On Shapley value for measuring importance of dependent inputs

Art B. Owen, Stanford

joint with

Clémentine Prieur, Grenoble

Based on forthcoming JUQ article (2017).

Outline

1) Variable importance
2) ANOVA and Sobol' indices
3) Shapley value
4) Dependent inputs
5) Special cases

You're invited to the

SAMSI workshop on Quasi-Monte Carlo

August 28 to September 1, 2017

Raleigh-Durham, NC

Google with **SAMSI** and **QMC** and **workshop**

Black box functions

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

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

Questions

- How important is x_j?
- How important is the set $\{x_j \mid j \in u\}$ for $u \subseteq \{1, 2, \ldots, d\} \equiv 1{:}d$?

Context

Models in science and engineering:

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

Statistical / machine learning models:

linear models, random forests, $\cdots$

Is importance important?

Sometimes measuring importance is a goal in its own right.

Other times it helps to focus on important variables.

It can however be costly.

Note added after talk

Measuring variable importance can be an expensive preliminary step. At this meeting Barry Nelson mentioned that in some operations research applications one main goal is to find the important inputs and the simulations are much faster/cheaper than the ones underlying engineering applications.

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

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

Notation

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

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

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

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

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

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

Sobol' indices

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

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

We also want to include σ_v^2 for $v \subset u$.

Sobol's (1993) importance measures

$$\underline{\tau}_u^2 = \sum_{v \subseteq u} \sigma_v^2 \qquad v \text{ contained in } u$$

$$\overline{\tau}_u^2 = \sum_{v \cap u \neq \varnothing} \sigma_v^2 \qquad v \text{ touches } u\text{, so interactions count}$$

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 $u = \{1, 2\}$

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

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

Identity

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

Variance explained

After some algebra:

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

So we can use $\underline{\tau}_u^2$ as a measure of how much variance $\boldsymbol{x}_u$ explains.

Hybrid points

Take $\boldsymbol{x}_u$ from $\boldsymbol{x}$ and $\boldsymbol{z}_{-u}$ from $\boldsymbol{z}$ to get $\boldsymbol{x}_u{:}\boldsymbol{z}_{-u}$.

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

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

Example

$$\begin{aligned} \boldsymbol{x} &= (0.1, 0.2, 0.3, 0.4, 0.5, 0.6) \\ &\quad\downarrow\ \downarrow \qquad \downarrow\ \downarrow \\ \boldsymbol{y} &= (0.1, 0.2, 0.7, 0.4, 0.5, 0.4) \equiv \boldsymbol{x}_{\{1,2,4,5\}}{:}\boldsymbol{z}_{\{3,6\}} \\ &\quad\uparrow \qquad\qquad \uparrow \\ \boldsymbol{z} &= (0.9, 0.8, 0.7, 0.6, 0.5, 0.4) \end{aligned}$$

: is a 'smush operator'. We smush together parts of $\boldsymbol{x}$ and $\boldsymbol{z}$.

Pick and freeze methods

Evaluate f at two points:

Freeze: keep some components $\boldsymbol{x}_u \to \boldsymbol{x}_u$

Pick: change the others $\boldsymbol{x}_{-u} \to \boldsymbol{z}_{-u}$

Identities

$$\underline{\tau}_u^2 = \int f(\boldsymbol{x})\big(f(\boldsymbol{x}_u{:}\boldsymbol{z}_{-u}) - f(\boldsymbol{z})\big)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}$$ Mauntz (2002), Saltelli (2001)

$$\overline{\tau}_u^2 = \frac{1}{2}\int \big(\big(f(\boldsymbol{x}) - f(\boldsymbol{x}_{-u}{:}\boldsymbol{z}_u)\big)^2\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{z}$$ Sobol' (1990 / 93)

They follow directly from ANOVA effect orthogonality.

Like tomography: global integrals reveal internal structure.

Estimation

Sample n points in $[0,1]^{2d}$, Monte Carlo or quasi-Monte Carlo

Next topic:

Shapley value

15 million Dollars

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 $
$\varnothing$	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 Dollars

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 $
$\varnothing$	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 $u \subseteq 1{:}d \equiv \{1, 2, \ldots, d\}$ create value **val**(u).

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

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

Shapley axioms

Efficiency	$\sum_{j=1}^{d} \phi_j =$ **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

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

Weighted average of value increments from j

For variable importance

Let variables $x_1, x_2, \ldots, x_d$ 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

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

Linear regression

Lindemann, Merenda and Gold (1980) used it with **val**$(u) = R_u^2$.

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 $j \in u$.

No nice estimation identities like Sobol's. (that we know of)

Requires $2^d - 1$ quantities.

