int \dot{f}(x)^2 dx < \infty\}$$

Hierarchical interaction space defined as

$$\mathcal{F} = (\mathcal{C}_1 \otimes \mathcal{C}_2) \oplus (\mathcal{F}_1 \otimes \mathcal{C}_2) \oplus (\mathcal{C}_1 \times \mathcal{F}_2) \oplus (\mathcal{F}_1 \otimes \mathcal{F}_2)$$

Hierarchical interaction model: $f \in \mathcal{F}$.

Theorem (kernel for hierarchical interaction space)

Suppose $\mathcal{F}_1$ and $\mathcal{F}_2$ are RKKSs with r.k.s h_1 and h_2, and $\mathcal{C}_1$ and $\mathcal{C}_2$ are RKHSs of constant functions with r.k.s equal to 1 everywhere. Then $\mathcal{F}$ is the RKKS with r.k. given by

$$h((x_1, x_2), (x_1', x_2')) = 1 + h_1(x_1, x_1') + h_2(x_2, x_2') + h_1(x_1, x_1')h_2(x_2, x_2')$$

Proposed estimation method

Model

$$y_i = f(x_i) + \varepsilon_i \quad f \in \mathcal{F}$$

where $\mathcal{F}$ is a hierarchical interaction space.

- f can be estimated in closed form by its posterior mean under the (Gaussian) I-prior.
- Any hyperparameters can be estimated by maximizing the marginal likelihood of f under the I-prior.

Parsimonious use of scale (hyper)parameters

Kernels can be multiplied by 'scale parameters' to potentially achieve (i) invariance to measurement units (e.g., fractional Brownian motion kernels), (ii) optimal convergence rates (van der Vaart and van Zanten, 2007, EJS).

Example (interaction kernel with scale parameters)

Suppose $\mathcal{F}_1$ and $\mathcal{F}_2$ are RKKSs with r.k. $\lambda_1 h_1$ and $\lambda_2 h_2$, $\lambda_k \in \mathbb{R}$ and h_1 and h_2 positive definite kernels. Then the hierarchical interaction space $\mathcal{F}$ is the RKKS with r.k. given by

$$h((x_1, x_2), (x_1', x_2')) = 1 + \lambda_1 h_1(x_1, x_1') + \lambda_2 h_2(x_2, x_2') + \lambda_1\lambda_2 h_1(x_1, x_1')h_2(x_2, x_2')$$

Parsimony achieved because only one scale parameter is needed per covariate. Usual approach in literature: separate scale parameter for each interaction term.

Parsimonious use of scale (hyper)parameters

$$h((x_1, x_2), (x_1', x_2')) = 1 + \lambda_1 h_1(x_1, x_1') + \lambda_2 h_2(x_2, x_2') + \lambda_1\lambda_2 h_1(x_1, x_1')h_2(x_2, x_2')$$

Advantage of parsimony: different interaction models have the same number of fixed hyperparameters making model comparison simple: choose model with highest maximized marginal likelihood.

NB: interaction spaces inherit properties of main effect spaces; encodes that size of interactions are related to sizes of main effects (e.g., Andrew Gelman blog).

Why RKKSs rather than RKHSs?

Interaction kernel with h_1, h_2 positive definite:

$$h((x_1, x_2), (x_1', x_2')) = \lambda_1 h_1(x_1, x_1') + \lambda_2 h_2(x_2, x_2') + \lambda_1\lambda_2 h_1(x_1, x_1')h_2(x_2, x_2')$$

If at least one $\lambda < 0$, h defines an RKKS.

The scale parameters (kernel weights λ_k) determine the curvature of the likelihood: no reason these should be restricted to be positive.

Abitrarily restricting scale parameters to be positive potentially leads to lower maximum of marginal likelihood, typically reducing prediction quality.

NB: RKKS needed if and only if there are at least two covariates.

(Note for Tikhonov regularization the RKHS suffices.)

Why not fully Bayes?

- Scale parameters tend to converge to zero or infinity.
- Hence no prior belief about them can be coherent.
- Frequentist estimation equally impossible (NB maximum likelihood *not* frequentist in this case!)

