Sampling to understand high dimensional functions

Part I: review

Art B. Owen

Stanford University

This course

This is a two part course.

I) **Review** of material on:

   a) Functional ANOVA

   b) quasi-regression

   c) quasi-Monte Carlo sampling

II) **New** material on:

   a) new Sobol' index quantities

   b) effective dimension of some Sobolev spaces

   c) applications to missing heritability problems

Context

- We want to understand a black box computer program, $y = f(\boldsymbol{x})$
- We could sample it: $(\boldsymbol{x}_i, y_i)$ and analyze the data.
- We may want a fast surrogate $\widetilde{f}(\boldsymbol{x}) \approx f(\boldsymbol{x})$
- Then optimize, integrate, explore, invert $\widetilde{f}$

Alphabet soup

ANOVA DACE FAST SAMO UCM HDMR NPUA

Background

Work with electrical engineers using process, device, and circuit simulators

J. Schott, S. Sharifzadeh

We used Latin hypercubes McKay, Beckman, Conover (1979) and kriging as in Sacks, Ylvisaker (numerous papers and co-authors)

Search for better designs, randomized orthogonal arrays and ultimately randomized quasi-Monte Carlo.

Then some frequentist model fitting, mainly quasi-regression, with Koehler, An, Jiang.

Quadrature ↔ approximation

Given a quadrature oracle, we can approximate:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \beta_{\boldsymbol{k}} e^{2\pi i \boldsymbol{k}^{\mathsf{T}}\boldsymbol{x}}, \quad \text{where}$$

$$\beta_{\boldsymbol{k}} = \int_{[0,1]^d} f(\boldsymbol{x}) e^{-2\pi i \boldsymbol{k}^{\mathsf{T}}\boldsymbol{x}} \, \mathrm{d}\boldsymbol{x}.$$

Given an approximation oracle, we can integrate:

$$\mu = \int f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$\hat{\mu} = \int \widetilde{f}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \frac{1}{n}\sum_{i=1}^{n} \bigl(f(\boldsymbol{x}_i) - \widetilde{f}(\boldsymbol{x}_i)\bigr)$$

$$\mathbb{E}\bigl((\hat{\mu}-\mu)^2\bigr) = \frac{1}{n}\int (\widetilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x}.$$

Part I(a): ANOVA

Building on

Fisher

and

Hoeffding

and

Sobol'

ANOVA: starting with potatoes

Fisher & MacKenzie (1923)

Studies in crop variation II: The manurial response of different potato varieties

Hypothetical potato yields

Four varieties, and 3 manure levels

$$
\begin{array}{cc}
\text{Yield (kg)} & \begin{array}{cccc} \quad V_1 & \quad V_2 & \quad V_3 & \quad V_4 \end{array} \\
\begin{array}{c} M_1 \\ M_2 \\ M_3 \end{array} & \begin{bmatrix} 109.0 & 110.9 & 94.2 & 125.9 \\ 104.9 & 113.4 & 110.1 & 138.0 \\ 151.8 & 160.9 & 111.9 & 145.0 \end{bmatrix}
\end{array}
$$

Anova

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad \sum_{i=1}^{3} \alpha_i = 0 \quad \sum_{j=1}^{4} \beta_j = 0$$

- Y_{ij} is yield
- μ is overall average
- α_i is adjustment up/down for manure i
- β_j is adjustment up/down for variety j
- γ_{ij} is interaction (e.g. synergy)

$$\gamma_{i\bullet} := \sum_{j=1}^{4} \gamma_{ij} = 0 \quad \forall i \qquad\qquad \gamma_{\bullet j} := \sum_{i=1}^{3} \gamma_{ij} = 0 \quad \forall j$$

