Stolarsky Invariance and gene set testing

Art B. Owen

Stanford University

Joint with

Hera He, Kinjal Basu, Qingyuan Zhao

Paper to appear

Annals of Statistics

Co-authors

Hera He

PhD (2016)

Gene set tests

$\longrightarrow$ Two Sigma

Kinjal Basu

PhD (2016)

QMC out of cube

$\longrightarrow$ LinkedIn

Qingyuan Zhao

PhD (2016)

Causal inference

$\longrightarrow$ Wharton

Outline

1) We want a permutation test.

2) We don't want to do any permutations.

3) Reverse QMC: our answer is a proportion, the approximation is an integral.

Contributions

1) Test method.

2) New understanding of Stolarsky's invariance principle.

3) Additional connections.

Permutation tests

We have (X_i, Y_i) for $i = 1, \dots, n$ and wonder if X and Y have an association.

Let $\boldsymbol{X} = (X_1, \dots, X_n)^\mathsf{T}$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\mathsf{T}$.

Let π be a permutation of $\{1, 2, \dots, n\}$ and $\boldsymbol{X}_\pi$ have elements $X_{\pi(i)}$.

p-value

$T^{\mathrm{obs}} \equiv T(\boldsymbol{X}, \boldsymbol{Y})$ measures X, Y dependence

- $p = \dfrac{1}{n!} \sum_\pi 1_{T(\boldsymbol{X}_\pi, \boldsymbol{Y}) \geqslant T^{\mathrm{obs}}}$
- This p value is exact (vs independence).
- Based on group theory Lehmann & Romano (2005)
- Power depends on $T(\cdot, \cdot)$.
- Expensive to compute.

Are tests still a thing?

It seems they are, despite criticism and potential misuse.

The p-value measures the sample size, et cetera.

Why isn't everyone a Bayesian by now?

Brad Efron wrote about that.

George Box advocated Bayes for estimation and frequentism for tests.

Permutation issues

It is very expensive to enumerate all $N = n!$ permutations.

For binary X there are only $N = \binom{m_0+m_1}{m_0}$ distinct ones. **Still expensive.**

Parkinson's studies

Data for the paper.

$X_i = 1$ for Parkinsons's, 0 otherwise.

First author	n	m_1	m_0	$N = \binom{m_1+m_0}{m_1}$
Zhang	29	11	18	3.5×10^7
Moran	43	29	14	7.9×10^{10}
Scherzer	72	50	22	1.8×10^{18}

Monte Carlo sampling

$\pi_i \sim$ unif random permutation

$$\hat{p} = \frac{1 + \sum_{i=1}^{M} \mathbf{1}\{T(\boldsymbol{X}_{\pi_i}, \boldsymbol{Y}) \geqslant T^{\mathrm{obs}}\}}{M+1}$$

Adding 1 to numerator keeps $\hat{p} > 0$. 0'th permutation is original data.

Nearly exact

$$\hat{p} \mathrel{\dot{\sim}} \mathbf{U}\left\{\frac{1}{M}, \frac{2}{M}, \frac{3}{M}, \dots, 1\right\} \quad \text{for } M \ll N$$

Barnard (1963)

Granularity

True p **always** $\geqslant 1/N$, where

$$N = n! \quad \text{or} \quad \binom{m_0 + m_1}{m_0}$$

Sample granularity

Monte Carlo $\hat{p} \geqslant 1/M$.

Maybe $p \ll \hat{p}$.

Also

We **really** want $\hat{p} \geqslant 1/N$ and **never** $\hat{p} = 0$.

It is **still** expensive

If we need $\hat{p} \leqslant \epsilon$ to reject H_0.

Then we need $M \geqslant 1/\epsilon$.

Knijnenberg et al.'s (2009) rule of thumb:

$$M \geqslant 10/\epsilon$$

Why small ϵ?

Adjust for making many tests.

Genome wide association studies (GWAS) use $\epsilon = 5 \times 10^{-8}$:

$$M \geqslant 2 \times 10^8$$

Gene association tests

Expression Y_{gi} for gene g subject i.

Relate to $X_i \in \{0,1\}$

E.g., $X_i = 1$ if subject i has Parkinson's disease, 0 else.

An association **might** be causal.

Two cases

- X_i is **disease** status; maybe $Y_g \to X$
- X_i is a **treatment**; maybe $X \to Y_g$

Commonly measure association via correlation.