Outline

Simulation results: linear regressions

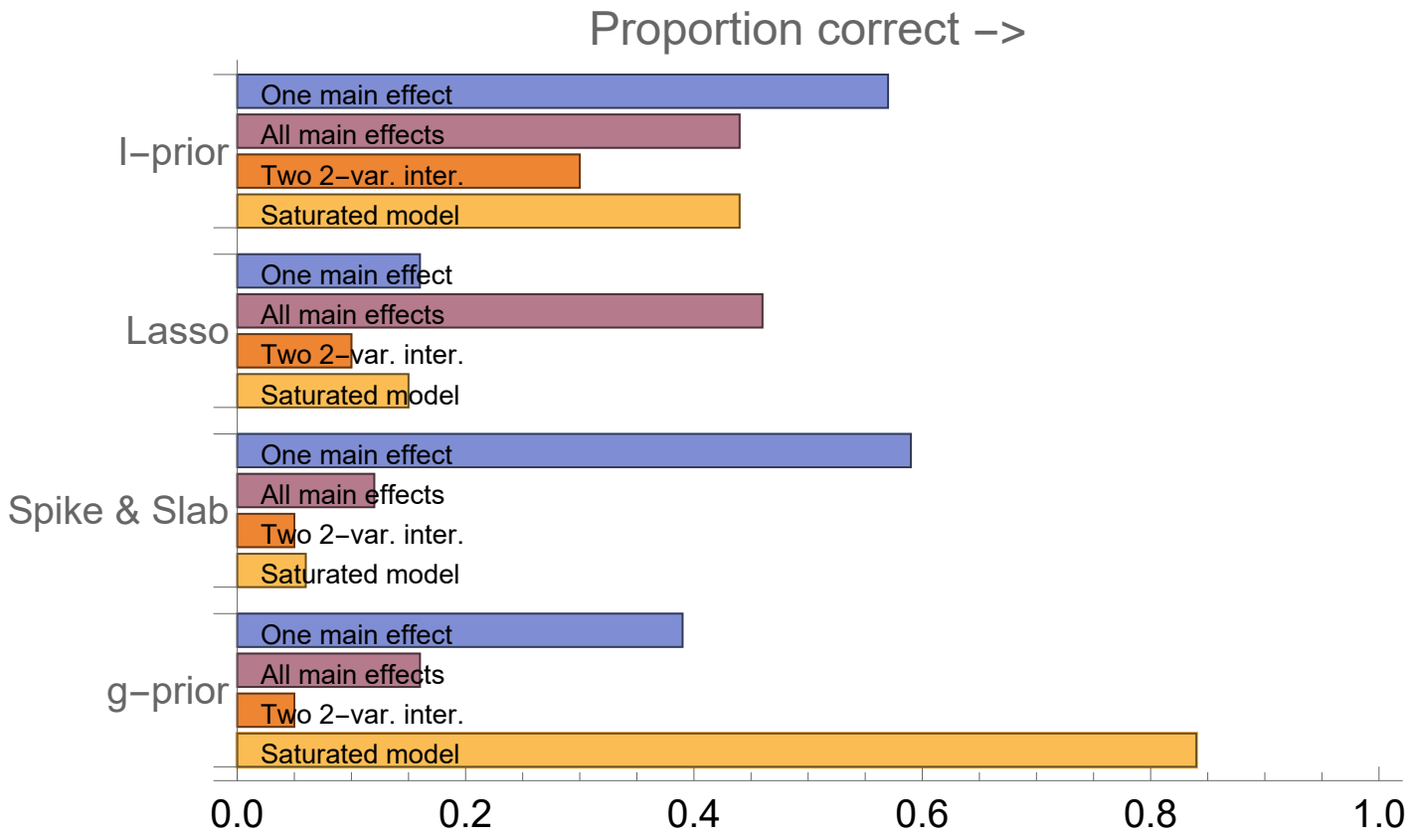
x_{1i}, x_{2i}, x_{3i} standard normal with correlation 0.5, $\varepsilon_i \sim_{\text{iid}} N(0, 3^2)$.

Saturated model: $y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{1i}x_{2i}\beta_{12} + \ldots + x_{1i}x_{2i}x_{3i}\beta_{123} + \varepsilon_i$

Two 2-way interactions: $y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{1i}x_{2i}\beta_{12} + x_{2i}x_{3i}\beta_{23} + \varepsilon_i$