Bracketing

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

By comparison

$$\underline{\tau}^2_{\{j\}} = \sigma^2_{\{j\}}$$
$$\overline{\tau}^2_{\{j\}} = \sum_{u\,:\,j\in u} \sigma_u^2$$

ANOVA for dependent inputs

For general $\boldsymbol{x} \sim p$ usual ANOVA can give $\int f_u(\boldsymbol{x}) f_v(\boldsymbol{x}) p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \neq 0$ for $u \neq v$.

Stone-Hooker

An ANOVA with $\int f_u(\boldsymbol{x}) f_v(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0$ for $u \subsetneq v$

Stone (1994), Hooker (2012)

Chastaing, Gamboa & Prieur (2012,2015) use it to get Sobol' indices.

Problems with Stone-Hooker approach

1) Can get negative importances.

2) Places strong restrictions on p.

Dependence

In a Stone-Hooker ANOVA, we need $p(\boldsymbol{x}) > \epsilon \prod_{j=1}^{d} q_j(x_j)$ some $\epsilon > 0$, densities q_j

Chastaing, Gamboa & Prieur (2012,2015)

Rules out

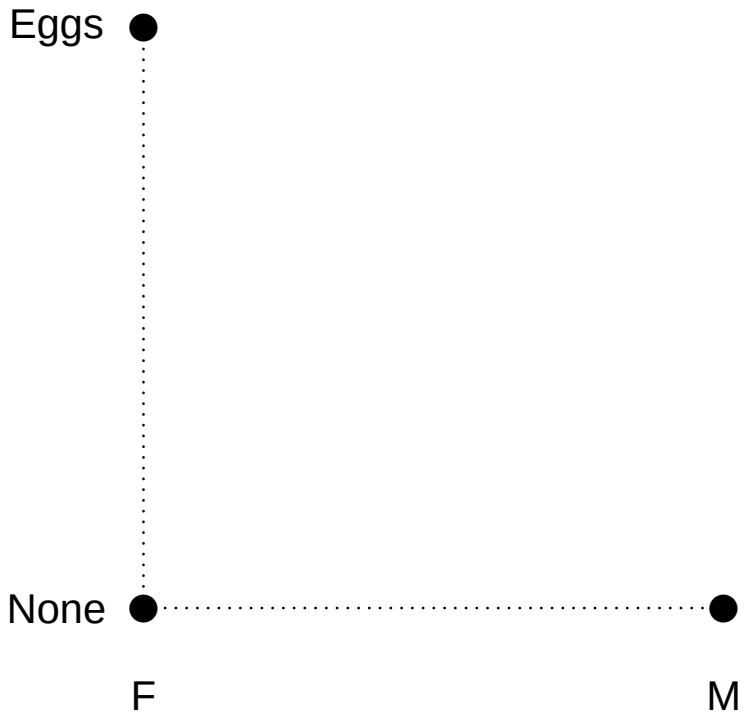
1) $\boldsymbol{x}$ uniform in $\triangle$ (or other non rectangular set)

2) $\boldsymbol{x}$ joint Gaussian $\rho \neq 0$

This condition rules out many (maybe most) dependent data scenarios.

Also, it is hard to keep pick-freeze in bounds.

Extreme case, 3 point support



x_1	x_2	y
0	0	y_{00}
0	1	y_{01}
1	0	y_{10}

3 kinds of turtle: male, female (no eggs), female with eggs

Now: $x_1 = 1_{\text{male}}$, $x_2 = 1_{\text{eggs}}$, $y =$ weight.

If $x_1 = 1$ then we cannot change x_2 at all. And vice versa.

This severely limits pick-freeze.

Shapley

Song, Nelson & Staum (2016)

- Advocate Shapley for dependent inputs.
- Computation is a challenge.
- They present an approach.
- Apply it to some real-world problems.

O & Prieur (2017)

- Verify that Shapley solve ‘conceptual’ issues.
- Give special cases.
- Computational problem remains.

Shapley for dependent $\boldsymbol{x}$

Assume only that $\underline{\tau}_u^2 \equiv \mathrm{Var}(\mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x}_u))$ exists.

Variables $\boldsymbol{x}_{u+j}$ explain at least as much as $\boldsymbol{x}_u$. Therefore $\underline{\tau}_{u+j}^2 - \underline{\tau}_u^2 \geqslant 0$ and so

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} (\underline{\tau}_{u+j}^2 - \underline{\tau}_u^2) \geqslant 0$$

Works even if

1) Support of x_2 depends on x_1
2) $x_1 = x_2$ or x_3 is constant
3) $\boldsymbol{x}$ Gaussian

Linear, independent

$$f(\boldsymbol{x}) = \sum_j \beta_j x_j, \qquad x_j \text{ independent}$$

$$\phi_j = \beta_j^2 \times \mathrm{Var}(x_j)$$

Rescaling

$$x_j \to c x_j, \quad \beta_j \to c^{-1}\beta_j \implies \phi_j \text{ unchanged}$$

Extreme dependent case

$$f(\boldsymbol{x}) = 10^6 x_1 + x_2$$

$$x_1 = 10^6 x_2$$

x_1 has 10^6-fold coefficient

and 10^6-fold range

However, Shapley says

$$\phi_1 = \phi_2$$

because conditioning on x_1 is the same as conditioning on x_2.

