QMC and Variable Importance

Art B. Owen

Stanford University

These slides are from a series of four lectures given at the Johann Radon Institute for Computational and Applied Mathematics (RICAM) held on March 24 and March 25 2021.

It was an honor to be asked to present on quasi-Monte Carlo (QMC) sampling in Austria, from where so much of QMC comes and has come. The talks were virtual; I would have otherwise made sure to get some Linzertorte. That will have to wait.

1. Quasi-Monte Carlo
2. Randomized Quasi-Monte Carlo
3. QMC Beyond the Cube
4. QMC and Variable Importance

A small number of corrections have been made since then.

Black box functions

$$y = f(\boldsymbol{x}), \quad \text{where}$$

$$\boldsymbol{x} = (x_1, x_2, \dots, x_d) \in \mathcal{X} \subseteq \prod_{j=1}^{d} \mathcal{X}_j$$

Questions

- How important is x_j?
- How important is $\boldsymbol{x}_u$ for $u \subset 1{:}d$?

Context

Models in science and engineering:

semiconductors, aerospace, malaria control, climate models, $\cdots$

Black box predictions:

AI / machine learning

Quasi-Monte Carlo

Effective dimension

What is importance?

A variable is important if changing it changes something else that is important.

Local sensitivity

$$\frac{\partial}{\partial x_j} f(\boldsymbol{x}_0) \qquad \text{et cetera}$$

Global sensitivity

Make random changes to $\boldsymbol{x}_u$ keeping $\boldsymbol{x}_{-u}$ fixed

study $\mathbb{E}((f(\boldsymbol{x}) - f(\boldsymbol{x}'))^2)$

where $\boldsymbol{x}_u \neq \boldsymbol{x}'_u$ and $\boldsymbol{x}_{-u} = \boldsymbol{x}'_{-u}$

Global sensitivity analysis

Global: consider **all** $\boldsymbol{x}$ not just focal $\boldsymbol{x}_0$

potentially **all** $\boldsymbol{x}' \neq \boldsymbol{x}$ not just local

Some references

For books giving context and uses see:

Fang, Li & Sudijanto (2010), Saltelli, Chan & Scott (2009), Saltelli, Ratto & Andres (2008), Cacuci, Ionescu-Bujor & Navon (2005), Saltelli, Tarantola & Campolongo (2004), Santner, Williams & Notz (2003)

Many scientific communities participate, many terms:

FANOVA DACE FAST SAMO MASCOT UCM HDMR NPUA UQ

Major survey paper

Razavi et al. (2021) 26 authors

Environmental Modelling and Software

Outline

1) ANOVA and notation
2) Sobol' indices and mean dimension
3) Mean dimension of ridge functions
4) Mean dimension of a neural network
5) Shapley value

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

Origins: Hoeffding (1948) Sobol' (1969) Efron & Stein (1981)

Notation

For $u \subseteq 1{:}d \equiv \{1,\dots,d\}$

$$|u| = \mathbf{card}(u)$$
$$-u = u^c = \{1,2,\dots,d\} - u$$

If $u = \{j_1, j_2, \dots, j_{|u|}\}$ then $\boldsymbol{x}_u = (x_{j_1}, \dots, x_{j_{|u|}})$ and $\mathrm{d}\boldsymbol{x}_u = \prod_{j\in u} \mathrm{d}x_j$

Decomposition

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

$f_u(\boldsymbol{x})$ only depends on x_j for $j \in u$.

ANOVA properties

$$j \in u \implies \int_0^1 f_u(\boldsymbol{x}) \, \mathrm{d}x_j = 0$$

$$u \neq v \implies \int f_u(\boldsymbol{x}) f_v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 0$$

Variances

$$\mathrm{Var}(f) \equiv \int (f(\boldsymbol{x}) - \mu)^2 \, \mathrm{d}\boldsymbol{x} = \sum_{u \subseteq 1:d} \sigma_u^2$$

$$\sigma_u^2 = \sigma_u^2(f) = \begin{cases} \int f_u(\boldsymbol{x})^2 \, \mathrm{d}\boldsymbol{x} & u \neq \varnothing \\ 0 & u = \varnothing. \end{cases}$$

General L^2

Independence is critical

Uniformity is not

Sobol' indices

$$\underline{\tau}_u^2 = \sum_{v\subseteq u} \sigma_v^2$$ v contained in u Sobol' (1993)

$$\overline{\tau}_u^2 = \sum_{v\cap u\neq\varnothing} \sigma_v^2$$ v 'touches' u Homma & Saltelli (1996)

Large $\underline{\tau}_u^2$ means $\boldsymbol{x}_u$ important

Small $\overline{\tau}_u^2$ means $\boldsymbol{x}_u$ unimportant can be frozen Sobol'

Normalization

$\dfrac{\underline{\tau}_u^2}{\sigma^2}$ and $\dfrac{\overline{\tau}_u^2}{\sigma^2}$ are like R^2 measures for $\boldsymbol{x}_u$

Examples

$d = 4$ and $u = \{1, 2\}$

$$\underline{\tau}^2_{\{1,2\}} = \sigma^2_{\{1\}} + \sigma^2_{\{2\}} + \sigma^2_{\{1,2\}}$$

$$\begin{aligned}\overline{\tau}^2_{\{1,2\}} = \sigma^2_{\{1\}} &+ \sigma^2_{\{2\}} + \sigma^2_{\{1,2\}} \\ &+ \sigma^2_{\{1,3\}} + \sigma^2_{\{1,4\}} + \sigma^2_{\{2,3\}} + \sigma^2_{\{2,4\}} \\ &+ \sigma^2_{\{1,3,4\}} + \sigma^2_{\{2,3,4\}} + \sigma^2_{\{1,2,3\}} + \sigma^2_{\{1,2,4\}} \\ &+ \sigma^2_{\{1,2,3,4\}}\end{aligned}$$

Identity

$$\underline{\tau}^2_u + \overline{\tau}^2_{-u} = \sigma^2$$

Variance explained

$\underline{\tau}_u^2$ is the variance 'explained by' $\boldsymbol{x}_u$

For $j \in v$, $\int_0^1 f_v(\boldsymbol{x})\,\mathrm{d}x_j = 0$. $(*)$

$$
\begin{aligned}
\mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x}_u) &= \sum_{v \subseteq 1:d} \mathbb{E}(f_v(\boldsymbol{x}) \mid \boldsymbol{x}_u) \\
&= \sum_{v \subseteq u} \mathbb{E}(f_v(\boldsymbol{x}) \mid \boldsymbol{x}_u) \quad \text{by } (*) \\
&= \sum_{v \subseteq u} f_v(\boldsymbol{x})
\end{aligned}
$$