All main effects: $y_i = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i$

One main effect: $y_i = \alpha + x_{1i}\beta_1 + \varepsilon_i$



Simulation results: smooth regressions

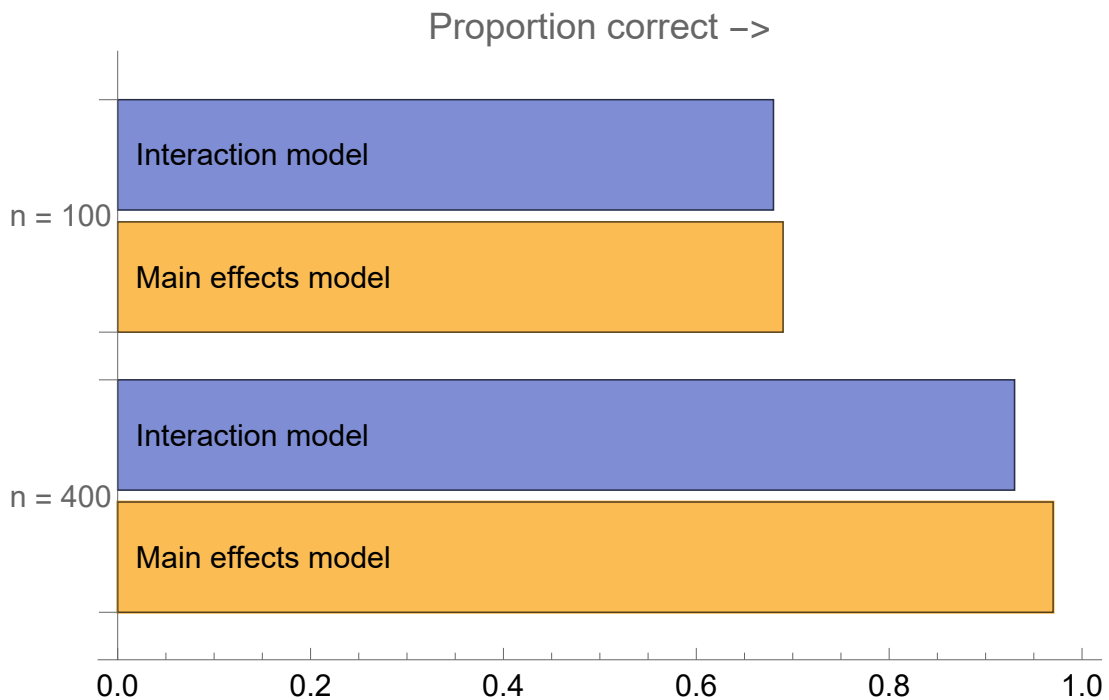
To illustrate: as $n \to \infty$, correct model is chosen with probability going to 1.

$\mathcal{F}_1 = \mathcal{F}_2 = \{f : \mathbb{R} \to \mathbb{R} | \int \ddot{f}(x)^2 dx < \infty\}$

f_1, f_2, f_{12} random functions in $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_1 \otimes \mathcal{F}_2$.

Interaction model: $y_i = f_1(x_{1i}) + f_2(x_{2i}) + f_{12}(x_{1i}, x_{2i}) + \varepsilon_i$

Main effects model: $y_i = f_1(x_{1i}) + f_2(x_{2i}) + \varepsilon_i$



Functional response model / longitudinal data

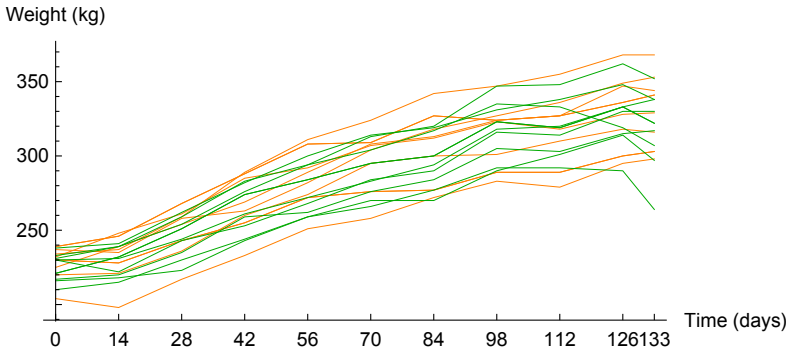


Figure: Cow growth under two treatments

Does treatment affect growth?

Permutation test for treatment effect based on Baringhaus-Franz 2-sample test statistic

$$T^2 = \sum\sum \left(2\|y_i - y_j\|_{\mathcal{F}} - \|y_i - y_j\|_{\mathcal{F}} - \|y_i - y_j\|_{\mathcal{F}} \right)$$

where y_i and y_j are growth curves for two treatments, and here $\|\cdot\|_{\mathcal{F}}$ is norm of fractional Brownian motion RKHS with (empirically determined) Hurst coefficient 0.3.

