Finding important variables and interactions in black boxes

Art B. Owen

Stanford University

owen@stat.stanford.edu

Tao Jiang

Stanford University

jiang@stat.stanford.edu

The borehole function

Morris, Mitchell, Ylvisaker

Flow from upper to lower aquifer:

$$\frac{2\pi T_u \left[H_u - H_l\right]}{\log\left(\frac{r}{r_w}\right)\left[1 + \frac{2LT_u}{\log\left(\frac{r}{r_w}\right) r_w^2 K_w} + \frac{T_u}{T_l}\right]}$$

r, r_w	Radii	borehole, basin
T_l, T_u	Transmissivities	upper and lower
H_l, H_u	Potentiometric heads	upper and lower
L, K_w	Length	and conductivity

Diaconis: closed form $\neq$ understanding

Which variables are important?

Which interact?

Black box functions

$$Y = f(X) \quad \text{Without} \quad \text{“}+ e\text{”}$$

Examples	X	Y
Semiconductors	Device design	Speed, heat
Aerospace	Wing shape	Lift, drag
Automotive	Auto Frame	Strength, weight
Statistics	Predictors	Responses

Used to design products. Cheaper than physical experiments. Costs from milliseconds to hours. Dimension from 3 to 300. Accuracy varies too.

Kriging widely used Journel, Huijbreghts, Sacks, Ylvisaker, Welch, Wynn, Mitchell

A small neural net

Venables, Ripley

Predict $\log_{10}(perf)$ from the others

perf	published performance of computer
syct	cycle time in nanoseconds
mmin	minimum main memory in kilobytes
mmax	maximum main memory in kilobytes
cach	cache size in kilobytes
chmin	minimum number of channels
chmax	maximum number of channels

Function found by training on 209 examples.

The n-net function

$$
\begin{aligned}
&0.46 - 1.21x_1 + 1.36x_2 + 1.42x_3 \\
&\quad - 1.01x_4 - 0.33x_5 + 0.30x_6 \\
&-2.82\varphi\Big(-1.12 + 0.45x_1 + 2.24x_2 + 2.51x_3 \\
&\quad - 1.63x_4 - 0.56x_5 + 0.43x_6\Big) \\
&+3.17\varphi\Big(-1.09 + 2.28x_1 - 0.10x_2 + 1.44x_3 \\
&\quad + 2.70x_4 + 1.24x_5 + 0.25x_6\Big) \\
&+0.39\varphi\Big(0.04 - 0.11x_1 + 0.11x_2 + 0.12x_3 \\
&\quad - 0.10x_4 - 0.04x_5 + 0.02x_6\Big)
\end{aligned}
$$

where $\varphi(z) = [1 + \exp(-z)]^{-1}$ is a sigmoidal function

ANOVA of $L^2[0,1]^d$

Hoeffding, Efron & Stein, Sobol'

Main effects and k–factor interactions generalizing familiar discrete ANOVA

$$f(x) = \sum_{u \subseteq \{1,2,\ldots,d\}} f_u(x)$$

- f_u depends only on x-components in set u
- $f_\emptyset = \int f(x)dx$ "grand mean"
- $\sigma^2(f) = \sum_{u \neq \emptyset} \int f_u(x)^2 dx$
- $\int f_u(x) f_v(x) dx = 0, u \neq v$

Given $f(x)$ on $[0,1]^d$

How can we tell if f is:

1. Nearly linear?
2. Nearly additive?
3. Nearly quadratic?
4. Has mostly 3 factor interactions or less?
5. Which variables matter most?
6. Which interactions matter most?

We would like:

1. a systematic approach
2. that predicts f
3. and predicts ANOVA components f_u

MC approximation

Let: $$y = f(x) = \sum_{r \in \mathcal{U}} \beta_r \psi_r(x)$$
$$= \sum_{r \in \mathcal{R}} \beta_r \psi_r(x) + \eta(x)$$

Orthonormal basis ψ_r

$\eta(x)$ a deterministic truncation error

Estimate β_r from $f(x_i)$ values, where $x_i \sim U(0,1)^d$ getting

$$\widetilde{f}(x) = \sum_{r \in \mathcal{R}} \widetilde{\beta}_r \psi_r(x)$$

Apply graphical and numerical interpretation to $\widetilde{f}$

Start with univariate basis functions

$$\phi_0,\ \phi_1,\ \phi_2,\ \ldots$$

First is constant, all are orthonormal

$$\begin{aligned} \phi_0(x) &= 1, \quad 0 \le x \le 1 \\ \int_0^1 \phi_j(x)dx &= 0, \quad j \ge 1 \\ \int_0^1 \phi_j(x)\phi_k(x)dx &= 1_{j=k} \end{aligned}$$

EG: orthogonal polynomials, sinusoids, wavelets,

Hermite($\Phi^{-1}(\cdot)$), Chebychev(qbeta$(\cdot)$)

Tensor product basis

$$
\begin{aligned}
x &= (x_1, x_2, \ldots, x_d) \in [0,1]^d \\
r &= (r(1), r(2), \ldots, r(d)) \in \{0, 1, 2, \ldots\}^d \\
\psi_r(x) &= \prod_{j=1}^{d} \phi_{r(j)}(x_j)
\end{aligned}
$$

Finite subset of basis:

$$
\begin{array}{lcllcl}
\text{Rank}(r) & \equiv & \|r\|_0 & = \sum_{j=1}^{d} 1_{r(j)>0} & \leq & B_0 \\
\text{Degree}(r) & \equiv & \|r\|_1 & = \sum_{j=1}^{d} r(j) & \leq & B_1 \\
\text{Order}(r) & \equiv & \|r\|_\infty & = \max_{1 \leq j \leq d} r(j) & \leq & B_\infty
\end{array}
$$

Polynomials ψ_r, with . . .

Rank $\|r\|_0 \leq 3$, Degree $\|r\|_1 \leq 4$, Order $\|r\|_\infty \leq 3$.