ANOVA for potatoes

$$
\begin{bmatrix} 109.0 & 110.9 & 94.2 & 125.9 \\ 104.9 & 113.4 & 110.1 & 138.0 \\ 151.8 & 160.9 & 111.9 & 145.0 \end{bmatrix} = \begin{bmatrix} 123 & 123 & 123 & 123 \\ 123 & 123 & 123 & 123 \\ 123 & 123 & 123 & 123 \end{bmatrix}
$$

$$
+ \begin{bmatrix} -13.0 & -13.0 & -13.0 & -13.0 \\ -6.4 & -6.4 & -6.4 & -6.4 \\ 19.4 & 19.4 & 19.4 & 19.4 \end{bmatrix} + \begin{bmatrix} -1.1 & 5.4 & -17.6 & 13.3 \\ -1.1 & 5.4 & -17.6 & 13.3 \\ -1.1 & 5.4 & -17.6 & 13.3 \end{bmatrix}
$$

$$
+ \begin{bmatrix} 0.1 & -4.5 & 1.8 & 2.6 \\ -10.6 & -8.6 & 11.1 & 8.1 \\ 10.5 & 13.1 & -12.9 & -10.7 \end{bmatrix}
$$

The decomposition is interpretable

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

Goes back to Hoeffding (1948)

$$
\begin{aligned}
f(\boldsymbol{x}) &= f_{()}() + \sum_{j=1}^{d} f_{(j)}(x_j) + \sum_{j<k} f_{(j,k)}(x_j, x_k) + \cdots + f_{(1,2,\dots,d)}(x_1,\dots,x_d) \\
&= f_{()} + \sum_{r=1}^{d} \sum_{1\leqslant j_1<j_2<\cdots<j_r\leqslant d} f_{(j_1,j_2,\dots,j_d)}(x_{j_1}, x_{j_2}, \dots, x_{j_d})
\end{aligned}
$$

More simply

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

One term for each subset of $\mathcal{D} = \{1,2,\dots,d\}$

generalizes easily to product domains, i.e. independent inputs

Notation

For $u \subseteq \{1, \ldots, d\}$

$$|u| = \mathbf{card}(u)$$
$$-u = u^c = \{1, 2, \ldots, d\} - u$$
$$v \subset u \qquad \text{\textbf{strict} subset i.e. } \subsetneq$$

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

Frankenpoints

For $\boldsymbol{x}, \boldsymbol{y} \in [0, 1]^d$, $\quad \boldsymbol{z} = \boldsymbol{x}_u{:}\boldsymbol{y}_{-u}$ means

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

We glue together part of $\boldsymbol{x}$ and part of $\boldsymbol{y}$ to form $\boldsymbol{z} = \boldsymbol{x}_u{:}\boldsymbol{y}_{-u}$.

Recursive definition

For $u \subseteq \{1,\dots,d\}$, $f_u(\boldsymbol{x})$ only depends on x_j for $j \in u$.

I.e. $f(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u}) = f(\boldsymbol{x}) \quad \forall \boldsymbol{x}, \boldsymbol{z} \in [0,1]^d$

Overall mean $$f_\varnothing(\boldsymbol{x}) = \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

Main effect j $$f_{\{j\}}(\boldsymbol{x}) = \int \bigl(f(\boldsymbol{x}) - f_\varnothing(\boldsymbol{x})\bigr)\,\mathrm{d}\boldsymbol{x}_{-\{j\}}$$

Interaction u $$f_u(\boldsymbol{x}) = \int \Bigl(f(\boldsymbol{x}) - \sum_{v \subset u} f_v(\boldsymbol{x})\Bigr)\,\mathrm{d}\boldsymbol{x}_{-u} = \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}_{-u} - \sum_{v \subset u} f_v(\boldsymbol{x})$$

For example

$$\int_0^1 x\cos(y)\,\mathrm{d}x = \frac{1}{2}\cos(y)$$ a function of y alone

Sobol's decomposition

Let $\phi_0, \phi_1, \phi_2 \ldots$ be a complete orthonormal basis of $L^2[0,1]$ with $\phi_0(x) \equiv 1$.

$$\int_0^1 \phi_r(x)\,\mathrm{d}x = 1_{r=0}, \qquad \int_0^1 \phi_r(x)\phi_s(x)\,\mathrm{d}x = 1_{r=s}, \qquad \phi_{\boldsymbol{r}}(\boldsymbol{x}) \equiv \prod_{j=1}^{r} \phi_{r_j}(x_j)$$

Tensor product basis

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{r}\in\mathbb{N}^d} \beta_{\boldsymbol{r}}\phi_{\boldsymbol{r}}(\boldsymbol{x}), \qquad \beta_{\boldsymbol{r}} = \int f(\boldsymbol{x})\phi_{\boldsymbol{r}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

$$f_u(\boldsymbol{x}) = \sum_{\boldsymbol{r}_u\in(\mathbb{N}-\{0\})^{|u|}} \beta_{\boldsymbol{r}_u:\boldsymbol{0}_{-u}} \prod_{j\in u} \phi_{r_j}(x_j)$$

"Decomposition into summands of different dimension"

Sobol' (1967) used Haar functions.