$p < .001$

Is there an interaction effect cow x treatment?

Response: growth curve

Covariates: C is cow index, X is treatment

RKHS for growth curve: fractional Brownian motion with Hurst coefficient 0.3.

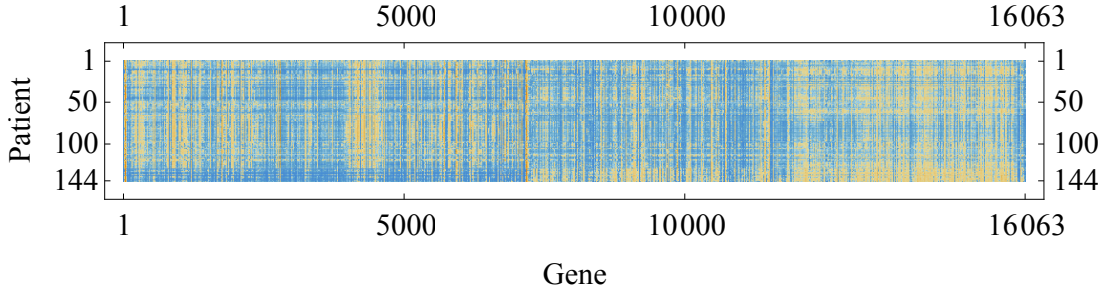
Interpretation	Model	Log-likelihood	Error standard deviation	Number of scale parameters
Identical growth all cows	$\{\}$	-2792.8	16.3	1
Treatment effect	$\{X\}$	-2792.7	16.3	2
Cow effect	$\{C\}$	-2266.4	2.7	2
Cow + treatment effects	$\{C, X\}$	-2242.3	2.5	3
Interaction effect	$\{CX\}$	-2251.3	3.3	3

No evidence treatment affects cows differentially! (Model $\{CX\}$.)

Conclusion

- I-prior methodology for estimating parametric and nonparametric interaction models
- Proper prior, hence admissible estimator
- Parsimonious use of scale parameters allows simple model comparison
- Simple EM algorithm
- Good comparative simulation performance
- Methodology generalizable to other estimation problems

Classifying cancers using gene expressions



$x_i \in \mathbb{R}^{16,063}$: vector of gene expressions for patient i

$$y_{ij} = \begin{cases} 1 & \text{patient } i \text{ has cancer } j \text{ (14 cancers)} \\ 0 & \text{patient } i \text{ does not have cancer } j \end{cases}$$

Linear model: $y_{ij} = \alpha_j + \sum_{k=1}^{16,063} x_{ik}\beta_{jk} + \varepsilon_{ij}$

I-prior for $y_{ij} = f(x_i, j) + \varepsilon_{ij}$ where

- x_i has linear or smooth effect
- j has nominal effect
- interaction effect present

Cancer classification: I-prior versus other methods

Method	Training errors Out of 144	Test errors Out of 54
Nearest neighbors	41	26
L^2-penalized discriminant analysis	25	12
Support vector classifier	26	14
Lasso	30.7	12.5
L^1 penalized multinomial	17	13
Elastic net penalized multinomial	22	11.8
SCRDA (Guo, Hastie, Tibsh 2007)	24	8
Scout (Witten, Tibshirani, 2011)	21	8

Cancer classification: I-prior versus other methods

Method	Training errors Out of 144	Test errors Out of 54
Nearest neighbors	41	26
L^2-penalized discriminant analysis	25	12
Support vector classifier	26	14
Lasso	30.7	12.5
L^1 penalized multinomial	17	13
Elastic net penalized multinomial	22	11.8
SCRDA (Guo, Hastie, Tibsh 2007)	24	8
Scout (Witten, Tibshirani, 2011)	21	8
I-prior (linear)	0	12

Cancer classification: I-prior versus other methods

Method	Training errors Out of 144	Test errors Out of 54
Nearest neighbors	41	26
L^2-penalized discriminant analysis	25	12
Support vector classifier	26	14
Lasso	30.7	12.5
L^1 penalized multinomial	17	13
Elastic net penalized multinomial	22	11.8
SCRDA (Guo, Hastie, Tibsh 2007)	24	8
Scout (Witten, Tibshirani, 2011)	21	8
I-prior (linear)	0	12
I-prior (smooth, $\hat{\gamma} = 0.8$)	0	10

I-priors for density estimation

Model $p(x) \propto e^{f(x)}$, f has L_2 derivative.