Single gene tests

Center and scale:

$$X^\mathsf{T}\mathbf{1} = 0,\ X^\mathsf{T}X = 1, \quad Y_g^\mathsf{T}\mathbf{1} = 0,\ Y_g^\mathsf{T}Y_g = 1, \quad g \in \mathcal{G}$$

Correlation

$$\hat\rho_g = X^\mathsf{T}Y_g$$

t statistic

$$t_g = \sqrt{n-2}\frac{\hat\rho_g}{\sqrt{1-\hat\rho_g^2}}$$

Gene set tests

Set of $\mathcal{G}$ of G related genes: $\mathcal{G} = \{g_1, g_2, \dots, g_G\}$

E.g., GO groups, KEGG

Benefits of gene set tests

1) **Interpretation:** the genes have similar function.

2) **Power:** combined evidence of individually small effects.

Plethorae

Ackermann & Strimmer (2009):

Compare $\mathbf{261}$ gene set testing proposals.

Extensive simulations.

They find two clear (families of) winners.

The winners are very simple.

The winners

Linear winners

$$\sum_{g\in\mathcal{G}} t_g \quad \& \quad \sum_{g\in\mathcal{G}} X^\mathsf{T} Y_g$$

Quadratic winners

$$\sum_{g\in\mathcal{G}} t_g^2 \quad \& \quad \sum_{g\in\mathcal{G}} (X^\mathsf{T} Y_g)^2$$

Some other winners tied these.

p-values came from permutations of X wrt Y.

$\sum t_g$ used by Tian et al. (2005)

$\sum t_g$ is the JG score of Jiang & Gentleman (2007).

Why so similar?

The main use case is for many small ρ_g.

Taylor expansion

$$t_g \doteq \sqrt{n-2}\Big(\hat\rho_g + \frac{1}{2}\hat\rho_g^{3}\Big)$$

Permutations for linear statistics

$$\sum_{g\in\mathcal{G}} X^{\mathsf{T}} Y_g = X^{\mathsf{T}} Y \quad \text{for} \quad Y_i \equiv \sum_{g\in\mathcal{G}} Y_{gi}$$

So we will approximate the permutation distribution of a **linear** test statistic.

Quadratic case

Dissertation of Hera He (2016) addresses the quadratic case.

Importance sampling and sequential Monte Carlo

Larson & O (2015) fit gamma distributions.

There is no (known) saddlepoint approximation.

Approximate permutation tests

In addition to plain Monte Carlo tests there are

Saddlepoint approximations

Survey in Reid (1988).

For permutations of linear statistics: Robinson (1982)

Moment and extreme value approximations

Eden & Yates (1933)

Zhou, Wang & Wang (2009)

Larson & O (2015)

Knijnenberg, Wessels, Reinders, Shmulevich (2009)

Not accurate enough, or lacking theory, or numerically problematic.

The sphere

Let $\boldsymbol{X} \in \{0,1\}^n$ with $m_0 > 0$ 0s and $m_1 > 0$ 1s.
Let $\boldsymbol{Y} \in \mathbb{R}^n$ (not constant).

Center and scale

$$\boldsymbol{x} \equiv \frac{\boldsymbol{X} - \bar{X}}{s_X\sqrt{n}} \quad \boldsymbol{y} \equiv \frac{\boldsymbol{Y} - \bar{Y}}{s_Y\sqrt{n}}$$

$$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^d \equiv \{\boldsymbol{z} \in \mathbb{R}^{d+1} \mid \boldsymbol{z}^\mathsf{T}\boldsymbol{z} = 1\} \quad d = n-1$$

Subsphere

$$\begin{aligned}\boldsymbol{x}, \boldsymbol{y} \in &\{\boldsymbol{z} \in \mathbb{R}^{n-1} \mid \boldsymbol{z}^\mathsf{T}\boldsymbol{z} = 1 \ \& \ \boldsymbol{z}^\mathsf{T}\mathbf{1} = 0\} \\ \equiv\ &\mathbb{S}^{n-2}\end{aligned}$$

So $d = n-2$ for the ‘equator’ of $\mathbb{S}^{n-1}$.

Permutation p-value

Original data $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$, $\hat{\rho} = \boldsymbol{x}_0^\mathsf{T}\boldsymbol{y}_0$.