As a Sobol' index

$$
\mathrm{Var}\big(\mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x}_u)\big) = \mathrm{Var}\Bigg(\sum_{v \subseteq u} f_v(\boldsymbol{x})\Bigg) = \sum_{v \subseteq u} \sigma_v^2 \equiv \underline{\tau}_u^2
$$

Hybrid points

$\boldsymbol{y} = \boldsymbol{x}_u{:}\boldsymbol{z}_{-u}$ means

$$y_j = \begin{cases} x_j, & j \in u \\ z_j, & j \notin u. \end{cases}$$

Example

$$\boldsymbol{x} = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$$
$$\boldsymbol{z} = (0.9, 0.8, 0.7, 0.6, 0.5, 0.4)$$
$$\boldsymbol{x}_{\{1,2,4,5\}}{:}\boldsymbol{z}_{\{3,6\}} = (0.1, 0.2, 0.7, 0.4, 0.5, 0.4)$$

Glue

: is a 'glue operator'.

We glue $\boldsymbol{x}_u$ to $\boldsymbol{z}_{-u}$ to get $\boldsymbol{x}_u{:}\boldsymbol{z}_{-u} \in [0,1]^d$

An identity

For $\boldsymbol{x}, \boldsymbol{z} \overset{\text{iid}}{\sim} \mathbf{U}[0,1]^d$

$$
\begin{aligned}
\mathbb{E}(f(\boldsymbol{x})f(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u})) &= \sum_{v\subseteq 1:d}\sum_{v'\subseteq 1:d} \mathbb{E}(f_v(\boldsymbol{x})f_{v'}(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u}))\\
&= \sum_{v\subseteq 1:d} \mathbb{E}(f_v(\boldsymbol{x})f_v(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u})) \qquad \text{orthogonality}\\
&= \sum_{v\subseteq u} \mathbb{E}(f_v(\boldsymbol{x})f_v(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u})) \qquad \text{line integral on } z_j\\
&= \sum_{v\subseteq u} \mathbb{E}(f_v(\boldsymbol{x})^2) \qquad f_v \text{ does not depend on } \boldsymbol{z}_{-u}\\
&= \mu^2 + \underline{\tau}_u^2
\end{aligned}
$$

From Sobol' (1993)

Pick and freeze methods

Evaluate f at two points:

Freeze:	keep some components	e.g., $\boldsymbol{x}_u \to \boldsymbol{x}_u$
Pick:	change the others	e.g., $\boldsymbol{x}_{-u} \to \boldsymbol{z}_{-u}$

Identities

$$\underline{\tau}_u^2 = \int f(\boldsymbol{x}) f(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} - \mu^2 \qquad \text{Sobol' (1993)}$$

$$\overline{\tau}_u^2 = \frac{1}{2}\int ((f(\boldsymbol{x}) - f(\boldsymbol{x}_{-u}{:}\boldsymbol{z}_u))^2 \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} \qquad \text{Jansen (1999)}$$

Use MC, QMC, RQMC

There are many such identities

Like tomography

global integrals reveal internal structure

we don't have to estimate any f_u

Mean dimension

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

Low mean dimension $\implies$ f dominated by low dimensional aspects

Identity from Liu & O (2006)

$$\begin{aligned}\sum_{j=1}^{d} \overline{\tau}_j^2 &= \sum_{j=1}^{d} \sum_{u \subseteq 1:d} 1_{j \in u} \sigma_u^2 \\ &= \sum_{u \subseteq 1:d} \sum_{j=1}^{d} 1_{j \in u} \sigma_u^2 \\ &= \sum_{u \subseteq 1:d} |u| \sigma_u^2\end{aligned}$$

Answer from 2^d variance components equals

sum of d Sobol' indices

Example

Kuo, Schwab, Sloan (2011)

$$f(\boldsymbol{x}) = \frac{1}{1 + \sum_{j=1}^{500} x_j / j!}, \quad \boldsymbol{x} \sim \mathbf{U}[0,1]^{500}$$

Find numerically that

$$1.00356 \leqslant \nu(f) \leqslant 1.00684 \qquad \text{(99\% confidence)}$$

vs effective dimension

1) easier to compute
2) not defined via 0.99 or other threshold
3) not restricted to integer values

Ridge functions

$$f(\boldsymbol{x}) = g(\theta^\mathsf{T}\boldsymbol{x}), \qquad \boldsymbol{x}, \theta \in \mathbb{R}^d$$

More generally

$$f(\boldsymbol{x}) = g(\Theta^\mathsf{T}\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^d$$

$$\Theta \in \mathbb{R}^{d\times r}, \qquad r < d$$

$$g : \mathbb{R}^r \to \mathbb{R}$$

We're interested in

$$r \ll d$$

Normalization

$$\Theta^\mathsf{T}\Theta = I_r \qquad \theta^\mathsf{T}\theta = 1$$

Ridge functions and QMC

1) P. Constantine (2015) ridge functions are ubiquitous in physical sciences and engineering **(active subspaces)**

2) Integrands that are dominated by low dimensional aspects are favorable for QMC

3) Smooth ridge functions are dominated by low dimensional aspects

Therefore

QMC should often be very effective in the physical sciences

Finer print

1) usually $f(\boldsymbol{x}) \approx g(\Theta^{\mathsf{T}}\boldsymbol{x})$

2) the low dimensional parts ought to be regular

Commonly true Griebel, Kuo, Sloan (2013, 2017)

3) we use mean dimension

Spoiler

Sometimes mean dim is $O(1)$ as $d \to \infty$

Sometimes mean dim is $O(\sqrt{d})$ as $d \to \infty$

What matters

1) smoothness of $g(\cdot)$
2) sparsity of θ or Θ

Gaussian setting

$$\boldsymbol{x} \sim \mathcal{N}(0, I_d) \implies \boldsymbol{z} = \Theta^{\mathsf{T}}\boldsymbol{x} \sim \mathcal{N}(0, I_r)$$
$$\implies \boldsymbol{z} = \theta^{\mathsf{T}}\boldsymbol{x} \sim \mathcal{N}(0, 1)$$