Rank	Deg	Order	$\psi_r(x)$	#
0	0	0	Const	1
1	1	1	Linear	d
	2	2	Quad	d
	3	3	Cubic	d
2	2	1	Lin $\times$ Lin	$\binom{d}{2}$
	3	2	Lin $\times$ Quad	$d(d-1)$
	4	3	Lin $\times$ Cubic	$d(d-1)$
	4	2	Quad $\times$ Quad	$\binom{d}{2}$
3	3	1	Lin $\times$ Lin $\times$ Lin	$\binom{d}{3}$
	4	2	Lin $\times$ Lin $\times$ Quad	$3\binom{d}{3}$

$$p = 1 + 3d + 3d(d-1) + (2/3)d(d-1)(d-2)$$

Interpretation

Variance of f is $\sum_{r \neq 0} \beta_r^2 + \int \eta(x)^2 dx$

Importance of $\mathcal{S}$ is $\sum_{r \in \mathcal{S}} \beta_r^2$

Estimate by $\sum_{r \in \mathcal{S}} \widetilde{\beta}_r^2 - \widehat{\mathrm{Var}}(\widetilde{\beta}_r)$

Subsets of interest include:

$\{r \mid r(1) > 0\}$	involves x_1
$\{r \mid r(1) = 0\}$	does not involve x_1
$\{r \mid \|r\|_0 = 1\}$	additive part
$\{r \mid 0 < \|r\|_0 \leq k\}$	interactions up to order k
$\{r \mid 0 < \|r\|_1 \leq k\}$	of degree at most k
$\{r \mid r(j) = 0, j > 3\}$	uses only first 3 inputs,

Approximation through integration

Define: $Z(x) = (\psi_0(x), \ldots, \psi_{p-1}(x))^T$

Optimal β is

$$\beta^* = \arg\min_{\beta} \int \left(f(x) - Z(x)^T\beta\right)^2 dx$$

$$= \left(\int Z(x)Z(x)^T dx\right)^{-1} \int Z(x)f(x)dx$$

also,

$$ISE = \int (f(x) - Z(x)^T\beta)^2 dx$$

Regression and quasi-regression

$$\beta^* = \left(\int Z(x)Z(x)^T dx\right)^{-1} \int Z(x)f(x)dx$$

$$= \int Z(x)f(x)dx \quad \text{by orthogonality}$$

Observations

$$x_i \sim U[0,1]^d, 1 \le i \le n, \quad \text{IID}$$

Regression

$$\hat{\beta} = \left(\mathcal{Z}^T\mathcal{Z}\right)^{-1}\mathcal{Z}^T\mathcal{Y} \qquad \mathcal{Z}_{n\times p} \quad \mathcal{Y}_{n\times 1}$$

Quasi-Regression

$$\widetilde{\beta} = \frac{1}{n}\mathcal{Z}^T\mathcal{Y}$$

Precursors of quasi-regression

Quasi-interpolation

Chui & Diamond, Wang

"Ignore the denominator" ($\mathcal{Z}^T\mathcal{Z}$) to get fast approximate interpolation.

Computer experiments

Koehler and Owen 1996 advocate quasi-regression for computer experiments

Efromovich 1992 applies qr to sinusoids on $[0,1]$.

Owen 1992 describes quasi-regression for Latin hypercube sampling

Fast stable updates

Define:

$$\widetilde{\beta}_r^{(n)} \equiv \frac{1}{n}\sum_{i=1}^{n}\psi_r(x_i)f(x_i)$$

$$S_r^{(n)} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\psi_r(x_i)f(x_i) - \widetilde{\beta}_r^{(n)}\right)^2$$

Then:

$$\widetilde{\beta}_r^{(n)} = \widetilde{\beta}_r^{(n-1)} + \frac{1}{n}\left[\psi_r(x_i)f(x_i) - \widetilde{\beta}_r^{(n-1)}\right]$$

$$S_r^{(n)} = \frac{n-1}{n}S_r^{(n-1)} + \frac{n-1}{n^2}\left[\psi_r(x_i)f(x_i) - \widetilde{\beta}_r^{(n-1)}\right]^2$$

Chan, Golub, Leveque who use $nS_r^{(n)}$

$$E\left(\tfrac{n}{n-1}S_r^{(n)}\right) = \mathrm{Var}(\widetilde{\beta}_r^{(n)})$$

Updatable accuracy estimates

Predict $f(x_n)$ by $\widetilde{f}_{n-1}(x_n)$ $\quad x_n$ indep of $\widetilde{f}_{n-1}$

Average recent squared errors

$$\widehat{ISE}(n_m) = \frac{1}{n_m - n_{m-1}} \sum_{i=n_{m-1}}^{n_m} \left(f(x_i) - \widetilde{f}_{i-1}(x_i) \right)^2$$

on subsequence $n_m = m(m+1)/2$

estimates avg ISE over recent $\approx \sqrt{2n}$ values

Diagnostic:

Large LOF and small $\sum_r \widehat{\mathrm{Var}}(\widetilde{\beta}_r) \implies$ need bigger basis

Presented as lack-of-fit:

$$1 - R^2$$

$$\text{LOF} = \frac{ISE}{Var} \qquad \widehat{\text{LOF}} = \frac{\text{AVG}(f - \widetilde{f})^2}{\text{AVG}(f - \widetilde{\beta}_0)^2}$$

$\log_{10}(LOF)$	R^2
-4	99.99%
-3	99.9%
-2	99%
-1	90%
0	0%
1	-900%

Costs of algebra

	Time	Space	Footprint
Kriging	$O(n^3 + p^3)$	$O(n^2 + p^2)$	$O(n^5 + p^5)$
Reg.	$O(np^2)$	$O(p^2)$	$O(np^4)$
Q.–Reg.	$O(np)$	$O(p)$	$O(np^2)$

Quasi-reg allows larger n or much larger p

$p = 1,000,000$ doable by quasi-reg., not by reg.