N permutations

$$N = \binom{n}{m_0} = \binom{n}{m_1} = \binom{m_0 + m_1}{m_1}$$

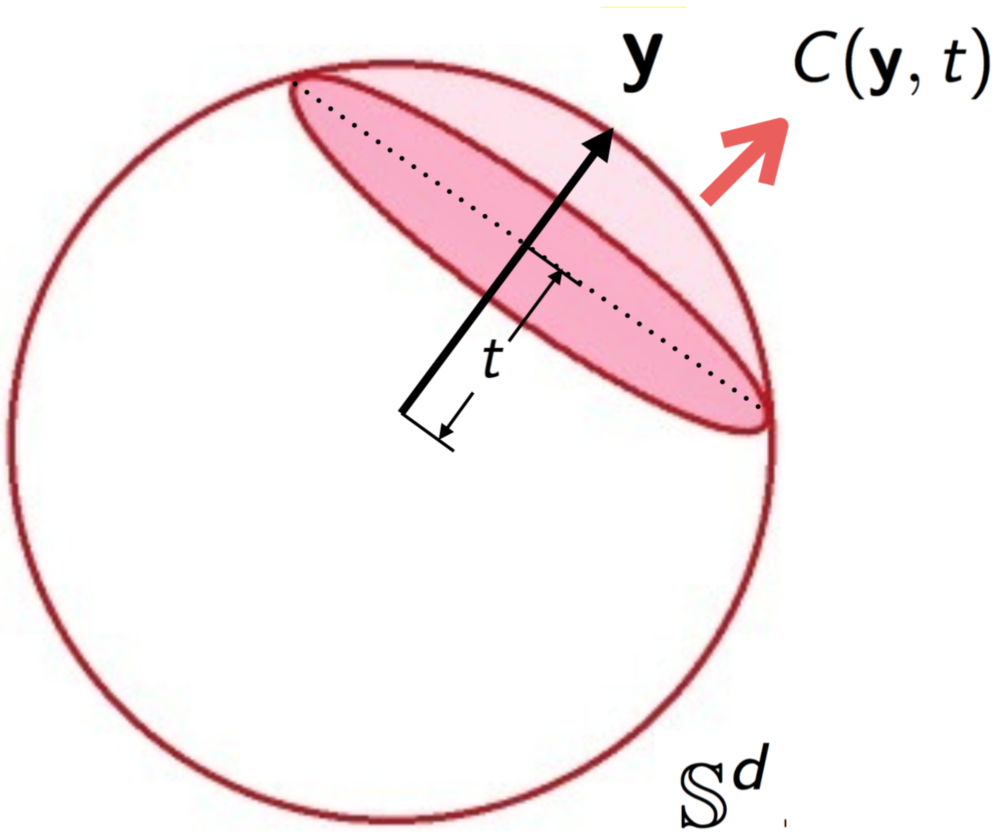
$$p = \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}\{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{y}_0 \geqslant \hat{\rho}\}.$$

For permutations $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{N-1}$ of $\boldsymbol{x}_0$.

Spherical caps

Cap with center $\boldsymbol{y} \in \mathbb{S}^d$ and “height” t, $-1 \leqslant t \leqslant 1$:

$$C(\boldsymbol{y}; t) = \{\boldsymbol{z} \in \mathbb{S}^d \mid \boldsymbol{y}^\mathsf{T}\boldsymbol{z} \geqslant t\}$$



Source: http://hubpages.com/education/Volume-of-a-Spherical-Cap-Formula-and-Examples

Geometry of p-values

$$\hat{p} = \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}\{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{y}_0 \geqslant \hat{\rho}\} = \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{1}\{\boldsymbol{x}_i \in C(\boldsymbol{y};\hat{\rho})\}$$

First approximation

The p-value is the fraction of permuted points in $C(\boldsymbol{y}_0;\hat{\rho})$.