Thanks to A. Chouldechova for translation.

ANOVA properties

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

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

$$\& \quad \int f_u(\boldsymbol{x}) g_v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$$

Variances

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

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

Variable importance

How important is $\boldsymbol{x}_u$?

Larger σ_u^2 means that $f_u(\boldsymbol{x})$ contributes more.

Squared contributions are additive.

Sobol's importance measures

$$\underline{\tau}_u^2 = \sum_{v \subseteq u} \sigma_v^2 \qquad \text{Sobol's } D_u$$

$$\overline{\tau}_u^2 = \sum_{v \cap u \neq \varnothing} \sigma_v^2 \qquad \text{Sobol's } D_u^{\text{tot}}$$

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

Large $\underline{\tau}_u^2$ means joint effect of $\boldsymbol{x}_u$ is important

Small $\overline{\tau}_u^2$ means joint effect of $\boldsymbol{x}_u$ is **not** important

We can freeze these 'unessential' variables (Sobol')

Normalized versions: $\dfrac{\underline{\tau}_u^2}{\sigma^2}$ and $\dfrac{\overline{\tau}_u^2}{\sigma^2}$

Inversion

From $\underline{\tau}^2$ to σ^2

$$\begin{aligned}\sigma^2_{\{1,2,3\}} &= \underline{\tau}^2_{\{1,2,3\}} \\ &\quad - \underline{\tau}^2_{\{1,2\}} - \underline{\tau}^2_{\{1,3\}} - \underline{\tau}^2_{\{2,3\}} \\ &\quad + \underline{\tau}^2_{\{1\}} + \underline{\tau}^2_{\{2\}} + \underline{\tau}^2_{\{3\}}\end{aligned}$$

Generally

$$\sigma^2_u = \sum_{v \subseteq u} (-1)^{|u-v|} \underline{\tau}^2_v$$

though this might involve lots of cancellation.

More derived importance measures

Superset importance

$$\Upsilon_u^2 = \sum_{v \supseteq u} \sigma_v^2$$

Liu & O (2006)

Small Υ_u^2 means deleting f_u and higher order interactions makes little difference. Relevant to Hooker (2004)'s simplifications of black box functions.

Mean dimension

$$\frac{1}{\sigma^2} \sum_u |u| \sigma_u^2$$

Measures 'dimensionality' of f. Liu & O (2006)

Higher dimensionality makes for harder numerical handling.

Many quadrature problems have mean dimension near 1

Estimation of $\underline{\tau}_u^2$ and $\overline{\tau}_u^2$

Naive approach for $\underline{\tau}_u^2$:

1) Sample $\boldsymbol{x}_i$ and get $y_i = f(\boldsymbol{x}_i)$ for $i = 1, \dots, n$
2) Somehow estimate $f_v(\boldsymbol{x})$ for all necessary v by $\hat{f}_v$
3) Put $\hat{\sigma}_v^2 = \int \hat{f}_u(\boldsymbol{x})^2 \, \mathrm{d}\boldsymbol{x}$, $u \neq \varnothing$
4) Sum: $\underline{\tau}_u^2 = \sum_{v \subseteq u} \hat{\sigma}_v^2$

This is expensive and has many biases.

Sobol' has a much better way.

Better estimates

$$
\begin{aligned}
\iint f(\boldsymbol{x}) f(\boldsymbol{x}_u\!:\!\boldsymbol{z}_{-u})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}
&= \iint \Bigl(\sum_v f_v(\boldsymbol{x})\Bigr)\Bigl(\sum_w f_w(\boldsymbol{x}_u\!:\!\boldsymbol{z}_{-u})\Bigr)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} && \text{(anova)}\\
&= \sum_v \sum_w \iint f_v(\boldsymbol{x}) f_w(\boldsymbol{x}_u\!:\!\boldsymbol{z}_{-u})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} && \text{(linearity)}\\
&= \sum_v \iint f_v(\boldsymbol{x}) f_v(\boldsymbol{x}_u\!:\!\boldsymbol{z}_{-u})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} && \text{(orthogonality)}\\
&= \sum_{v\subseteq u} \iint f_v(\boldsymbol{x}) f_v(\boldsymbol{x}_u\!:\!\boldsymbol{z}_{-u})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z} && \text{(line integrals)}\\
&= f_\varnothing^2 + \sum_{v\subseteq u} \sigma_u^2 \equiv f_\varnothing^2 + \underline{\tau}_u^2.
\end{aligned}
$$

