Variable importance measures in high dimensional data sets

Art B. Owen

Stanford University

Based on joint work with:

Naofumi Hama, Hitachi, Ltd.

Masayoshi Mase, Hitachi, Ltd.

Benjamin Seiler, Stanford Statistics

Christopher Hoyt, Stanford ICME

Opinions are my own, and not those of Stanford, the NSF, or Hitachi, Ltd.

These are the slides I presented at ICIAM 2023 in Tokyo. The goal was to present variable importance ideas to an audience of experts in QMC and related areas because tools from those areas are useful in quantifying variable importance.

I wrote on the board about the important example of the COMPAS data like:

Race	Gender	Age	Felony	Priors	Predicted to reoffend
B	M	25	1	3	1
⋮	⋮	⋮	⋮	⋮	⋮
W	F	19	0	2	0

and a question about what variables were important to the prediction and how that connects to fairness. Later slides have a toy example, but our papers have a deep dive into this real example. Here it is critical to not mix and match arbitrary variable combinations, e.g., greatest number of prior convictions with least age. The black box models will not have been trained on data like that.

Connection to QMC

Methods from QMC

ANOVA decomposition, anchored decomposition

can measure variable importance in statistics and machine learning models

Invitation

Apply QMC / cubature / approximation / complexity

to problems in machine learning

explanation, fairness, privacy

Overview

- Black box models can be the most accurate
- How do humans interpret them?

Methods from (R)QMC are useful

- ANOVA: underlies Sobol' indices
- Anchored decomposition: underlies Shapley value

Variable importance

A step towards explanation

Predict Y by $f(\boldsymbol{x})$

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$$

How important is x_j?

Causal: x_j affects Y (in real life)

Mechanical: x_j affects $f(\boldsymbol{x})$ (in silico)

Local vs global

Local: x_{tj} affects $f(\boldsymbol{x}_t)$ (subject t)

Global: variable j needed for model accuracy

Some notation

$1{:}d = \{1, 2, \ldots, d\}$

$u \subseteq 1{:}d$ has

cardinality $|u|$

complement $-u = 1{:}d \setminus u$

$\boldsymbol{x}_u = (x_j)_{j \in u}$

Hybrid points

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}_u{:}\boldsymbol{x}'_{-u} \quad \Longrightarrow \quad \tilde{x}_j = \begin{cases} x_j, & j \in u \\ x'_j, & j \notin u \end{cases}$$

Abbreviation

x_j is $x_{\{j\}}$

x_{-j} is $x_{-\{j\}}$

ANOVA

Independent x_j and $0 < \mathbb{E}(f(\boldsymbol{x})^2) < \infty$

$$f(\boldsymbol{x}) = \sum_{u \subseteq 1:d} f_u(\boldsymbol{x})$$

Effect $f_u(\boldsymbol{x})$ depends **only** on $\boldsymbol{x}_u$

$$f_\varnothing(\boldsymbol{x}) = \mathbb{E}(f(\boldsymbol{x})) \equiv \mu \qquad \text{constant}$$

Variance components

$$\sigma^2 = \mathrm{Var}(f(\boldsymbol{x})) = \sum_{u \subseteq 1:d} \sigma_u^2$$

$$\sigma_u^2 = \mathrm{Var}(f_u(\boldsymbol{x}))$$

Fisher & MacKenzie (1923), Hoeffding (1948), Sobol' (1969), Efron & Stein (1981)

Mean dimension

$$\nu(f) \equiv \frac{1}{\sigma^2} \sum_{u \subseteq 1:d} |u| \sigma_u^2$$

Identities

$$\nu(f) = \frac{1}{\sigma^2} \sum_{j=1}^{d} \bar{\tau}_j^2 \qquad \text{Liu \& O (2006)}$$

$$\tau_j^2 = \frac{1}{2} \mathbb{E}\big((f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}}_j{:}\boldsymbol{x}_{-j}))^2 \big) \qquad \text{Jansen (1999)}$$

$$\tilde{\boldsymbol{x}} \overset{\mathrm{d}}{=} \boldsymbol{x} \quad \text{indep}$$

τ_j^2 is a Sobol' index

Survey: Razavi et al. (2021)

MNIST classifier

Hoyt & O (2021) SIAM/ASA JUQ

Neural network to classify handwritten digits

$28 \times 28 = 784$ gray level pixels

Pre-softmax function

$f : [0,1]^{28\times 28} \to \mathbb{R}^{10}$

has mean dimensions 1.5 to 2.