Moments

Let $\varphi(\boldsymbol{x})$ be $\mathcal{N}(0, I)$ density

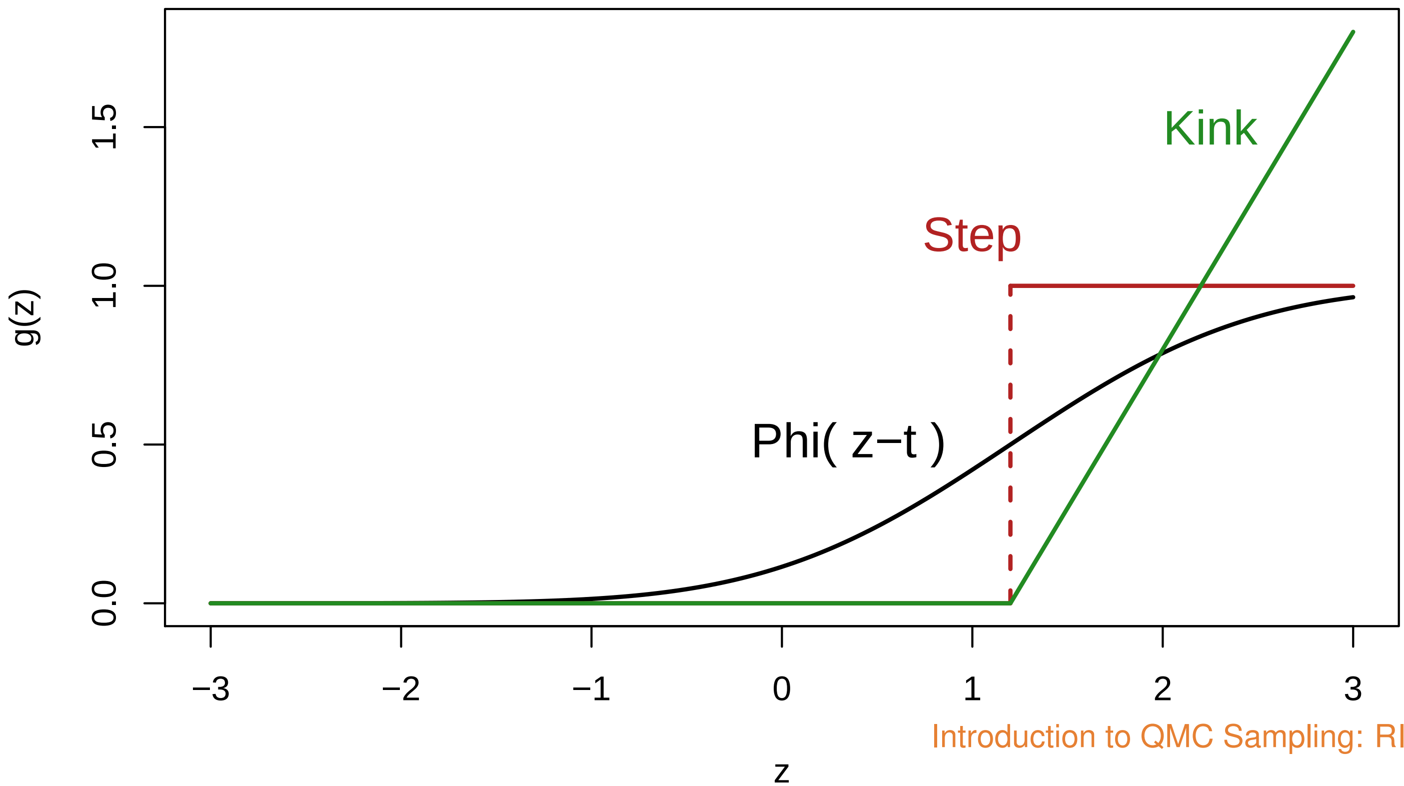
$$\mu = \int_{\mathbb{R}^d} g(\Theta^{\mathsf{T}}\boldsymbol{x})\varphi_d(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^r} g(\boldsymbol{z})\varphi_r(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$$
$$\sigma^2 = \int_{\mathbb{R}^d} \big(g(\Theta^{\mathsf{T}}\boldsymbol{x}) - \mu\big)^2\varphi_d(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^r} \big(g(\boldsymbol{z}) - \mu\big)^2\varphi_r(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$$

Monte Carlo rate independent of d

One dimensional ridge functions

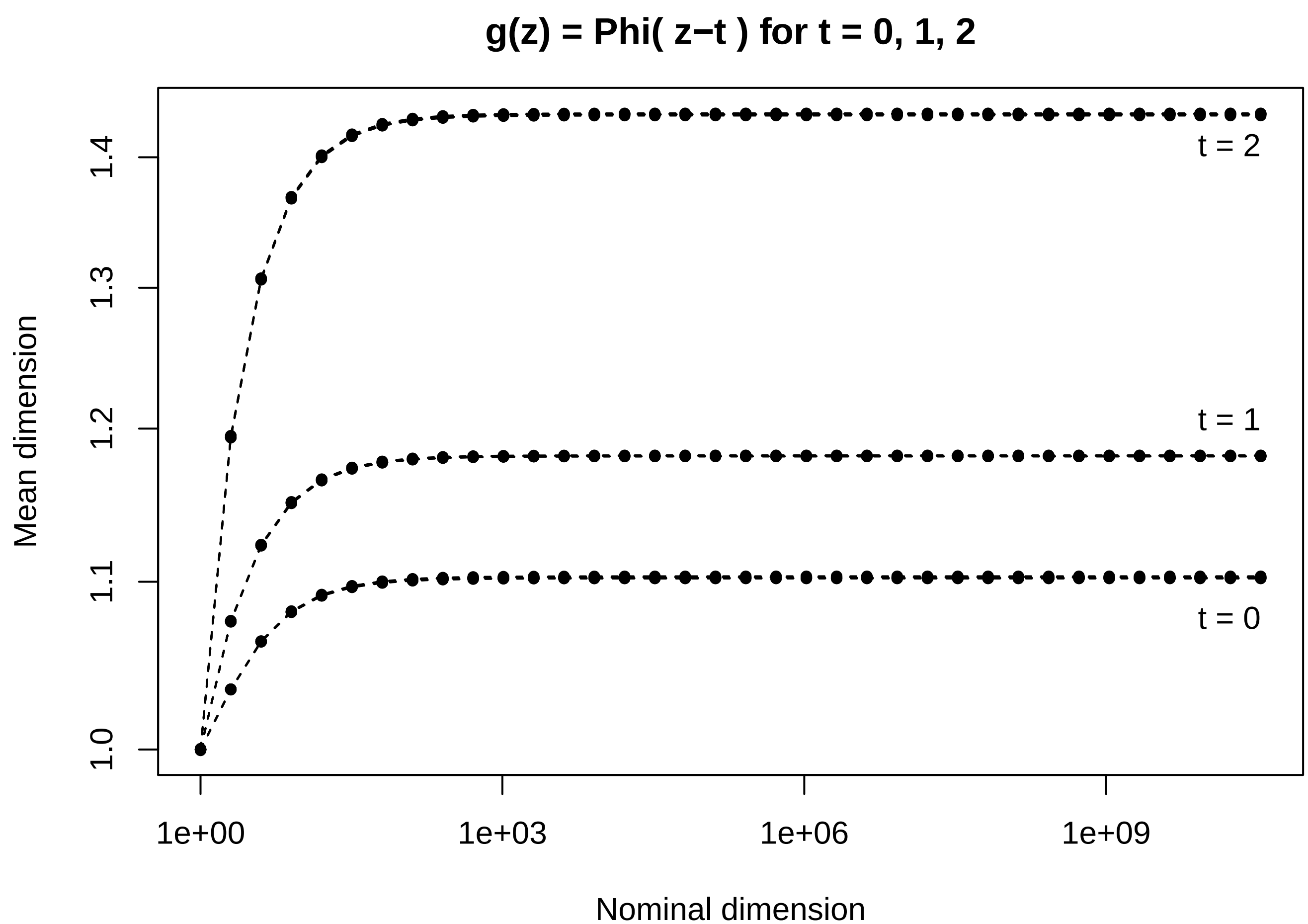
$$\boldsymbol{x} \sim \mathcal{N}(0, I) \qquad f(\boldsymbol{x}) = g\biggl(\frac{1}{\sqrt{d}}\sum_{j=1}^{d} x_j\biggr) = g\bigl(\mathcal{N}(0,1)\bigr) \qquad \theta = \frac{1_d}{\sqrt{d}}$$