Owen, Ann Stat 2000

Cost of f	Dimension	
Low	Low	Easy
High	Low	Kriging
Low	High	(Quasi-)regression
High	High	[good luck]

Incorporating shrinkage

Hoerl, Kennard, Efromovich, Donoho, Johnstone, Beran . . .

$$\widetilde{f}_{\gamma,n}(x) = \sum_r \gamma_{r,n}\widetilde{\beta}_{r,n}\psi_r(x), \quad \gamma_{r,n} \in [0,1]$$

Optimally

$$\gamma_{r,n} = \frac{\beta_r^2}{\beta_r^2 + \mathsf{Var}(\widetilde{\beta}_{r,n})}$$

Shrinkage can reduce prediction variance.

We use data to estimate $\gamma_{r,n}$

$$\text{e.g.} \quad \hat{\gamma}_{r,n} = \frac{\widetilde{\beta}_r^{(n-1)^2}}{\widetilde{\beta}_r^{(n-1)^2} + S_r^{(n-1)}}$$

Exploiting residuals

For $r \neq 0$: $\beta_r(f) = \beta_r(f - c)$, for $c \in \mathbb{R}$

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_r(x_i)(f(x_i) - c)\right)$$ depends on c

Try $c \approx \beta_0$

More generally

$$\widetilde{\beta}_r^{(n)} \equiv \frac{1}{n}\sum_{i=1}^{n} \psi_r(x_i)\left(f(x_i) - \sum_{s \neq r} \lambda_{s,i-1}\widetilde{\beta}_s^{(i-1)}\psi_s(x_i)\right)$$

Original quasi-reg: $\lambda_{r,i} = 0$ or $1_{r==0}$ $\gamma_{r,i} = 1_{r \in \mathcal{R}}$

Self-consistent quasi-reg: $\lambda_{r,i} = \gamma_{r,i} \in [0, 1]$

Bounding $\int \widetilde{f}^2$ by sample variance . . . eliminates explosive feedback

Still updatable $\widetilde{\beta}_r$ and S_r

NB: $n(\widetilde{\beta}_r^{(n)} - \beta_r)$ is a martingale in n

N-net example

$f(x)$ is prediction of $\log_{10}(\textit{perf})$

$d = 6$ ϕ_r are Legendre polynomials

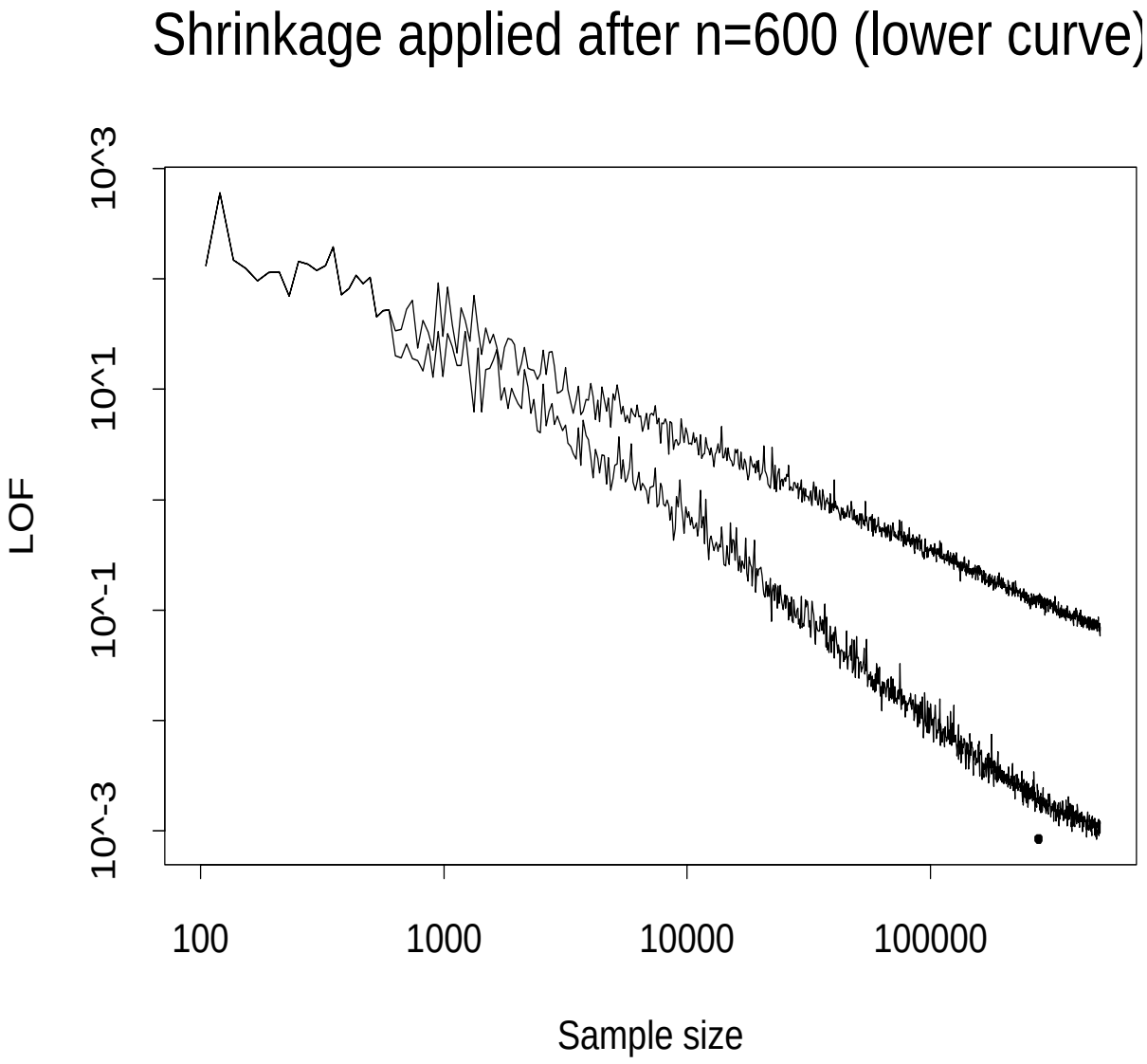
ψ_r are tensor products

$$\|r\|_0 \leq 3 \quad \|r\|_1 \leq 8 \quad \|r\|_\infty \leq 4 \implies p = 1145$$