We could approximate it by

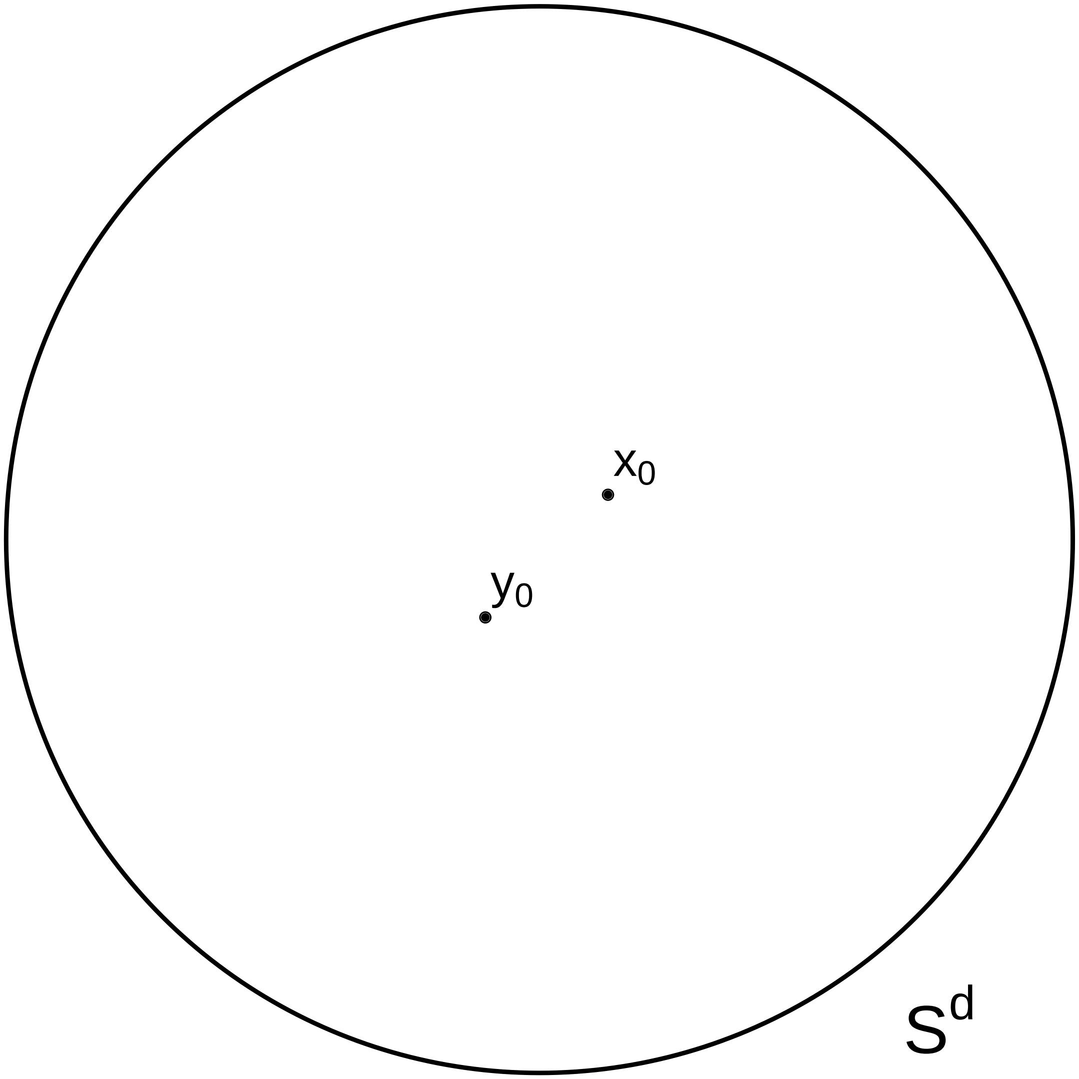
$$\hat{p}_1 = \frac{\mathbf{vol}(C(\boldsymbol{y}_0;\hat{\rho}))}{\mathbf{vol}(\mathbb{S}^d)}$$

This first approximation is not so good. (It is actually the t-test.)

It leads us to better ones:

$\hat{p}_2$ and $\hat{p}_3$ (below).