$$
\hat{\underline{\tau}}_u^2 = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) f(\boldsymbol{x}_{i,u}\!:\!\boldsymbol{z}_{i,-u}) - \Bigl(\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)\Bigr)^2
$$

Even better

$$\underline{\tau}_u^2 = \iint f(\boldsymbol{x})\bigl(f(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u}) - f(\boldsymbol{z})\bigr)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}$$

$$\hat{\underline{\tau}}_u^2 = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\bigl(f(\boldsymbol{x}_{i,u}{:}\boldsymbol{z}_{i,-u}) - f(\boldsymbol{z}_i)\bigr)$$

This avoids subtracting $\hat{f}_{\varnothing}^2$. It is unbiased: $\mathbb{E}\bigl(\hat{\underline{\tau}}_u^2\bigr) = \underline{\tau}_u^2$

Kucherenko, Feil, Shah, Mauntz (2011)

Improved statistical efficiency

$$\hat{\underline{\tau}}_u^2 = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) f(\boldsymbol{x}_{i,u}{:}\boldsymbol{z}_{i,-u}) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{f(\boldsymbol{x}_i) + f(\boldsymbol{x}_{i,u}{:}\boldsymbol{z}_{i,-u})}{2}\right)^2$$

Janon, Klein, Lagnoux, Nodet & Prieur (2012)

For $\overline{\tau}_u^2$

$$
\begin{aligned}
&\frac{1}{2}\iint \bigl(f(\boldsymbol{x}) - f(\boldsymbol{x}_{-u}{:}\boldsymbol{z}_u)\bigr)^2 \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}\\
&= \frac{1}{2}\Bigl(\sigma^2 + f_\varnothing^2 - 2\bigl(\underline{\tau}_{-u}^2 + f_\varnothing^2\bigr) + \sigma^2 + f_\varnothing^2\Bigr)\\
&= \sigma^2 - \underline{\tau}_{-u}^2\\
&= \overline{\tau}_u^2.
\end{aligned}
$$

Sobol's estimates are like tomography: integrals reveal internal structure.

For mean dimension

$$\sum_{j=1}^{d} \overline{\tau}_j^2 = \sum_{j=1}^{d} \sum_{v \cap \{j\} \neq \varnothing} \sigma_v^2$$

$$= \sum_{v} \sigma_v^2 \sum_{j=1}^{d} 1_{v \cap \{j\} \neq \varnothing}$$

$$= \sum_{v} |v| \sigma_v^2$$

From Liu & O (2006)

Generalizes to $\sum_{v} |v|^k \sigma_v^2$ for $k \geqslant 1$.

For superset importance

$$\Upsilon_u^2 = \frac{1}{2^{|u|}} \iint \Bigl(\sum_{v\subseteq u} (-1)^{|u-v|} f(\boldsymbol{x}_v : \boldsymbol{z}_{-v})\Bigr)^2 \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}$$

Mean of a square of differences $\cdots$ better than differences of means of squares.

Includes $2^{|u|}$ terms.

Reasonable for estimating $\sigma^2_{\{1,2,\dots,d\}}$ when d is not too large.
(2^d terms per integrand versus 2^d integrands.)

From Liu & O (2006)

Generalizes $\overline{\tau}_u^2$ formula from 2 terms to $2^{|u|}$ terms.