Net is fast, so $n = 500,000$

(about 3min on 800Mhz PC in java)

Neural net accuracy



Additive component of $\widetilde{f}$

Input	Main effect
syct	0.520
mmin	0.011
mmax	0.088
cach	0.131
chmin	0.037
delch	0.009
Total	0.797

Function is 79.7% additive, mostly from syct

Neural net results

```
Number of bases is 1145

######  Anova at Iteration 500000  ######

1-RSquare (LOF) is 0.0011707 at iteration 499500

Beta[0] (constant factor) is 2.0717
Sample mean is 2.0719,  sample variance is 0.1435

Unbiased estimates of dimension variances
0.11441    0.026592   0.0027723  0.0        0.0

Dimension Probabilities
(Ratios of dimension variances to sample variance

0.79676    0.18518    0.019307   0.0        0.0
```

Neural net results, ctd

Normalized Variances:

main effects on diagonal

two factor interactions below

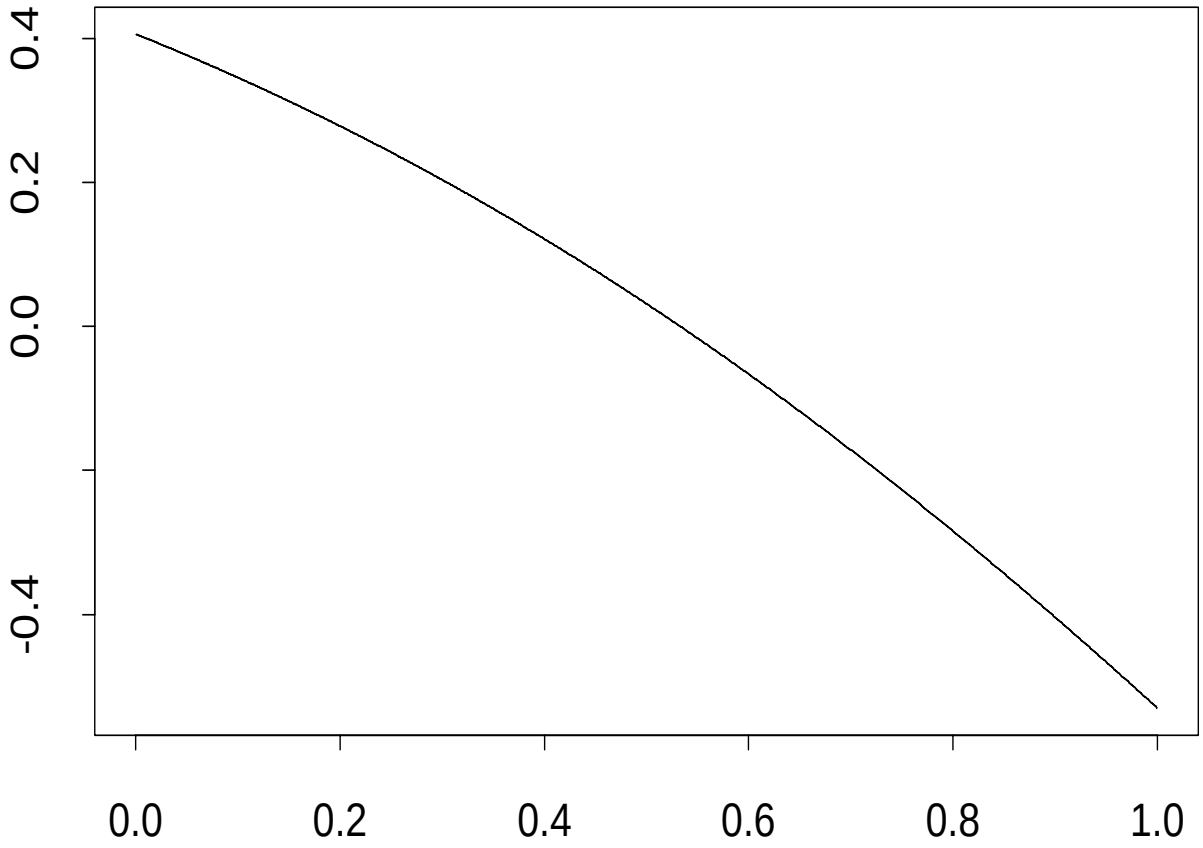
```
syct   mmin   mmax   cach chmin deltach

0.520
0.001 0.011
0.009 0.026 0.088
0.055 0.006 0.055 0.131
0.011 0.001 0.009 0.010 0.037
0.000 0.001 0.000 0.001 0.000 0.009
```

Biggest main effect: *syct* is 52%

Biggest interaction *syct*$\times$*cach* is 5.53%

Next interaction *syct*$\times$*mmax* is 5.48%

Effect of *cycle time*



Degree	1	2	3	4
Coef	-0.272	-0.030	0.00242	.0000777
% of $\widetilde{f}$	51.38	0.630	0.00041	0.000004

Caveats

- Important variables in $E(Y \mid X = x)$ are not necessarily causal
- Same for $f(x) = \hat{E}(Y \mid X = x)$ and $\widetilde{f}$
- Training x not from a product measure (nor are test x)