Examples of $g(z)$



Smooth g

$$f(\boldsymbol{x}) = g\Bigl(\frac{1}{\sqrt{d}}\sum_{j=1}^{d} x_j\Bigr) \qquad g(z) = \Phi(z-t) = \int_{-\infty}^{z-t} \varphi(y)\,\mathrm{d}y$$

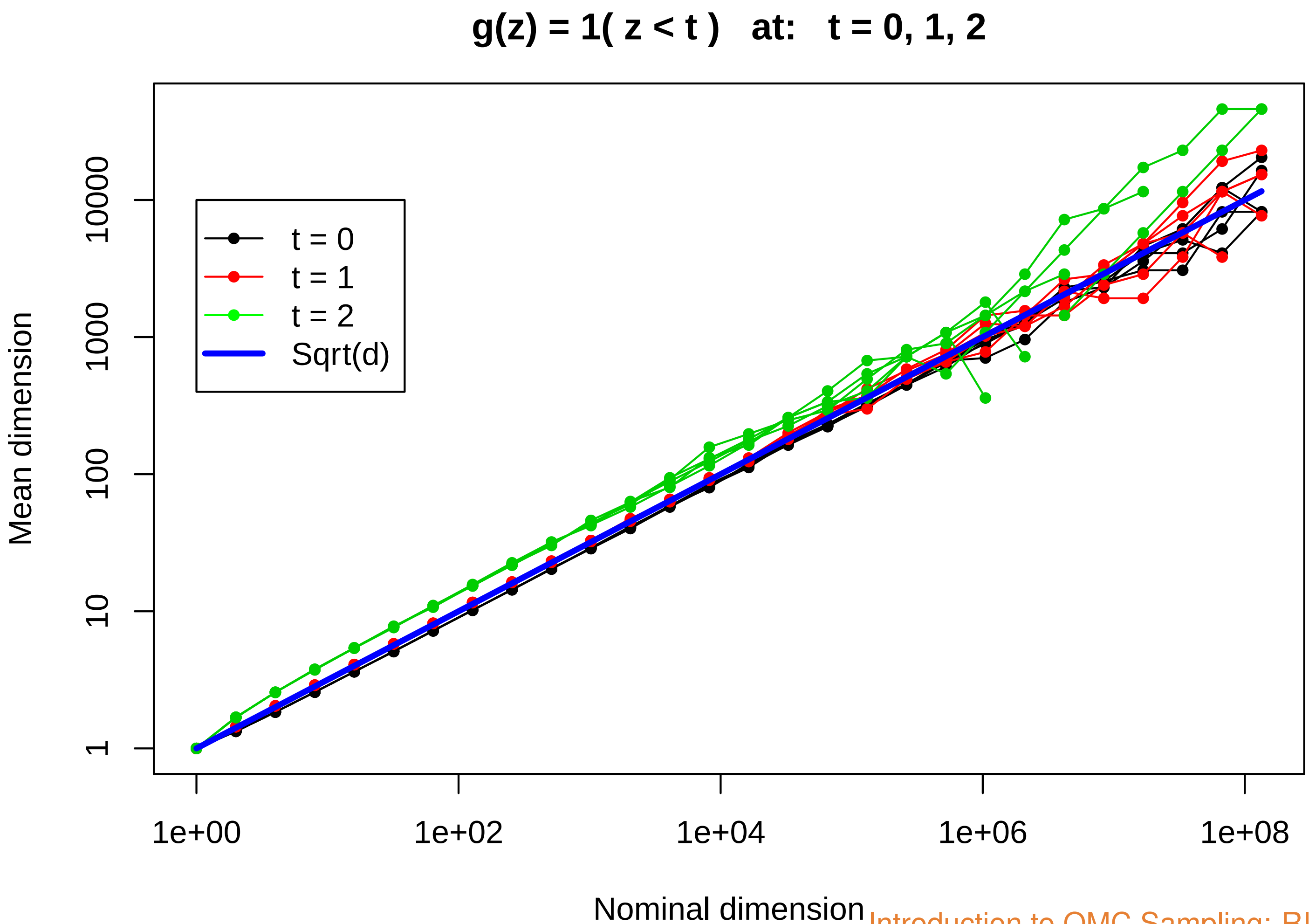


Rand. Sobol' points on Sobol' index $n = 2^{14}$; 5 repeats.

$1 \leqslant \log_2(d) \leqslant 35 \qquad \overline{\tau}_j^2$ all the same

Jump g

$$f(\boldsymbol{x}) = g\biggl(\frac{1}{\sqrt{d}}\sum_{j=1}^{d} x_j\biggr) \qquad g(z) = 1\{z < t\}$$



Via randomized Sobol' $n = 2^{16}$; 5 repeats.

Kink g

$$f(\boldsymbol{x}) = g\biggl(\frac{1}{\sqrt{d}}\sum_{j=1}^{d} x_j\biggr)$$

$$g(z) = \max(0, z - t)$$

Puzzler

Is a kink like a step or like $\Phi(z-t)$?