ANOVA grand challenge

What if $\boldsymbol{x} \sim w$ but w is not a product measure? **Very hard**. See:

- Stone (1984)

Retains $\int f_u(\boldsymbol{x}) f_v(\boldsymbol{x}) w(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = 0$ for $u \subset v$

- Hooker (1987)

Applies to machine learning functions

- Chastaing, Gamboa & Prieur (2012)

New estimation methods for generalized indices

- Kucherenko, Tarantola & Annoni (2012)

Use Gaussian copula

Sampling designs

- Stein (1987) LHS for dependent data
- Petelet, Iooss, Asserin & Loredo (2010) Linearly constrained LHS

Part I(b): quasi-regression

- Like regression but faster
- Can use for estimation of $f_u(\boldsymbol{x})$

Quasi-regression

Suppose that $f(\boldsymbol{x})$ is inexpensive:

1) f may be the surrogate function

2) f may be a prediction rule in statistical machine learning

Then we may be able to take millions of points $\boldsymbol{x}_i \in [0,1]^d \dots$

and get an interpretable approximation.

Basis functions

For $\boldsymbol{k} \in \{0,1,2,\dots\}^d$, let $\phi_{\boldsymbol{k}}(\boldsymbol{x}) \in L^2 : [0,1]^d$ satisfy

$$\phi_{\mathbf{0}}(\boldsymbol{x}) = 1$$

$$\int \phi_{\boldsymbol{k}}(\boldsymbol{x})\phi_{\boldsymbol{j}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1_{\boldsymbol{j}=\boldsymbol{k}}$$

Examples

Fourier, Wavelets, Walsh, polynomials

Truncated model

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\boldsymbol{K}} \beta_{\boldsymbol{k}}\phi_{\boldsymbol{k}}(\boldsymbol{x}) + \eta(\boldsymbol{x}), \qquad \boldsymbol{x} \sim \mathbf{U}[0,1]^d$$

$$\equiv \Phi(\boldsymbol{x})^{\mathsf{T}}\beta + \eta(\boldsymbol{x})$$

$$\Phi(\boldsymbol{x}) = \left(\phi_{(0,\dots,0)}, \phi_{(1,0,\dots,0)}, \dots\right)^{\mathsf{T}} \in \mathbb{R}^{|\boldsymbol{K}|}$$

Choosing $\boldsymbol{K}$

Keep $\boldsymbol{k}$ with small $\|\boldsymbol{k}\|_1$ and/or $\|\boldsymbol{k}\|_0$ and/or $\|\boldsymbol{k}\|_\infty$.

Best $\beta \in \mathbb{R}^{|\boldsymbol{K}|}$

$$\beta^* = \mathbb{E}\big(\Phi(\boldsymbol{x})\Phi(\boldsymbol{x})^{\mathsf{T}}\big)^{-1}\mathbb{E}\big(\Phi(\boldsymbol{x})f(\boldsymbol{x})\big)$$

$$= \mathbb{E}\big(\Phi(\boldsymbol{x})f(\boldsymbol{x})\big) \qquad \text{(by orthogonality)}$$

Regression and quasi-regression

$$\hat{\beta} = \left(\sum_{i=1}^{n} \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_i)^{\mathsf{T}}\right)^{-1} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i) f(\boldsymbol{x}_i) \quad \text{(regression)}$$

$$\tilde{\beta} = \left(nI_{|\boldsymbol{K}|}\right)^{-1} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i) f(\boldsymbol{x}_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i) f(\boldsymbol{x}_i) \quad \text{(quasi-regression)}$$

Comparison for $p = |\boldsymbol{K}|$ basis functions

Method	Time	Space
Regression	$O(np^2)$	$O(p^2)$
Quasi-regression	$O(np)$	$O(p)$

Use n obs for regression or $n' = O(np)$ obs for quasi-regression

Quasi-regression can have $p \gg n$

Regression vs. quasi-regression

If $p \gg n$, then (statistically) regression is better than quasi-regression at estimating β given $\boldsymbol{x}_i$, and $y_i = f(\boldsymbol{x}_i)$ for $i = 1, \dots, n$

Computationally, quasi-regression allows much larger bases

Both are faster than kriging when n is large

Black box approximation

1) Select large basis $\Phi_{\boldsymbol{k}}$ for $\boldsymbol{k} \in \boldsymbol{K} \subset \mathbb{Z}^d$
2) Set all $\widetilde{\beta}_{\boldsymbol{k}} = 0$
3) Get stream of $\boldsymbol{x}_i$ for $i \geqslant 1$
4) Form unbiased estimates $\widetilde{\beta}_{\boldsymbol{k}}$ from the stream
5) Shrink: $\widetilde{f}_n(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \gamma_{\boldsymbol{k},n} \widetilde{\beta}_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}),\quad$ for $0 \leqslant \gamma_{\boldsymbol{k},n} \leqslant 1$
6) Estimate error by averaging $(f(\boldsymbol{x}_n) - \widetilde{f}_{n-1}(\boldsymbol{x}_n))^2$

The $(\boldsymbol{x}_i, y_i)$ flow through the algorithm. No point is looked at twice.