Open issue

Low mean dimension under 13 different independent distributions

But . . . data near low dim manifold

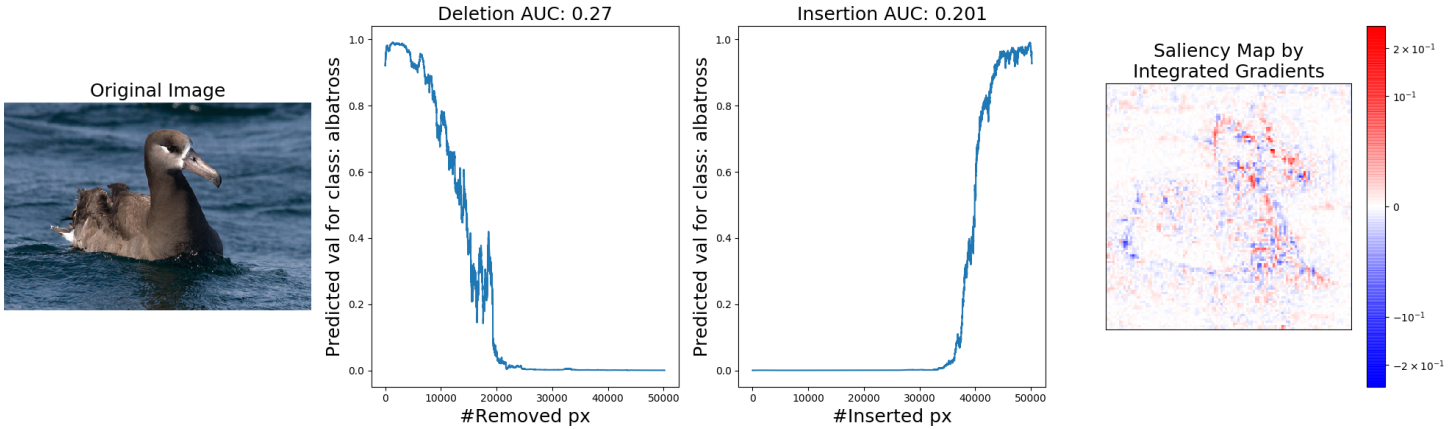
So $\cdots$ how to quantify mean dimension under dependence?

Deletion measures

Petsiuk, Das, Saenko (2018)

Rank variables from most to least important.

‘Destroy’ them in that order. Watch how fast classification deteriorates.



Hama, Mase, O JMLR (tentatively accepted)

Insertion

Include variables in importance order

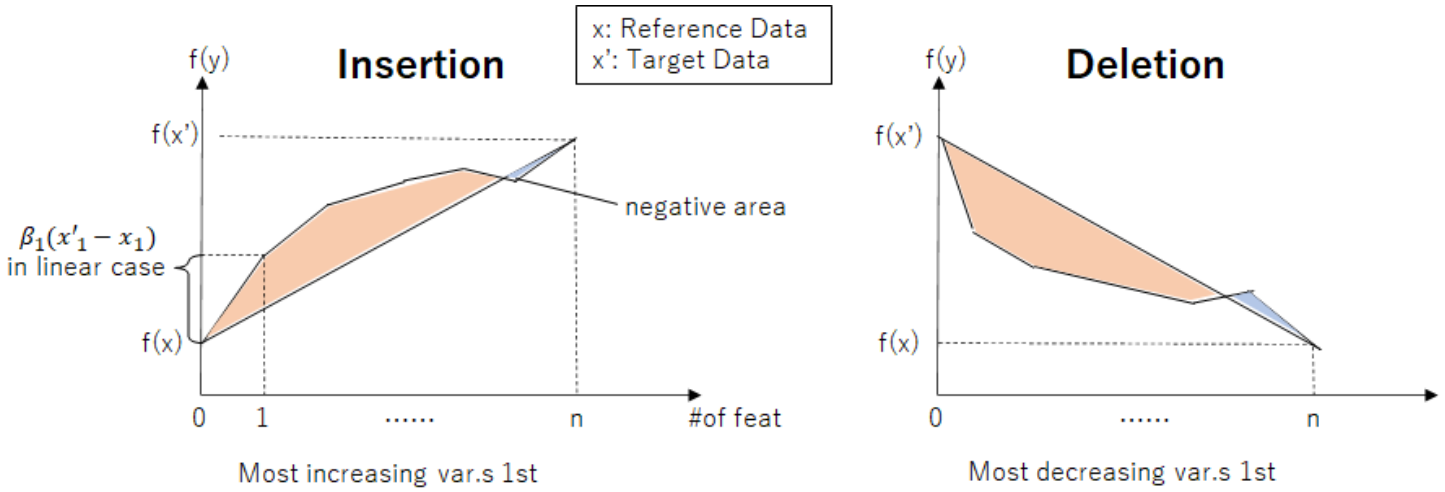
Keeping score

Area under curve measures quality

Area between the curves

Change x_j to x'_j one at a time

From 'most increasing' to 'least increasing'