Bijections always have equal value

$$x_j = \tau(x_k) \quad \& \quad x_k = \tau^{-1}(x_j)$$

$$\implies \quad \phi_j = \phi_k$$

O & Prieur (2017)

Easily from definitions

Transformation invariance

$$z_j = \tau_j(x_j) \quad \& \quad x_j = \tau_j^{-1}(z_j)$$

$$\begin{aligned}\tilde{f}(z_1, \ldots, z_d) &\equiv f(\tau_1^{-1}(z_1), \ldots, \tau_d^{-1}(z_d)) \\ &= f(x_1, \ldots, x_d)\end{aligned}$$

$$\Longrightarrow \quad \tilde{\phi}_j = \phi_j$$

O & Prieur (2017)

Easily from definitions

Linear Gaussian case

$$\boldsymbol{x} \sim \mathcal{N}(0, \Sigma), \quad f(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\beta = \boldsymbol{x}_u^{\mathsf{T}}\beta_u + \boldsymbol{x}_{-u}^{\mathsf{T}}\beta_{-u}$$

Chastaing, Gamboa & Prieur (2012,2015) find that Stone-Hooker ANOVA does not yield Sobol' indices for this case.

After some algebra

$$\phi_j = \frac{1}{d}\sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \underbrace{\mathrm{Var}(\boldsymbol{x}_{-u}^{\mathsf{T}}\beta_{-u} \mid \boldsymbol{x}_u)}_{\text{unexplained by } \boldsymbol{x}_u} \times \underbrace{\mathrm{Corr}^2(x_j, \boldsymbol{x}_{-u}^{\mathsf{T}}\beta_{-u} \mid \boldsymbol{x}_u)}_{R^2 \text{ from } x_j}$$

Even if $\beta_j = 0$ we can have $\phi_j > 0$:

$$f(\boldsymbol{x}) = x_1 \qquad \boldsymbol{x} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) \quad \Longrightarrow \quad \phi_2 = \frac{\rho^2}{2}$$

Three level data

p	x_1	x_2	y	$\Delta \equiv y - y_{00}$
p_0	0	0	y_{00}	0
p_1	0	1	y_{01}	Δ_1
p_2	1	0	y_{10}	Δ_2

If $\sigma^2 = \mathrm{Var}(y) > 0$ and $\min(p_1, p_2) > 0$, then

$$\frac{\phi_1}{\sigma^2} = \frac{1}{2}\left(1 + \frac{p_0}{\sigma^2} \times \frac{p_1(1-p_1)\Delta_1^2 - p_2(1-p_2)\Delta_2^2}{(1-p_1)(1-p_2)}\right)$$

$p_0 = 0 \implies \phi_1 = 1/2$ (bijection)

$p_1 = p_2 \implies$ larger Δ_j^2 more important

$\Delta_1 = \Delta_2 \implies$ larger p_j more important

O & Prieur (2017)

Bivariate setting

$$\frac{\phi_1}{\sigma^2} = \frac{1}{2}\Big(1 + \frac{\mathrm{Var}(\mathbb{E}(Y\,|\,x_1)) - \mathrm{Var}(\mathbb{E}(Y\,|\,x_2))}{\sigma^2}\Big)$$
$$= \frac{1}{2}\Big(1 + \frac{\mathbb{E}(\mathrm{Var}(Y\,|\,x_2)) - \mathbb{E}(\mathrm{Var}(Y\,|\,x_1))}{\sigma^2}\Big).$$

Special cases

Bivariate Gaussian and $f(\boldsymbol{x}) = \exp(\boldsymbol{x}^\mathsf{T}\beta)$

Bivariate Farlie-Gumbel-Morgenstern copula and $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\beta$

O & Prieur (2017)

Gaussian $\boldsymbol{x}$, exponential f

$$f(\boldsymbol{x}) = e^{\beta_1 x_1 + \beta_2 x_2} \qquad \boldsymbol{x} \sim \mathcal{N}(0, \Sigma) \qquad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Relative importance of x_1

$$\frac{\phi_1}{\sigma^2} = \frac{1}{2}\left(1 + \frac{e^{(\beta_1+\beta_2\rho)^2} - e^{(\beta_2+\beta_1\rho)^2}}{e^{\beta_1^2+\beta_2^2+2\rho\beta_1\beta_2} - 1}\right)$$