Example

Neural net on $d = 6$ inputs Jiang & O (2001)

ϕ_j Legendre polynomial of degree j

$\phi_{\boldsymbol{k}} = \prod_{j=1}^{d} \phi_{r_j}(x_j)$

$$\boldsymbol{K} = \Bigl\{ \boldsymbol{k} \mid \sum_{j=1}^{d} r_j \leqslant 8,\ \sum_{j=1}^{d} 1_{r_j \neq 0} \leqslant 3,\ \max_{1 \leqslant j \leqslant d} r_j \leqslant 4 \Bigr\}$$

$p = |\boldsymbol{K}| = 1145$, $n = 500{,}000$ $\quad$ 318 seconds in java circa 2000

Results

$$\sigma^2_{\{1\}} \doteq 0.520\,\sigma^2 \qquad \sum_{j=1}^{d} \sigma^2_{\{j\}} \doteq 0.797\,\sigma^2$$

$$\sum_{|u| \leqslant 2} \sigma^2_u \doteq 0.982\,\sigma^2 \qquad \sigma^2_{1,4} \doteq \sigma^2_{3,4} \doteq 0.055\,\sigma^2$$

One also gets plottable estimates of $f_u(\boldsymbol{x})$. $f_{\{1\}}$ was nearly linear.

Estimating linearity

$$f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{d} \beta_j \phi_j(x_j) + \eta(\boldsymbol{x})$$

$$\int_0^1 \phi_j(x)\,\mathrm{d}x = 0 \qquad \int_0^1 \phi_j(x)^2\,\mathrm{d}x = 1 \qquad \int_0^1 \phi_j(x)^4\,\mathrm{d}x < \infty$$

$$R^2 \equiv \frac{\sigma_{\mathrm{L}}^2}{\sigma_{\mathrm{L}}^2 + \int \eta^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}$$

$$\sigma_{\mathrm{L}}^2 \equiv \sum_{j=1}^{d} \beta_j^2$$

$R^2 \approx 1$ makes f easier to integrate, approximately maximize, or visualize

Estimating linearity ctd.

$$\widetilde{\beta}_j = \frac{1}{n}\sum_{i=1}^{n}\phi_j(x_{ij})f(\boldsymbol{x}_i), \qquad \mathbb{E}(\widetilde{\beta}_j) = \beta_j$$

$$\hat{\sigma}^2_{\mathrm{L}} = \sum_{j=1}^{d}\widetilde{\beta}_j^2$$ (severe but correctable bias)

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{L}}) = \frac{n-1}{n}\sigma^2_{\mathrm{L}} + \frac{1}{n}\mathbb{E}\Bigl(f(\boldsymbol{x})^2\sum_j \phi_j(x_j)^2\Bigr)$$

$$\widehat{B}_{\mathrm{L}} = \frac{1}{n^2}\sum_{i=1}^{n} f(\boldsymbol{x}_i)^2\Bigl(\sum_j \phi_j(x_j)^2\Bigr)$$

Bias corrected quasi-regression

$$\hat{\sigma}^2_{\mathrm{L,BC}} = \frac{n}{n-1}\Bigl(\hat{\sigma}^2_{\mathrm{L}} - \widehat{B}_{\mathrm{L}}\Bigr)$$

We get

$$\mathbb{E}\bigl(\hat{\sigma}^2_{\mathrm{L,BC}}\bigr) = \sigma^2_{\mathrm{L}}$$
$$\mathbb{E}\Bigl(\bigl(\hat{\sigma}^2_{\mathrm{L,BC}} - \sigma^2_{\mathrm{L}}\bigr)^2\Bigr) = O\Bigl(\frac{1}{n} + \frac{d^2}{n^3}\Bigr)$$

even when β is **not sparse** if $n \gg d^{2/3}$ O (2000)

also work in progress with Zuk

Workable for $d = 1{,}000{,}000$ and $n = 100{,}000$.