Infinite Hardy-Krause variation like the step

Has weak derivative like $\Phi(z-t)$

Kink is a once integrated step

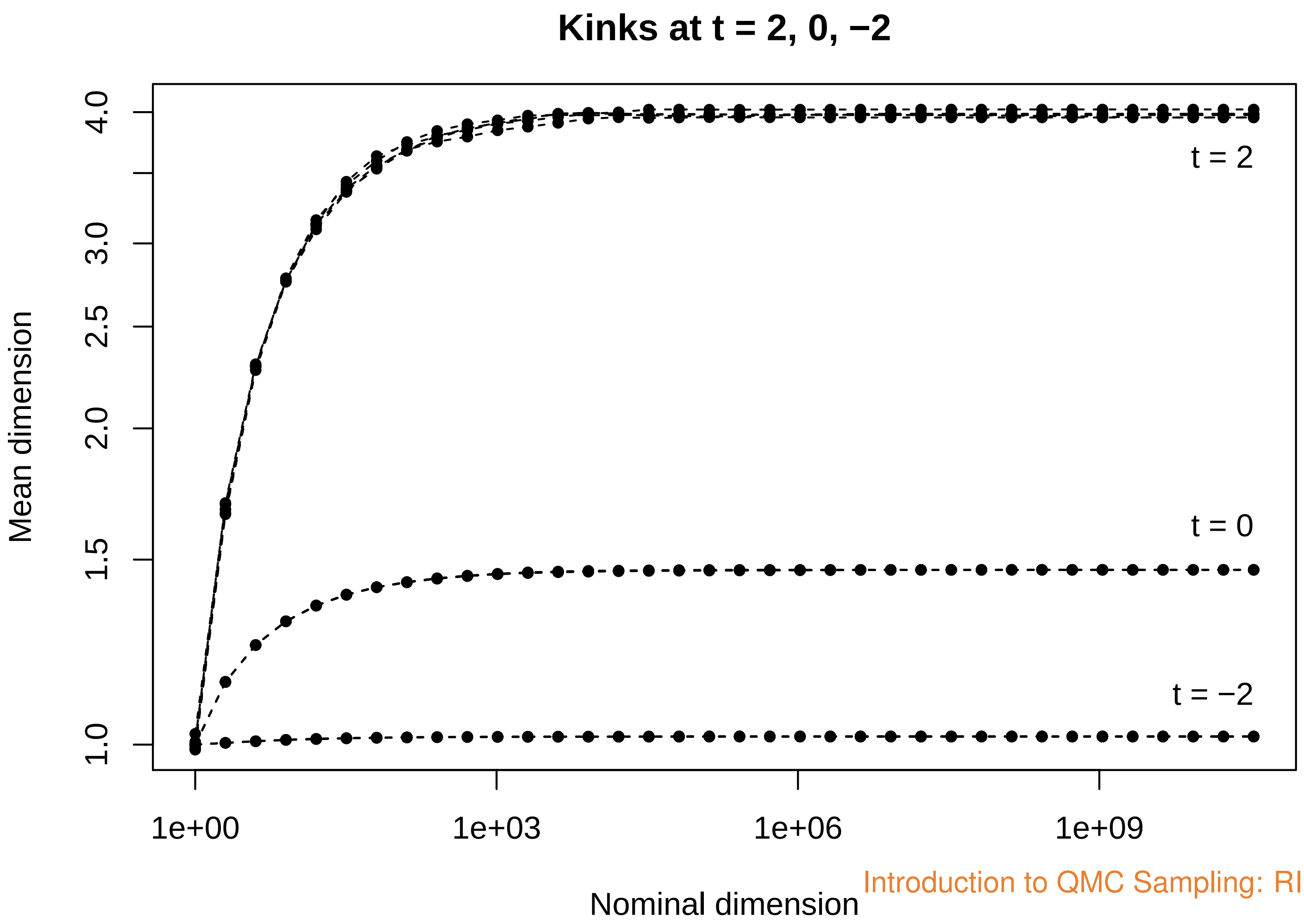
see Griewank, Kuo, Leövey & Sloan (2018)

For jump or kink: f_u is smooth for $|u| < d$

Griebel, Kuo & Sloan (2013, 2017)

Kink g

$$f(\boldsymbol{x}) = g\Bigl(\frac{1}{\sqrt{d}}\sum_{j=1}^{d} x_j\Bigr) \qquad g(z) = \max(z-t, 0)$$



Theorem

Theorem 3.1 of Hoyt & O (2020) upper bound for $\nu(f)$ when

$f(\boldsymbol{x}) = g(\boldsymbol{x}^\mathsf{T}\Theta)$ and $g(\cdot)$ is Hölder α

Lipschitz corollary

If $f(\boldsymbol{x}) = g(\Theta^\mathsf{T}\boldsymbol{x})$ for $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$ with $\Theta^\mathsf{T}\Theta = I_r$ then

$$\nu(f) \leqslant r \times \frac{C^2}{\sigma^2}$$

where g is Lipschitz C with $\sigma^2 = \mathrm{Var}(f(\boldsymbol{x})) = \mathrm{Var}(g(\boldsymbol{z}))$.

$r = 1$ corollary

For $d \geqslant r = 1$ and $|g(\boldsymbol{z}) - g(\boldsymbol{z}')| \leqslant C\|\boldsymbol{z} - \boldsymbol{z}'\|^\alpha \quad 0 < \alpha \leqslant 1$

$$\nu(f) = O(d^{1-\alpha})$$

Implied constants in Hoyt & O (2020)

Mean dimension of a neural network

Paper is Hoyt & O (2021)

Journal of Uncertainty Quantification

Theorem is about how to compute

$$\sum_{j=1}^{d} \hat{\overline{\tau}}^{2}_{\{j\}}$$

efficiently handling $O(d^2)$ correlations

Variant of **Winding stairs** works best

Jansen, Rossing, Daamen (1994)

Example

Mean dimension of a neural network on $28 \times 28 = 768$ pixels

MNIST data

Digits $0, 1, \ldots, 9$ in 28×28 gray level images

$70{,}000$ images from LeCun

Neural net architecture from Yalcin (2018)

Second last layer produces $g_0(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_9(\boldsymbol{x})$.