Large area $\Longrightarrow$ good variable ranking

$f(x_t)$ could be higher or lower than $f(x_b)$

Anchored decomposition

Kuo, Sloan, Wasilkowski, Woźniakowski (2010),

Alis and Rabitz (2001), Sobol' (2003)

x_j not independent $\cdots$ not even random

Anchor at $\mathbf{0}$

$$
\begin{aligned}
g(\boldsymbol{x}) &= \sum_{u \subseteq 1:d} g_u(\boldsymbol{x}) \\
g_\varnothing(\boldsymbol{x}) &= g(\mathbf{0}) \\
g_j(\boldsymbol{x}) &= g(\boldsymbol{x}_j{:}\mathbf{0}_{-j}) - g(\mathbf{0}) \\
g_{j,k}(\boldsymbol{x}) &= g(\boldsymbol{x}_{j,k}{:}\mathbf{0}_{-\{j,k\}}) - g_j(\boldsymbol{x}) - g_k(\boldsymbol{x}) - g(\mathbf{0})
\end{aligned}
$$

Generally

$$
\begin{aligned}
g_u(\boldsymbol{x}) &= g(\boldsymbol{x}_u{:}\mathbf{0}_{-u}) - \sum_{v \subsetneq u} g_v(\boldsymbol{x}) \\
&= \sum_{v \subseteq u} (-1)^{|u-v|} g_v(\boldsymbol{x}_v{:}\mathbf{0}_{-v})
\end{aligned}
$$

Target vs baseline

Variable j: Target $\boldsymbol{x}_j$ OR Baseline $\boldsymbol{x}'_j$

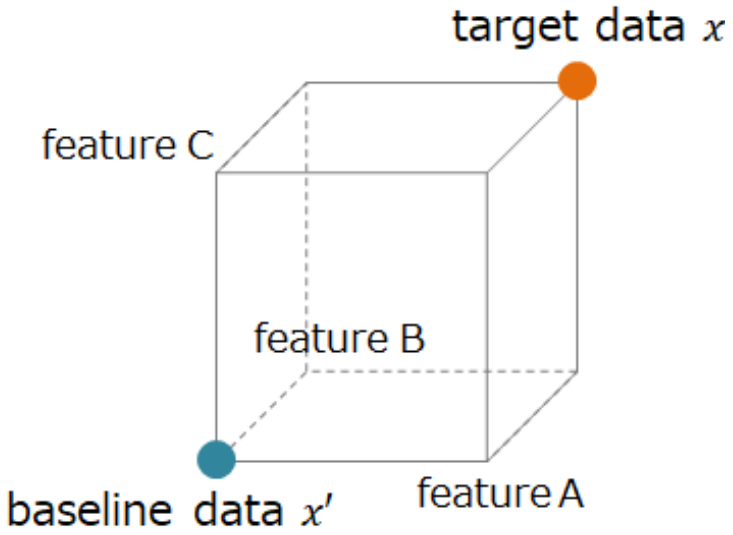
2^d 'corners' $\boldsymbol{z} \in \{0,1\}^d$

$$g(\boldsymbol{z}) = f(\boldsymbol{z}\boldsymbol{x} + (\mathbf{1} - \boldsymbol{z})\boldsymbol{x}')$$ 'corner' function

$z_j = 1 \implies$ use target $\boldsymbol{x}_j$

$z_j = 0 \implies$ use baseline $\boldsymbol{x}'_j$

From Hama, Mase, O (2022)



Area under insertion curve

Define

$$\Delta_u = \sum_{v \subseteq u} (-1)^{|u-v|} f(\boldsymbol{x}'_v{:}\boldsymbol{x}_{-v})$$

via anchored decomp. of $g(\boldsymbol{z}) = f(\boldsymbol{z}\boldsymbol{x} + (\boldsymbol{1} - \boldsymbol{z})\boldsymbol{x}')$