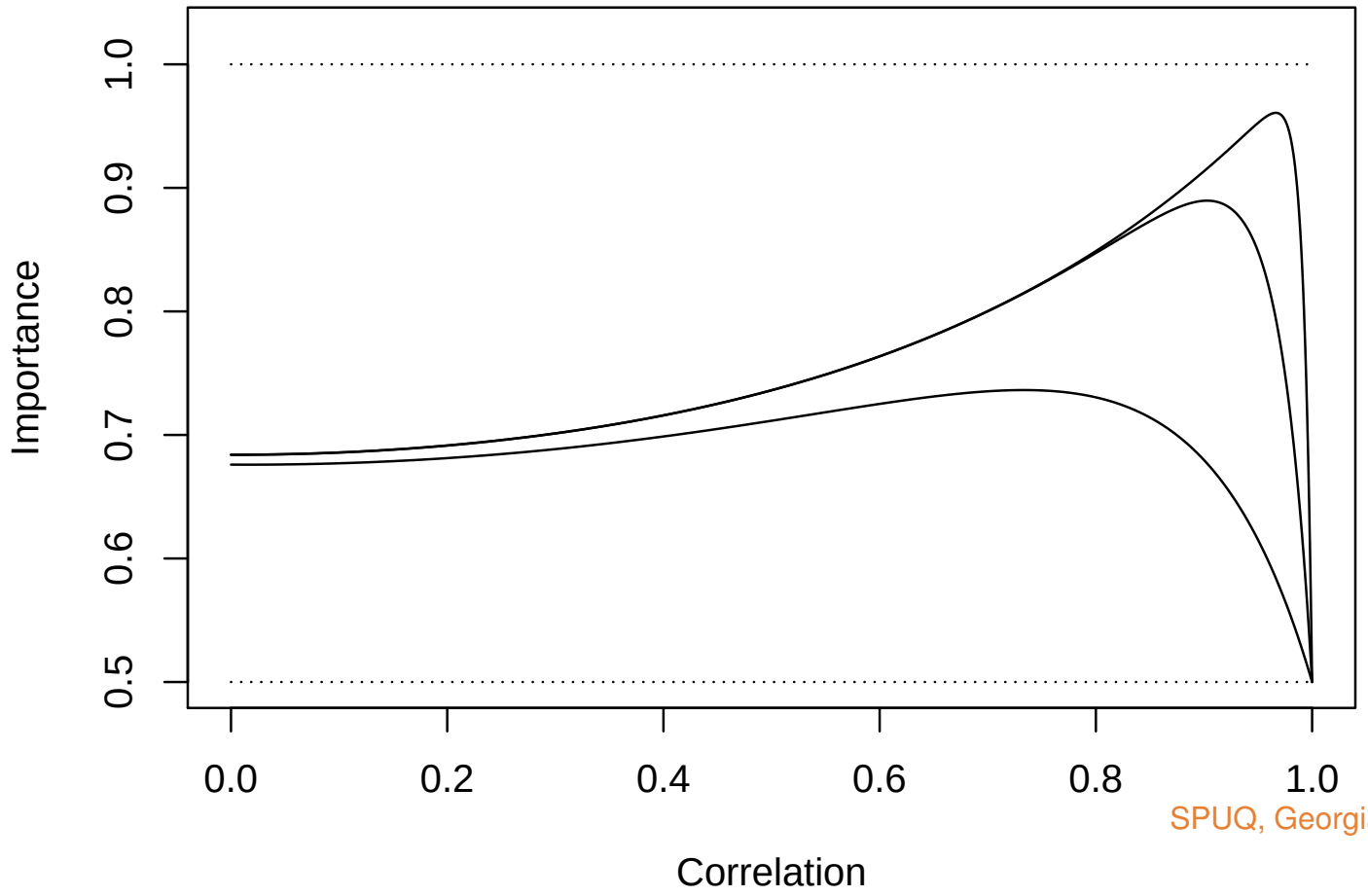
We can easily find

- ϕ_1/σ^2 grows with β_1
- same for ρ and $-\rho$ (transformation invariance)
- equals $\frac{1}{2}$ for $\rho = \pm 1$ (bijection)

But what is the effect of $|\rho|$?

Importance for $e^{\beta^\mathsf{T}\mathcal{N}(0,\Sigma)}$

$\frac{\phi_1}{\sigma^2}$ versus ρ. Top to bottom: $\beta = \begin{pmatrix} 8 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$



Thanks

- Clémentine Prieur, co-author
- Peter Qian, invitation
- V. R. Joseph, D. Higdon, M. Plumlee, P. Qian & B. Haaland, organizers
- H. Sharp, S. Jacobson, local support
- S. Mak, A. Krishna, C.-L. Sung, F. Cao, W. Wang, L.-H. Lin, local support
- NSF DMS-1521145