Non-product measure issues

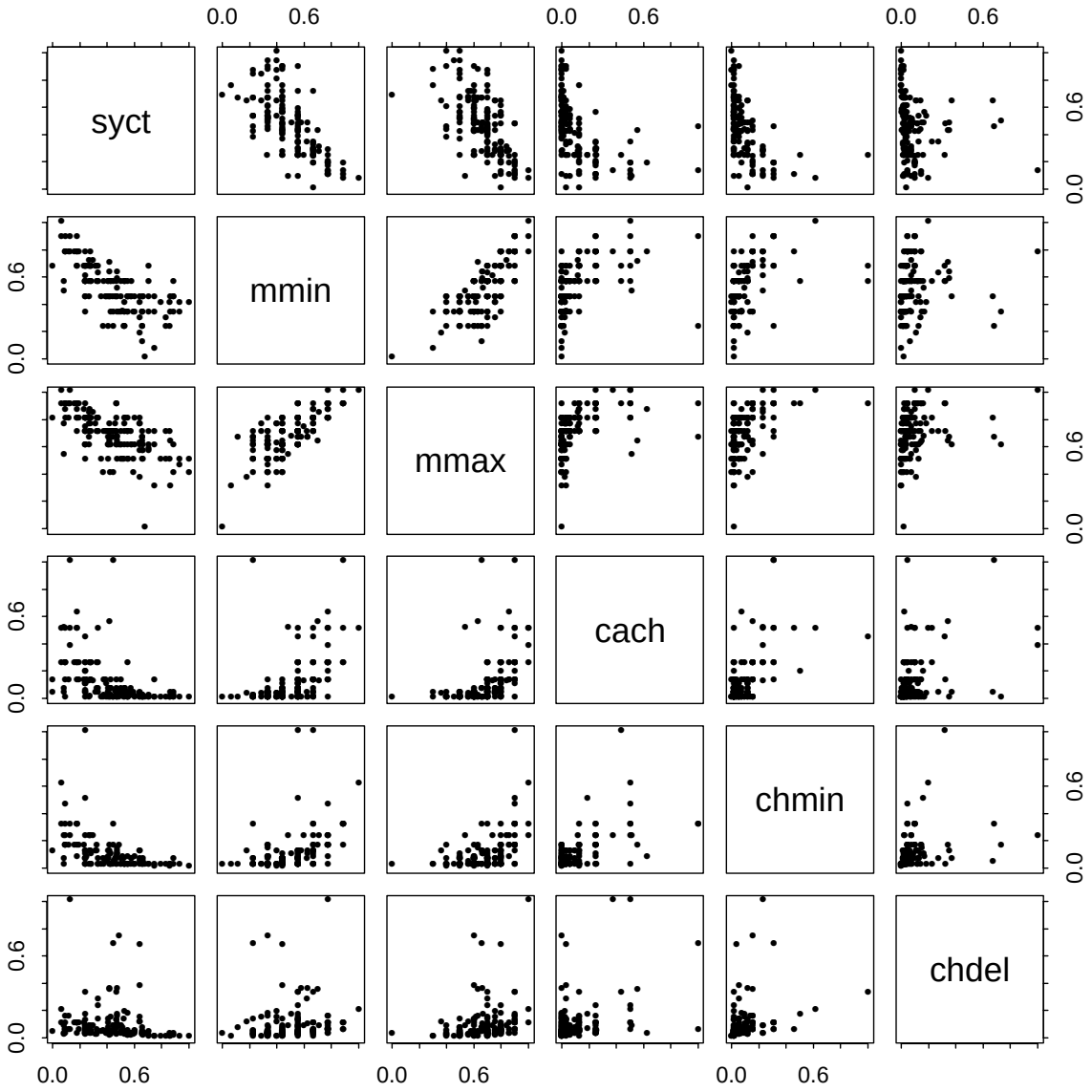
False positives: $f, \widetilde{f}$ might have large structure in region with no data

False negatives: $\int (\widetilde{f} - f)^2 dx$ might be dominated by x away from data. Small error and simple model might mask poor fit in training region. (Easy to compare f and $\widetilde{f}$ on training data.)

Functions ψ_r and estimated anova components correlated on empirical distribution

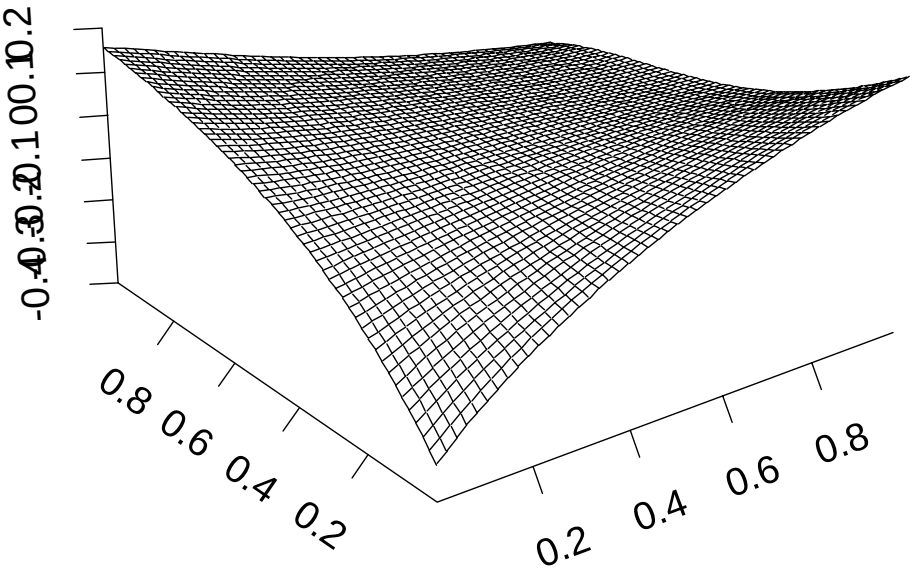
Using product of empirical margins mitigates problem (only slightly)