Part I(c): quasi-Monte Carlo

- Monte Carlo
- Latin hypercube sampling
- Quasi-Monte Carlo
- Randomized quasi-Monte Carlo

Sampling to estimate integrals

$$\mu = \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

$$\hat{\mu} = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i), \quad \boldsymbol{x}_i \in [0,1]^d$$

Law of large numbers

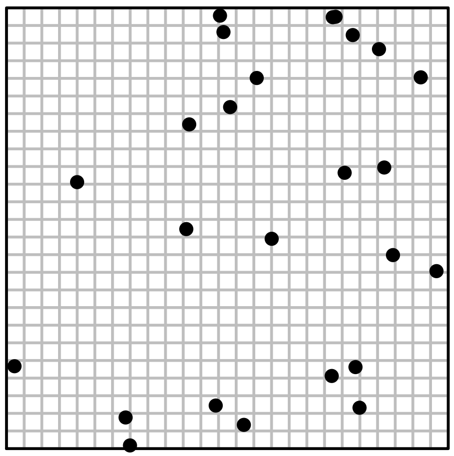
If $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \mathbf{U}[0,1]^d$ and $\int |f(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} < \infty$ then $\Pr\bigl(\hat{\mu}_n \to \mu\bigr) = 1$.

Improvements

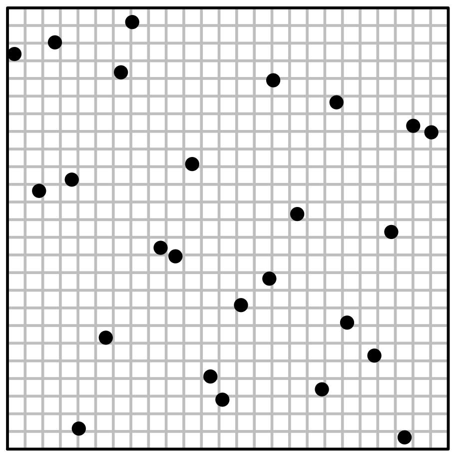
Choose $\boldsymbol{x}_i$ more strategically

MC, LHS, QMC

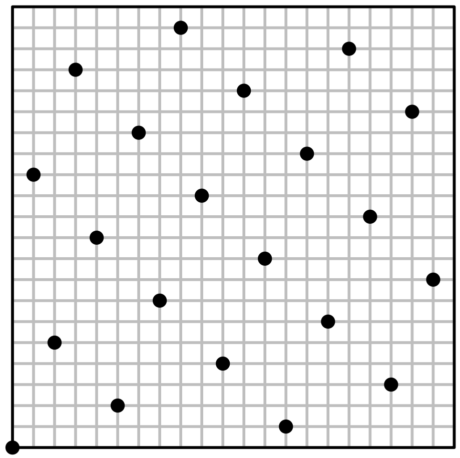
Monte Carlo



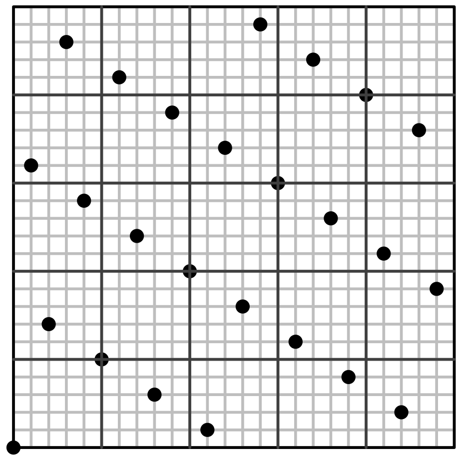
Latin hypercube



QMC (lattice)



QMC (Faure)



QMC goal

QMC makes $\mathbf{U}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\} \approx \mathbf{U}[0,1]^d$. Then

$$\hat{\mu} = \int f(\boldsymbol{x})\,\mathrm{d}\mathbf{U}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$$
$$\mu = \int f(\boldsymbol{x})\,\mathrm{d}\mathbf{U}[0,1]^d$$

will be close.