Last layer (softmax) is

$$f_y(\boldsymbol{x}) = \frac{\exp(g_y(\boldsymbol{x}))}{\sum_{j=0}^{9} \exp(g_j(\boldsymbol{x}))}$$

Mean dimension

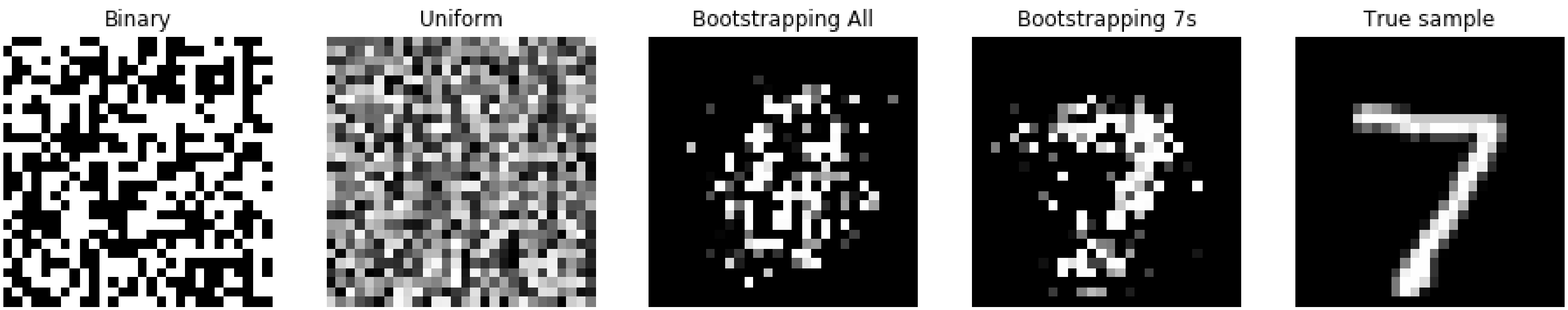
The g_j had modest mean dimension

The f_j not so much

Problem

Very unrealistic independence models on the 768 pixel values

Independent pixel distributions



From left to right

$\mathbf{U}\{0,1\}^{768}$

$\mathbf{U}[0,1]^{768}$

resample pixels independently from data

resample pixels (one class)

Example image

$\nu(f_y)$ with softmax

Samp.	0	1	2	3	4	5	6	7	8	9
binary	11.07	936.04	10.43	9.92	18.69	10.22	13.27	13.37	8.67	16.54
unif.	6.92	4,108.99	7.28	6.60	9.90	7.03	6.92	8.03	5.61	9.48
comb.	8.77	4.68	4.06	3.95	4.56	5.11	7.62	4.62	3.43	7.39
0	3.52	6.81	3.48	7.20	6.56	5.78	7.54	4.67	4.04	9.08
1	36.12	2.88	6.00	3.43	7.75	3.76	8.74	7.60	2.83	5.58
2	10.03	3.86	3.68	4.70	8.23	12.27	12.57	7.20	4.31	17.23
3	23.20	4.69	5.95	4.10	6.96	6.72	13.63	7.10	4.42	9.00
4	7.42	8.39	7.59	9.96	3.81	7.63	8.57	5.35	3.86	6.82
5	8.12	4.77	5.72	4.82	5.60	3.48	7.61	7.28	3.54	7.87
6	9.22	5.65	4.36	6.52	4.31	6.67	3.57	6.43	4.28	11.99
7	8.57	5.85	4.42	4.09	4.66	5.09	3.59	3.59	4.29	5.58
8	19.58	6.06	4.54	4.77	8.21	6.28	13.15	6.72	4.20	10.11
9	7.47	7.00	5.25	4.96	3.15	4.52	7.34	3.74	2.92	3.48

$\nu(g_y)$ without softmax

Samp.	0	1	2	3	4	5	6	7	8	9
binary	1.66	1.76	1.74	1.72	1.73	1.79	1.75	1.69	1.74	1.79
unif.	1.65	1.62	1.66	1.66	1.67	1.71	1.71	1.61	1.68	1.70
comb.	1.79	1.77	1.70	1.73	1.73	1.90	1.88	1.78	1.90	1.89
0	1.92	1.65	1.68	1.69	1.65	1.80	1.86	1.56	1.68	1.81
1	1.48	1.56	1.35	1.61	1.62	1.57	1.49	1.42	1.56	1.50
2	1.55	1.66	1.62	1.74	1.57	1.72	1.67	1.61	1.78	1.59
3	1.56	1.65	1.59	1.58	1.63	1.85	1.59	1.64	1.67	1.66
4	1.87	1.62	1.61	1.55	1.70	1.75	1.76	1.66	1.57	1.78
5	1.71	1.60	1.59	1.63	1.72	1.78	1.74	1.62	1.76	1.90
6	1.65	1.60	1.60	1.66	1.68	1.70	1.65	1.60	1.54	1.63
7	1.73	1.59	1.61	1.63	1.60	1.62	1.65	1.57	1.59	1.63
8	1.73	1.65	1.60	1.64	1.66	1.78	1.75	1.64	1.84	1.75
9	1.86	1.68	1.61	1.63	1.73	1.80	1.86	1.67	1.69	1.82