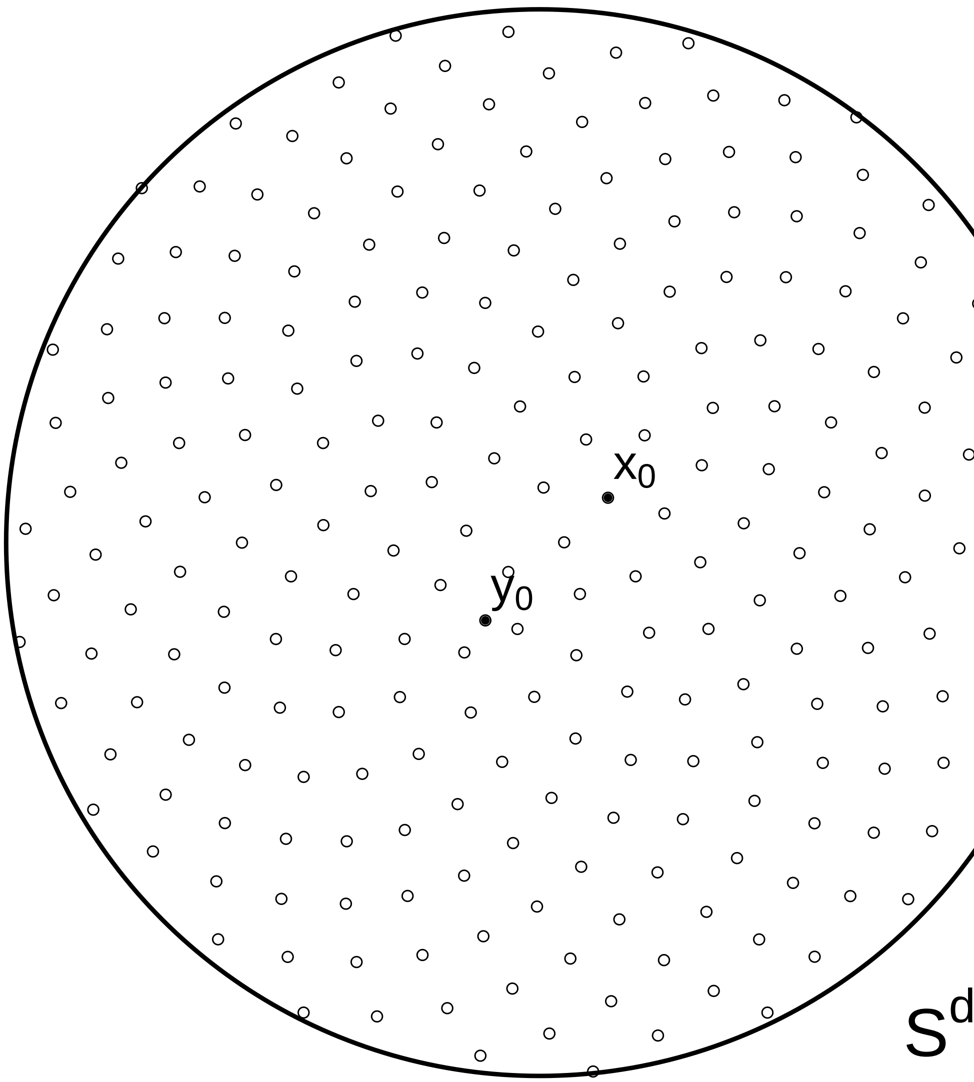
Data in $\mathbb{S}^d$



Binary $\boldsymbol{x}_0$

Real $\boldsymbol{y}_0$

Data in $\mathbb{S}^d$



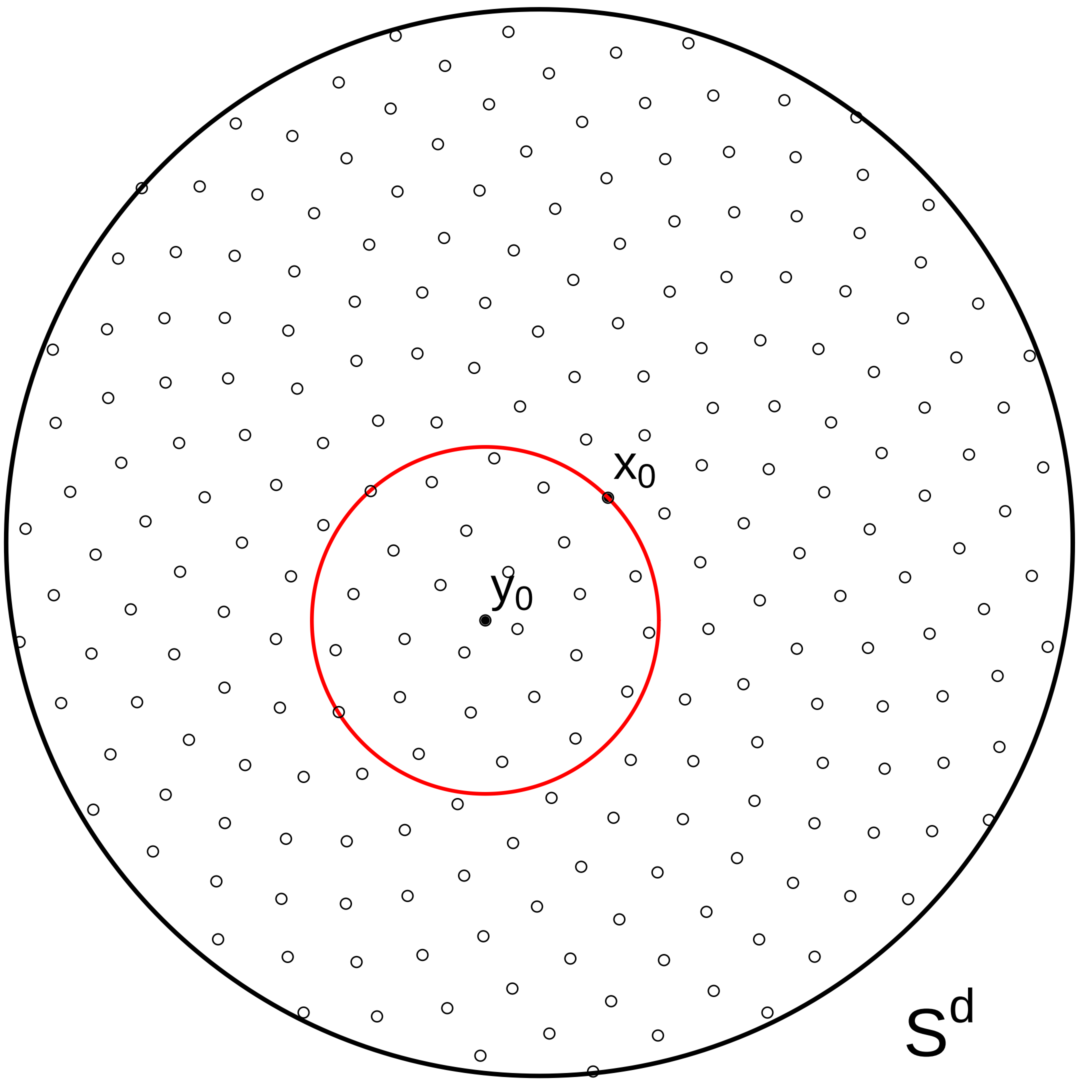
Binary $\boldsymbol{x}_0$

Real $\boldsymbol{y}_0$

Permutations:

$$\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}$$

Data in $\mathbb{S}^d$



Binary $\boldsymbol{x}_0$

Real $\boldsymbol{y}_0$

Permutations:

$$\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}$$

Spherical cap $C(\boldsymbol{y}_0; \hat{\rho})$

$$p = \frac{1}{N} \#\{\boldsymbol{x}_i \text{ in cap}\}$$

$$\hat{p}_1(\boldsymbol{y}; \hat{\rho}) = \frac{\mathbf{vol}(C(\boldsymbol{y}_0; \hat{\rho}))}{\mathbf{vol}(\mathbb{S}^d)}$$

Mean square discrepancy

Is $\hat{p}_1$ close to p?

$$\int_{-1}^{1}\int_{\mathbb{S}^d}(\hat{p}_1(\boldsymbol{y},t)-p(\boldsymbol{y},t))^2\,\mathrm{d}\sigma_d(\boldsymbol{y})\,\mathrm{d}t$$

$$=\int_{-1}^{1}\int_{\mathbb{S}^d}\Bigl(\sigma_d(C(\boldsymbol{y};t))-\frac{1}{N}\sum_{k=0}^{N-1}\mathbf{1}\{\boldsymbol{x}_k\in C(\boldsymbol{y};t)\}\Bigr)^2\,\mathrm{d}\sigma_d(\boldsymbol{y})\,\mathrm{d}t$$

Notes

σ_d is the uniform (Haar) measure on $\mathbb{S}^d$.

This compares $\hat{p}_1$ to p for all centers $\boldsymbol{y}\in\mathbb{S}^d$ all heights t.

We are more interested in accuracy of small caps: $p,\hat{p}\ll 1$.

Stolarsky's invariance principle

For **any** $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{N-1} \in \mathbb{S}^d$

$$\int_{-1}^{1} \int_{\mathbb{S}^d} (\hat{p}_1(\boldsymbol{y}, t) - p(\boldsymbol{y}, t))^2 \, \mathrm{d}\sigma_d(\boldsymbol{y}) \, \mathrm{d}t$$

$$= C_d \Bigl[\int_{\mathbb{S}^d} \int_{\mathbb{S}^d} \|\boldsymbol{x} - \boldsymbol{y}\| \, \mathrm{d}\sigma_d(\boldsymbol{x}) \, \mathrm{d}\sigma_d(\boldsymbol{y}) - \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \|\boldsymbol{x}_k - \boldsymbol{x}_\ell\| \Bigr]$$

$C_d = \omega_{d+1}/(d\omega_d)$ where ω_d is surface measure of $\mathbb{S}^d$

Notes

L^2 left side and L^1 right side.

Like Székely & Rizzo (2013) energy distance

Lowest discrepancy from widest spaced points.

Stolarsky (1973)

Weighted Stolarsky

Brauchart & Dick (2013)

$$\int_{-1}^{1} v(t) \int_{\mathbb{S}^d} \left| \sigma_d(C(\boldsymbol{y};t)) - \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{1}_{C(\boldsymbol{y};t)}(\boldsymbol{x}_k) \right|^2 \mathrm{d}\sigma_d(\boldsymbol{y})\, \mathrm{d}t$$

$$= \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} K_v(\boldsymbol{x}_k, \boldsymbol{x}_\ell) - \int_{\mathbb{S}^d} \int_{\mathbb{S}^d} K_v(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\sigma_d(\boldsymbol{x})\, \mathrm{d}\sigma_d(\boldsymbol{y})$$

Reproducing kernel

$$K_v(\boldsymbol{x}, \boldsymbol{y}) = \int_{-1}^{1} v(t) \int_{\mathbb{S}^d} \mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{x}) \mathbf{1}_{C(\boldsymbol{z};t)}(\boldsymbol{y})\, \mathrm{d}\sigma_d(\boldsymbol{z})\, \mathrm{d}t$$