CPU inputs

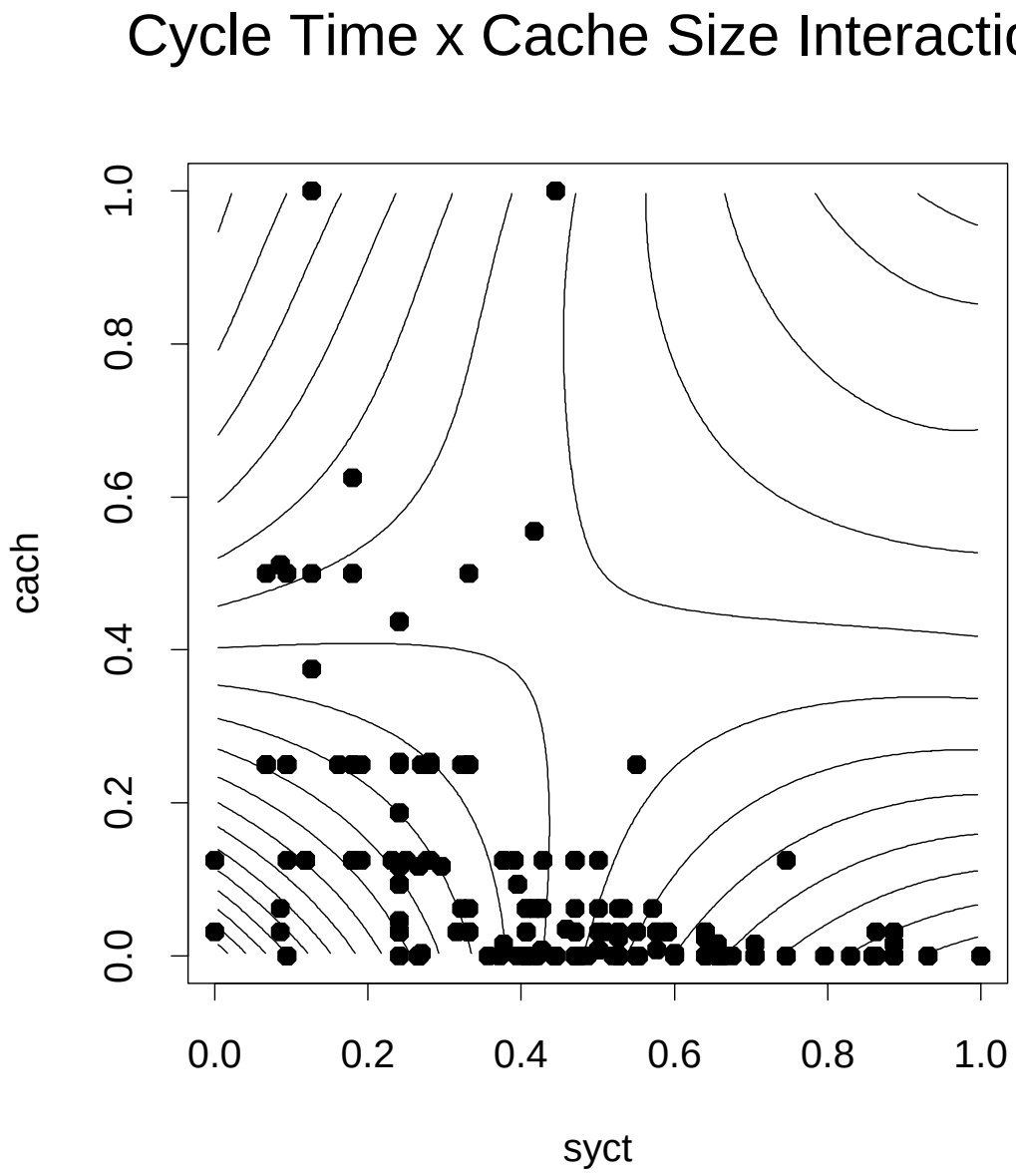


Biggest interaction

Cycle time $\times$ *Cache Size* $\approx$ 5.5% of $\widetilde{f}$