Koksma-Hlawka

$$|\hat{\mu} - \mu| \leqslant D_n^*(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)\|f\|_{\mathrm{HK}}$$

D_n^*, a discrepancy, that is

$$D_n^* = \|\mathbf{U}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\} - \mathbf{U}[0,1]^d\|$$

$\|\cdot\|_{\mathrm{HK}}$ total variation (Hardy-Krause)

We can make $D_n^* = O((\log n)^d/n)$

QMC rates

$O((\log n)^d/n)$ is really large for large d and feasible n.

n	$(\log n)^{10}/n$	$n^{-1/2}$
10^2	4.3×10^4	10^{-1}
10^4	4.4×10^5	10^{-1}
10^6	2.5×10^5	10^{-2}
10^8	4.5×10^4	10^{-3}
10^{10}	4.2×10^3	10^{-4}
10^{12}	2.6×10^2	10^{-5}

$d = 10$ $\quad \frac{(\log N)^{10}}{N} \approx N^{-1/2}$ some $10^{39} < N < 10^{40}$

$d = 20$ $\quad \frac{(\log N)^{20}}{N} \approx N^{-1/2}$ some $10^{93} < N < 10^{94}$

Taking account of the constants will change the numbers, but not the implication.

Surprise!

A famous 360 dimensional integrand (from finance) was successfully integrated by QMC

Paskov & Traub (1995)

QMC and ANOVA

$$f(\boldsymbol{x}) - \mu = \sum_{|u|>0} f_u(\boldsymbol{x})$$

$$|\hat{\mu} - \mu| \leqslant \sum_{|u|>0} \left| \frac{1}{n} \sum_{i=1}^{n} f_u(\boldsymbol{x}_i) \right|$$

$$\leqslant B \sum_{|u|>0} \frac{(\log n)^{|u|}}{n} \|f_u\|_{\mathrm{HK}}$$

Small error if each large $|u|$ has small $\|f_u\|_{\mathrm{HK}}$.

I.e., if f is dominated by its low order interactions.

no surprise

The 360 dimensional integrand (version in Caflisch, Morokoff & O)
was nearly additive $f(\boldsymbol{x}) \approx f_\varnothing + \sum_j f_{\{j\}}(x_j)$

Effective dimension

A function f has effective dimension s in the **truncation sense** if

$$\sum_{u\subseteq\{1,2,\dots,s\}} \sigma_u^2 \geqslant 0.99\sigma^2$$

A function f has effective dimension s in the **superposition sense** if

$$\sum_{|u|\leqslant s} \sigma_u^2 \geqslant 0.99\sigma^2$$

Superposition is a better description of QMC success than truncation.

Randomized QMC

QMC is deterministic. No practical error estimate.

Under randomized QMC (RQMC)

1) Each $\boldsymbol{x}_i \sim \mathbf{U}[0,1]^d$

2) $D_n^*(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) = O(n^{-1+\epsilon})$ (with prob. 1)

As a result $\mathbb{E}(\hat{\mu}) = \mu$.

Given independent replicates: $\hat{\mu}_1,\dots,\hat{\mu}_R$

$$\hat{\mu} = \frac{1}{R}\sum_{r=1}^{R}\hat{\mu}_r$$

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{R(R-1)}\sum_{r=1}^{R}(\hat{\mu}_r - \hat{\mu})^2.$$

Error cancellation

Random errors cancel (deterministic ones need not). Some RQMC methods attain

$$\mathbb{E}((\hat{\mu}-\mu)^2) = O(n^{-3+\epsilon}), \quad \text{any } \epsilon > 0$$

$$\hat{\mu}-\mu = O_p(n^{-3/2+\epsilon}), \quad \text{any } \epsilon > 0$$

Still requires low effective dimension.

$$\mathrm{Var}(\hat{\mu}) = \sum_{|u|>0} \mathrm{Var}\Bigl(\frac{1}{n}\sum_{i=1}^{n} f_u(\boldsymbol{x}_i)\Bigr).$$

Scrambled nets: O (1997)

Part II

Builds on and extends these topics in new directions

Merci

- GDR Coordinators: Clémentine Prieur, Bertrand Iooss, Fabien Mangeant
- Scientific and organizing committee
- Françoise Poggi
- National Science Foundation of the U.S. DMS-0906056