Setting $v(t) = 1$

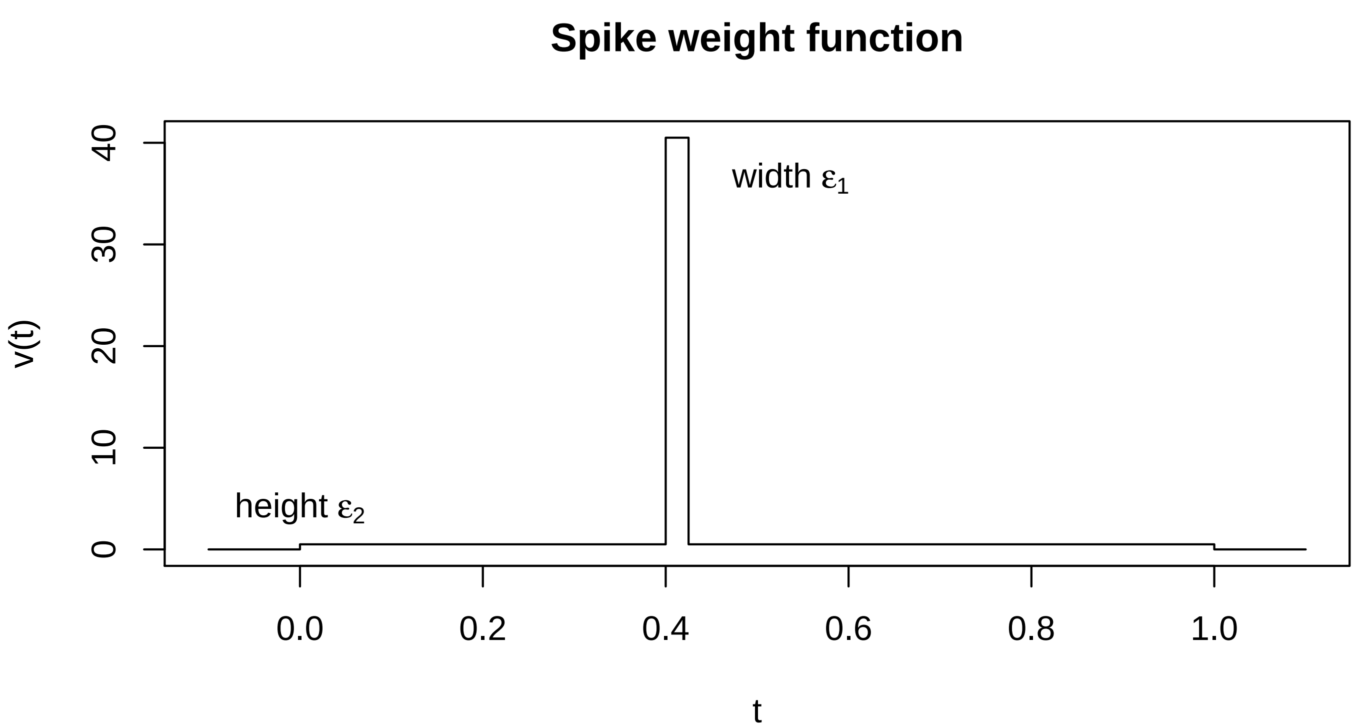
Yields $K_v(\boldsymbol{x}, \boldsymbol{y}) = 1 - C_d \|\boldsymbol{x} - \boldsymbol{y}\|$.

Recovers the original Stolarsky identity.

Size focussed Stolarsky

$$\int_{-1}^{1} v(t) \int_{\mathbb{S}^d} \left| \sigma_d(C(\boldsymbol{y};t)) - \frac{1}{N}\sum_{k=0}^{N-1} \mathbf{1}_{C(\boldsymbol{y};t)}(\boldsymbol{x}_k) \right|^2 \mathrm{d}\sigma_d(\boldsymbol{y})\,\mathrm{d}t$$

Choose $v(t)$ to zoom in on, e.g., $\hat\rho = 0.4$



NB: $\hat\rho = 0.4$ is pretty large. So p would be tiny for reasonable n.

Limiting argument

Let $v_\epsilon(t) = \frac{1}{\epsilon_1}\mathbf{1}\{\hat\rho \leqslant t \leqslant \hat\rho + \epsilon_1\} + \epsilon_2$

Take $\lim_{\epsilon_1 \to 0} \lim_{\epsilon_2 \to 0}$ both sides of weighted Stolarsky identity

What we get

RMSE of $\hat p_1 = \mathbf{vol}(C(\boldsymbol{y};\hat\rho))$ vs p over spherical caps of exactly the desired volume.

The volume $\hat p_1$ of the spherical cap $C(\boldsymbol{y};\hat\rho)$ does not depend on $\boldsymbol{y}$.