We find

$$\mathrm{AUC} = \sum_{u \subseteq 1:d} \big(n - \lceil u \rceil + 1\big)\Delta_u$$

$$\lceil u \rceil = \max\{1 \leqslant j \leqslant d \mid j \in u\}$$

Count Δ_u when its **last** variable has changed

Area under deletion curve

$$\mathrm{AUC}' = \sum_{u \subseteq 1:d} \lfloor u \rfloor \Delta_u$$

Count Δ_u when its **first** variable has changed

CERN example from Hama, Mase, O (2022)

Test Mode	Methods	Mean	Std. Error
Insertion	Kernel SHAP	18.535	0.215
	Integrated Gradients	18.289	0.213
	DeepLIFT	18.118	0.211
	Vanilla Grad	−1.310	0.252
	Input × Gradient	7.620	0.216
	LIME	17.319	0.209
	Random	−0.380	0.175
Deletion	Kernel SHAP	16.752	0.176
	Integrated Gradients	16.315	0.173
	DeepLIFT	16.646	0.176
	Vanilla Grad	0.226	0.256
	Input × Gradient	7.940	0.187
	LIME	15.845	0.170
	Random	−0.268	0.179

Mean insertion / deletion ABCs for 2000 CERN electron collision data point pairs

Shapley value

Method from game theory

Shapley (1953)

Very popular in local attribution

Lundberg, Lee (2017)

Global attribution

O (2014) JUQ

Song, Nelson, Staum (2016) JUQ

O and Prieur (2017) JUQ

Shapley handles dependence well

Shapley setup

Team $u \subseteq 1{:}d \equiv \{1, 2, \ldots, d\}$

creates value $\nu(u)$.

Total value is $\nu(1{:}d)$.

$\phi_j =$ 'fair' share for player j.

Incremental value of j given u

$$\nu(j \mid u) = \nu(u \cup \{j\}) - \nu(u)$$

Shapley axioms

Efficiency	$\sum_{j=1}^{d} \phi_j = \nu(1{:}d)$
Dummy	If $\nu(j \mid u) = 0$, all u then $\phi_j = 0$
Symmetry	If $\nu(i \mid u) = \nu(j \mid u)$, when $u \cap \{i, j\} = \varnothing$ then $\phi_i = \phi_j$
Additivity	If games ν, ν' have values ϕ, ϕ' then $\nu + \nu'$ has value $\phi_j + \phi'_j$

Shapley solution (unique)

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \nu(j \mid u)$$

Another formulation

Build a team from $\varnothing$ to $1{:}d$

in all $d!$ orders

ϕ_j is the average of $\nu(j \mid \cdot)$

Variable importance

Define $\nu(u)$ as predictive power for $\boldsymbol{x}_u$

very many choices

Shapley examples

$$\nu(u) = \mathrm{Var}(\mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x}_u)) \quad \text{Variance Shapley}$$

$$\nu(u) = f(\boldsymbol{x}_u{:}\boldsymbol{x}'_{-u}) \quad \text{Baseline Shapley}$$

$$\nu(u) = \mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x}_u) \quad \text{Cond. expect. Shapley}$$

See Najmi & Sundararajan (2020)

Cohort Shapley

Motivation:

avoid impossible combinations

only use actually observed combinations

Mase, Seiler, O (2019) arXiv:1911.00467 (for fairness)

Mase, Seiler, O (2024) Annual Rev Stat and its Applications

Similarity

Target has $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{td})$.

Define

$$z_{ij} = z_{ij}(t) = \begin{cases} 1, & x_{ij} \quad \text{'similar' to} \quad x_{tj} \\ 0, & \text{else.} \end{cases}$$

E.g.: $x_{ij} = x_{tj}$, or $|x_{ij} - x_{tj}| \leqslant \delta_j$

Toy example

	Subj	Color	Breakfast	$Z_{i1}(5)$	$Z_{i2}(5)$	$Z_{i,\{1,2\}}(5)$
	1	red	eggs	0	1	0
	2	red	cereal	0	0	0
	3	blue	cereal	1	0	0
	4	red	eggs	0	1	0
Target	5	blue	eggs	1	1	1