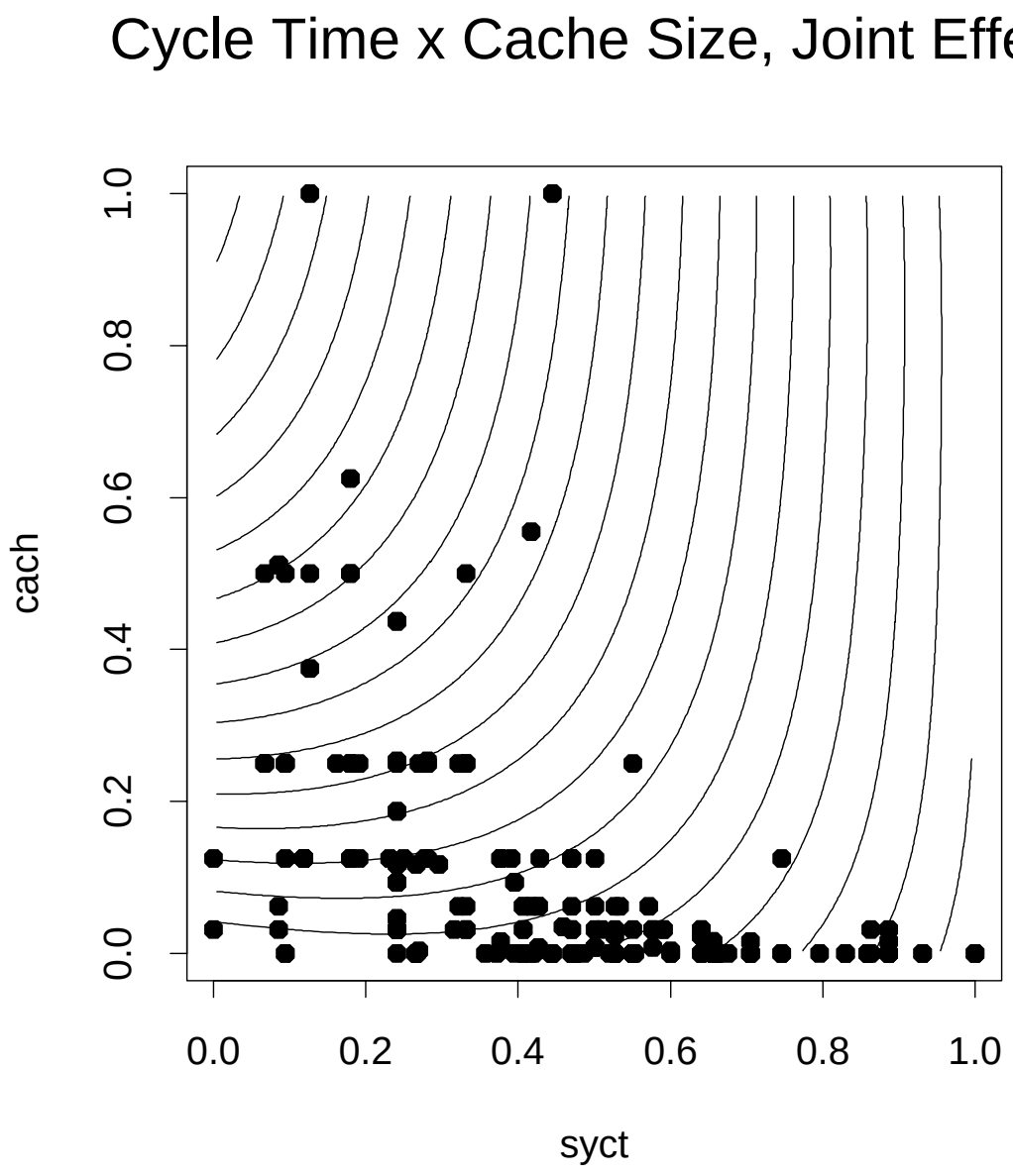


Biggest interaction



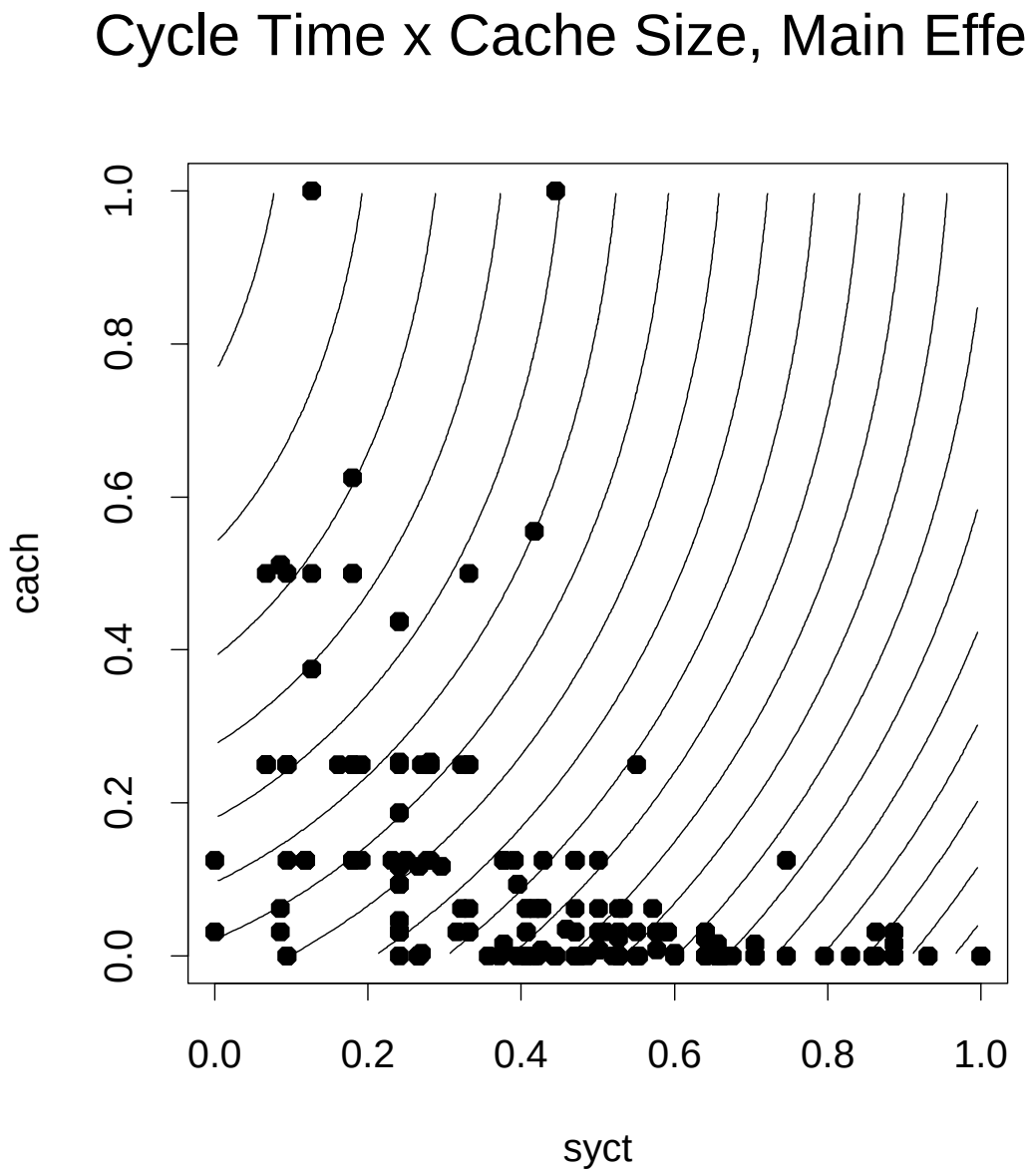
$$\|r\|_0 \leq 3 \quad \|r\|_1 \leq 8 \quad \|r\|_\infty \leq 4$$