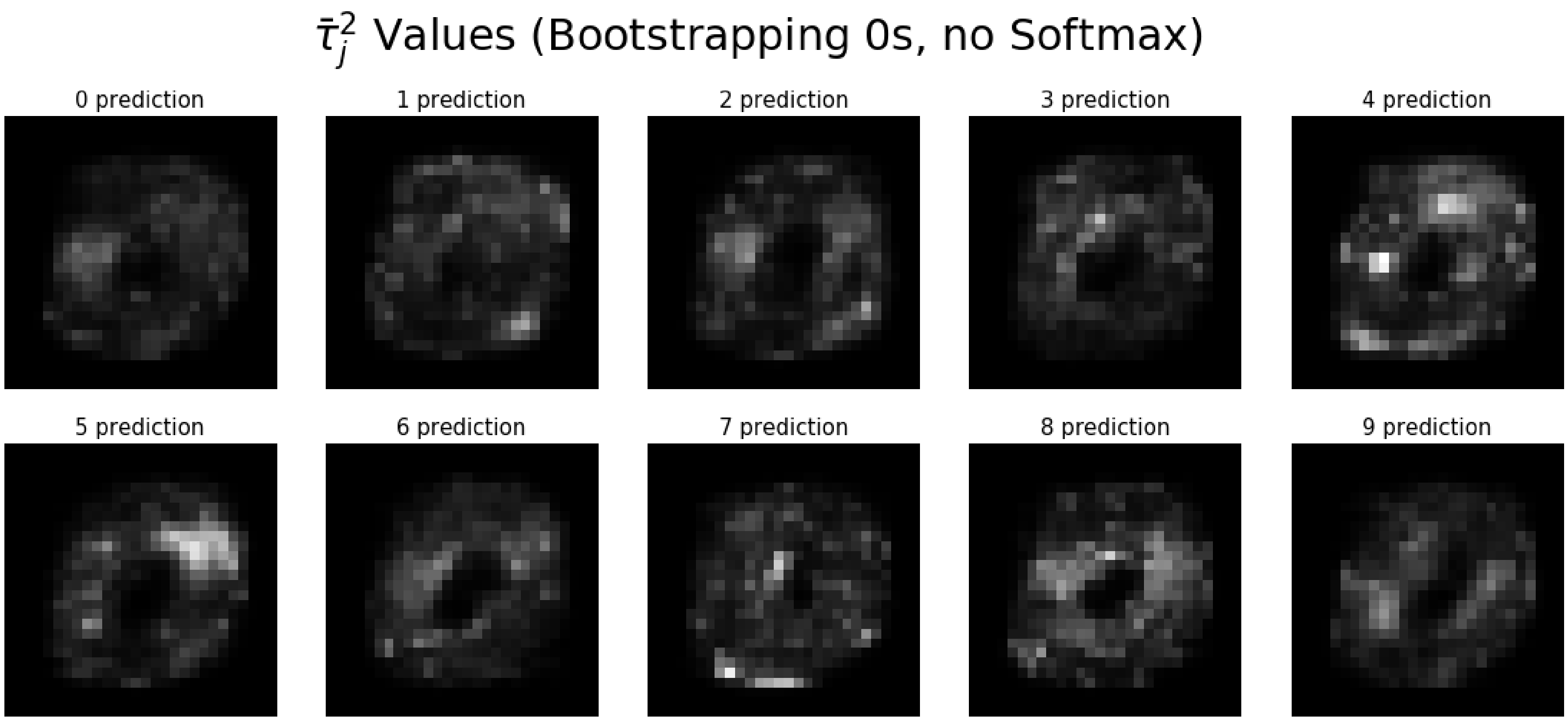
Pixel importance maps

Functions are g_y for predicting $Y = y$

Map shows $\overline{\tau}_j^2$ for 786 pixels j

Larger $=$ brighter

Resampling pixels from $y = 0$ images



Next topic:

Shapley value

1) connects to Sobol' indices
2) bridge to variable importance in machine learning

15 million Euros

Shapley's (1953) value can be used to quantify the contribution of members to a team.

We need to know what each subset of the team would have accomplished.

Example from Bank of International Settlement

Team	Output value in €
∅	0
A	4,000,000
B	4,000,000
C	4,000,000
A,B	9,000,000
A,C	10,000,000
B,C	11,000,000
A,B,C	15,000,000

Q: How should we split the €15,000,000 earned by A, B, C among them?

15 million Euros

Shapley's (1953) value can be used to quantify the contribution of members to a team.

We need to know what each subset of the team would have accomplished.

Example from Bank of International Settlement

Team	Output value in €
∅	0
A	4,000,000
B	4,000,000
C	4,000,000
A,B	9,000,000
A,C	10,000,000
B,C	11,000,000
A,B,C	15,000,000

Q: How should we split the €15,000,000 earned by A, B, C among them?

A: Shapley says: A gets €4,500,000, B gets €5,000,000, C gets €5,500,000

Shapley setup

Let team $u \subseteq 1{:}d \equiv \{1,2,\dots,d\}$ create value **val**(u).

Total value is **val**$(1{:}d)$.

We attribute ϕ_j of this to $j \in 1{:}d$.

Shapley axioms

Efficiency	$\sum_{j=1}^{d} \phi_j = \mathbf{val}(1{:}d)$
Dummy	If **val**$(u \cup \{i\}) =$ **val**(u), all u then $\phi_i = 0$
Symmetry	If **val**$(u \cup \{i\}) =$ **val**$(u \cup \{j\})$, all $u \cap \{i,j\} = \varnothing$ then $\phi_i = \phi_j$
Additivity	If games **val**, **val**$'$ have values ϕ, ϕ' then **val** $+$ **val**$'$ has value $\phi_j + \phi_j'$

Shapley (1953) shows there is a unique solution.

Shapley's solution

Letting $u + j = u \cup \{j\}$

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} (\textbf{val}(u+j) - \textbf{val}(u))$$

Weighted average of value increments from j

For variable importance

Let variables $x_1, x_2, \dots, x_d$ be team members trying to explain f.

The value of any subset u is how much can be explained by $\boldsymbol{x}_u$.

Choose **val**$(u) \equiv \underline{\tau}_u^2 = \sum_{v \subseteq u} \sigma_v^2$.

Shapley value

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} (\underline{\tau}_{u+j}^2 - \underline{\tau}_u^2)$$

After some algebra

$$\phi_j = \sum_{u:j\in u} \frac{1}{|u|}\sigma_u^2$$

O (2013)

Shapley shares σ_u^2 equally among all $j \in u$.

There seem to be no nice estimation identities like Sobol's.

Bracketing

$$\underline{\tau}_{\{j\}}^2 \leqslant \phi_j \leqslant \overline{\tau}_{\{j\}}^2$$

Shapley

Song, Nelson & Staum (2016) advocate Shapley for dependent case where ANOVA is problematic

- Computation is a challenge.
- They present an approach.
- Apply it to some real-world problems.

O & Prieur (2016)

- Verify that it handles dependence well
- Give special cases / properties

Shapley value for explainable AI

Why was target subject t

- denied a loan?
- sent to the emergency room?
- predicted to commit a crime?

Lundberg & Lee and Najmi & Sundararajan look at $f(\boldsymbol{x}_{t,u}{:}\boldsymbol{z}_{-u})$
for a baseline value such as $\boldsymbol{z} = (1/n)\sum_{i=1}^{n} \boldsymbol{x}_i$

Shapley value for variable j via

$$\textbf{val}(u) = f(\boldsymbol{x}_{t,u}{:}\boldsymbol{z}_{-u}) - f(\boldsymbol{z})$$
$$\textbf{val}(1{:}d) = f(\boldsymbol{x}_{t}) - f(\boldsymbol{z})$$

Cohort Shapley

Mase, O, Seiler (2020) don't like physically impossible combinations

Birth date after graduation

or logically impossible ones

Patient's maximum O_2 below average O_2

Cohort

$C_{t,u}$ subjects i with $\boldsymbol{x}_{i,u} = \boldsymbol{x}_{t,u}$

possibly rounding continuous x_{ij}

val$(u) =$ Average $f(\boldsymbol{x}_i)$ for $i \in C_{t,u}$

Important variables move the average subject predictions towards the target subject

Connection to QMC world

We make use of the **anchored decomposition**

alternative to ANOVA

See Kuo, Sloan, Wasilkowski, Wozniakowski (2010)

Thanks

- Johannes Kepler Universität Linz
- RICAM: Johann·Radon·Institute for Computational and Applied Mathematics
- MCQMC series & Harald Niederreiter
- U.S. NSF: grants up to and including IIS-1837931
- Invitation: Peter Kritzer, Gerhard Larcher, Lucia Del Chicca
- Introductions: Peter Kritzer, Gerhard Larcher, Gunther Leobacher
- Organization: Melanie Traxler

Especially Christopher Hoyt, Clémentine Prieur, Masayoshi Mase, Benjamin Seiler

Some support from Hitachi, Ltd.