Cohorts

$$\begin{array}{lcc} & \{1,4,5\} & \rightarrow & \{5\} \\ \text{similar food} & \uparrow & & \uparrow \\ & \{1,2,3,4,5\} & \rightarrow & \{3,5\} \end{array}$$

similar color

Toy continued

$$\begin{array}{lccc} & \{1,4,5\} & \rightarrow & \{5\} \\ \text{similar food} & \uparrow & & \uparrow \\ & \{1,2,3,4,5\} & \rightarrow & \{3,5\} \end{array}$$

similar color

Similarity constraints

$$\begin{array}{lccc} & \{2\} & \rightarrow & \{1,2\} \\ \text{similar food} & \uparrow & & \uparrow \\ & \varnothing & \rightarrow & \{1\} \end{array}$$

similar color

Value function

Cohorts

$$C_{t,u} = \{i \in 1{:}n \mid z_{ij}(t) = 1, \text{ all } j \in u\}$$

Cohort means

$$\nu(u) = \nu(u;t) \equiv \bar{y}_{t,u} = \frac{1}{|C_{t,u}|} \sum_{i \in C_{t,u}} f(\boldsymbol{x}_i)$$

Cohort refinement

Start with

$$C_{t,\varnothing} = \{1, 2, \ldots, n\}$$

Each j added to u refines the cohort by removing dissimilar subjects.

Important j move the cohort means the most

Integrated gradients

Aumann-Shapley value also Najmi and Sundararajan

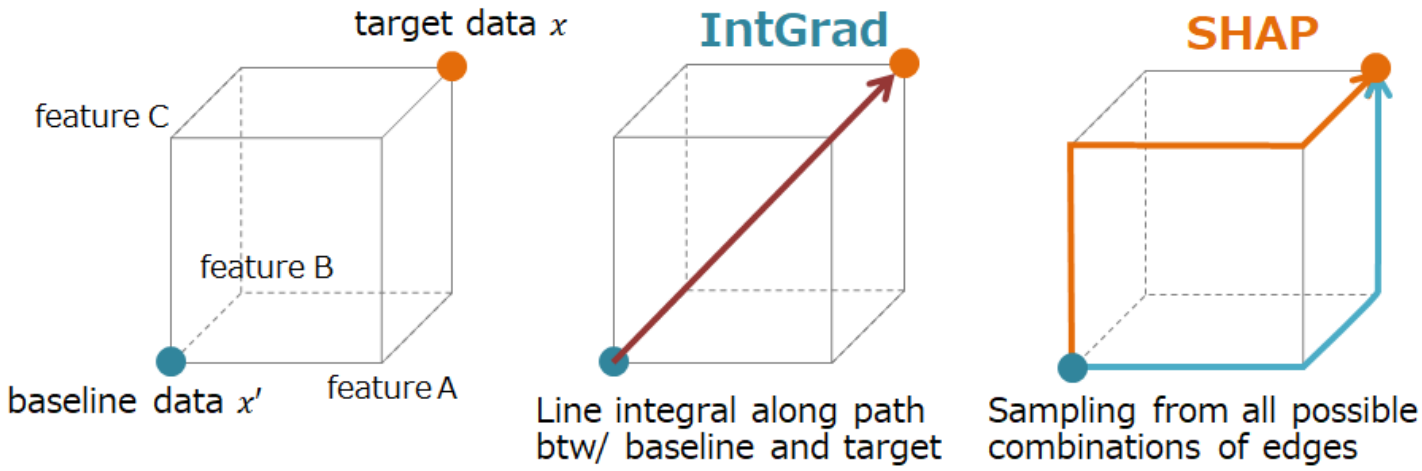
Replace $f : \{0,1\}^d \to \mathbb{R}$ by
smooth $f : [0,1]^d \to \mathbb{R}$

Approx Shapley value ϕ_j by

$$\psi_j = \int_0^1 \frac{\partial}{\partial x_j} f(t,t,\ldots,t)\,\mathrm{d}t$$

Integrate the gradient over the diagonal in $[0,1]^d$

Avoids exponential cost



Integrated gradients cohort Shapley

Introduce soft similarity in $[0,1]$

Replace $g : \{0,1\}^d \to \mathbb{R}$ by $\tilde{g} : [0,1]^d \to \mathbb{R}$

Use integrated gradients of $\tilde{g}$

Results

Find close match between CS and IGCS

for large d

Applications to physics and chemistry

Hama, Mase, O arXiv:2211.08414

Thanks

1) Co-authors:

Masayoshi Mase, Naofumi Hama, Christopher Hoyt, Ben Seiler

2) Funding:

NSF IIS-1837931 and DMS-2152780, and Hitachi, Ltd.

3) Invitation:

Takashi Goda, Yoshuhito Kazashi

Finally

There is no ground truth for variable importance!