Joint effect



$$\|r\|_0 \leq 3 \quad \|r\|_1 \leq 8 \quad \|r\|_\infty \leq 4$$

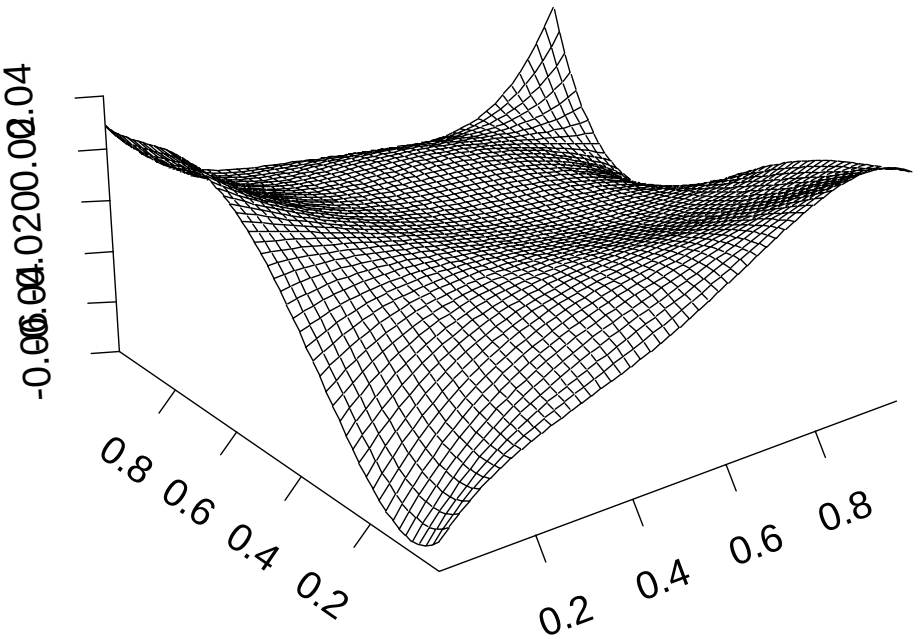
Two main effects



$$\|r\|_0 \leq 3 \quad \|r\|_1 \leq 8 \quad \|r\|_\infty \leq 4$$

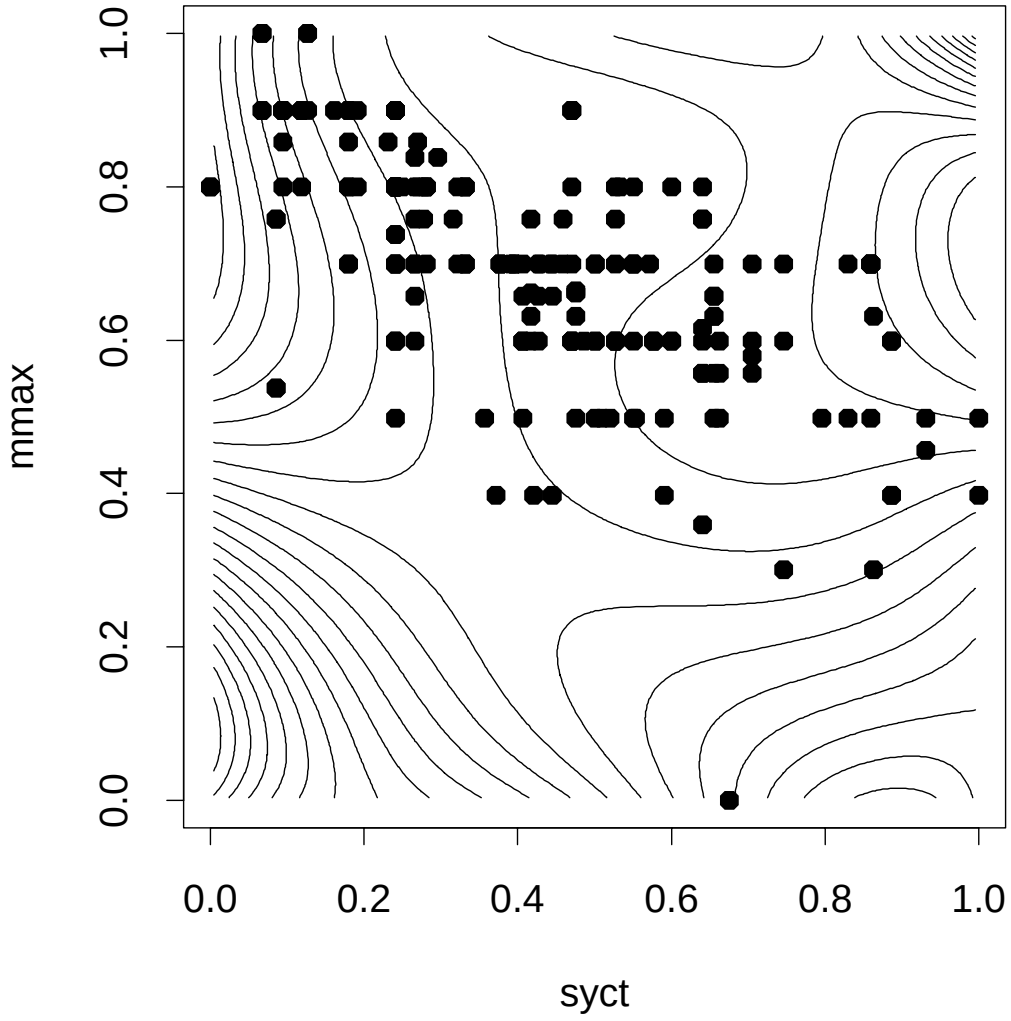
2nd biggest interaction

Cycle time$\times$*Main Memory Max* $\approx$ 5.4% of $\widetilde{f}$



2nd biggest interaction

Cycle Time x Max Main Memory Interaction



$$\|r\|_0 \leq 3 \quad \|r\|_1 \leq 8 \quad \|r\|_\infty \leq 4$$

N-net conclusions

1. $\widetilde{f}$ a fairly simple function wrt $U[0,1]^6$
2. x_1 most important, and nearly linear
3. One interaction not supported by data
4. But it counters the main effects there

Jiang's thesis

1. More basis functions
2. Effects of shrinkage rules
3. Mars-like dynamic choice of basis
4. Comparisons of f and $\widetilde{f}$ on training data

For more

Software and papers at:

http://www-stat.stanford.edu/$\sim$jiang/qra/

Or Google with:

quasi regression analzyer jiang