The next two backup slides did not get presented.

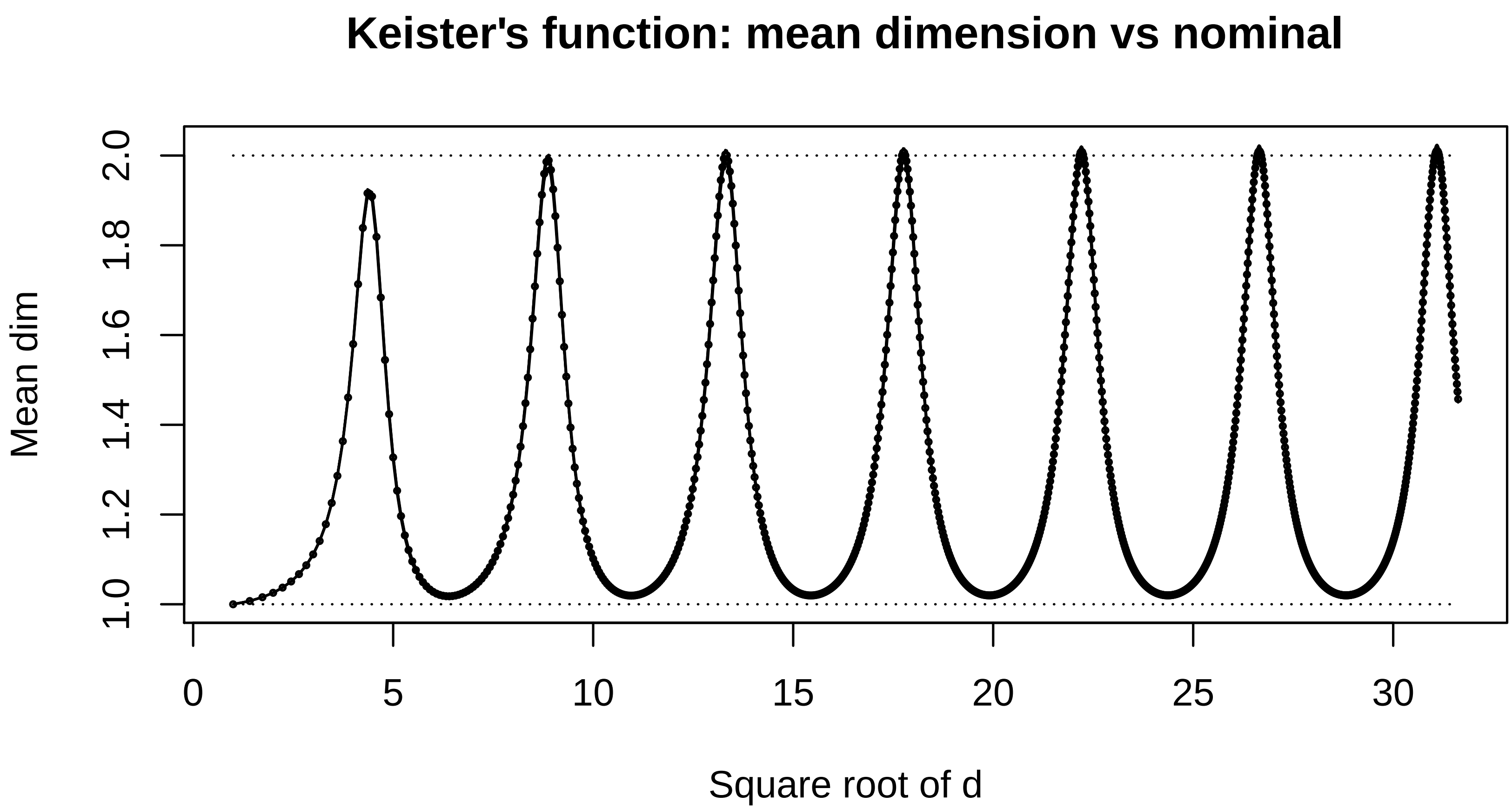
One shows mean dimension of a famous QMC integrand due to Keister.

Another describes how a pre-integration method of Griewank, Kuo, Leövey & Sloan (2018) can reduce mean dimension from $O(\sqrt{d})$ to $O(1)$.

Keister function

$$f(\boldsymbol{x}) = \cos\Bigl(\frac{\|\boldsymbol{x}\|}{2}\Bigr), \qquad \boldsymbol{x} \sim \mathcal{N}(0, I_d)$$

Hoyt & O (in preparation)



Reason: $\dfrac{\|\boldsymbol{x}\|}{2} \mathrel{\dot\sim} \mathcal{N}\Bigl(\dfrac{\sqrt{d}}{2}, \dfrac{1}{4}\Bigr)$ for large d

Preintegration

Integrate out **one** of the d variables

Griewank, Kuo, Leövey, Sloan (2018)

$$\tilde{f}(\boldsymbol{x}) = \tilde{f}_\ell(\boldsymbol{x}) = \int_{-\infty}^{\infty} f(\boldsymbol{x})\varphi(x_\ell)\,\mathrm{d}x_\ell$$

Handle x_ℓ in closed form or by quadrature

Consequences

$\tilde{f}(\boldsymbol{x}) = \tilde{g}(\tilde{\Theta}^\mathsf{T}\boldsymbol{x})$ is also a ridge function

Can get $\tilde{g}(\cdot)$ Lipschitz when $g(\cdot)$ is not. E.g. step function

Can even get

$$\nu(f) = O(\sqrt{d}\,) \quad\longrightarrow\quad \nu(\tilde{f}) = O(1)$$

Good to pre-integrate for $\ell = \arg\max_j |\theta_j|$.

Hoyt & O (2020)