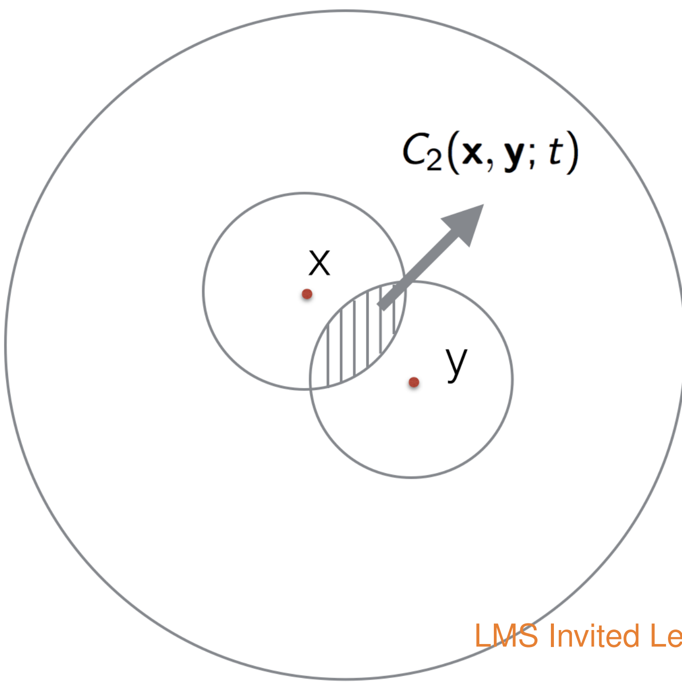
Result

He, Basu, Zhao & O (2016)

$$\int_{\mathbb{S}^d} |\hat{p}_1(\boldsymbol{y},t) - p(\boldsymbol{y},t)|^2 \,\mathrm{d}\sigma_d(\boldsymbol{y}) = \sum_{k=0}^{N-1}\sum_{\ell=0}^{N-1} \sigma_d\big(C_2(\boldsymbol{x}_k,\boldsymbol{x}_\ell;t)\big) - \hat{p}_1(t)^2$$

$$C_2(\boldsymbol{x}_k,\boldsymbol{x}_\ell;t) \equiv C(\boldsymbol{x}_k;t) \cap C(\boldsymbol{x}_\ell;t)$$

Intersection of caps



Probabilistic interpretation

We get a short probabilistic derivation of the Stolarsky invariance principle, using events

$$\boldsymbol{x}_k \in C(\boldsymbol{y};t) \qquad \text{i.e.} \qquad \boldsymbol{y} \in C(\boldsymbol{x}_k;t)$$

and

$$\boldsymbol{x}_k, \boldsymbol{x}_\ell \in C(\boldsymbol{y};t) \qquad \text{i.e.} \qquad \boldsymbol{y} \in C(\boldsymbol{x}_k;t) \cap C(\boldsymbol{x}_\ell;t)$$

See He, Basu, Zhao, O. (2016)

$$p(\boldsymbol{y}_0;\hat{\rho}) = \frac{1}{N}\sum_{k=0}^{N-1} \mathbf{1}\{\boldsymbol{x}_k \in C(\boldsymbol{y}_0;\hat{\rho})\} = \frac{1}{N}\sum_{k=0}^{N-1} Z_k$$

$$Z_k = Z_k(\hat{\rho}) = \mathbf{1}\{\boldsymbol{y}_0 \in C(\boldsymbol{x}_k;\hat{\rho})\}$$

Accuracy

$\hat{p}_1 = \mathbb{E}(p(\boldsymbol{y};\hat{\rho}))$ under $\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$. (Reference distribution 1) He, Basu, Zhao, O (2016)

RMSE = $\mathrm{Var}_{\text{Ref 1}}(\hat{p}_1) \equiv \mathrm{Var}_1(\hat{p}_1)$

Geometry and calculus

σ_d is volume element of unit sphere.

$$\sigma_d(C(\boldsymbol{x}_j;\hat{\rho})) = \text{calculus}$$

Intersection of two caps

$$\begin{aligned}\sigma_d(C_2(\boldsymbol{x}_j,\boldsymbol{x}_k;\hat{\rho})) &\equiv \sigma_d\Big(C(\boldsymbol{x}_j;\hat{\rho}) \cap C(\boldsymbol{x}_k;\hat{\rho})\Big) \\ &= \text{even more calculus}\end{aligned}$$

See He, Basu, Zhao, O (2018).

Incomplete Beta function.

Computation of $\mathrm{Var}_1(p)$

We need N^2 values of $\sigma_d\bigl(C_2(\boldsymbol{x}_k,\boldsymbol{x}_\ell;t)\bigr)$.

It depends on t and $\boldsymbol{x}_k^\mathsf{T}\boldsymbol{x}_\ell$

There are only $\underline{\mathrm{m}} = \min(m_0,m_1)$ distinct $\boldsymbol{x}_k^\mathsf{T}\boldsymbol{x}_\ell$ values.

Recall $\boldsymbol{x}_0$ is binary

$$\boldsymbol{X} = (\underbrace{0,0,\dots,0,}_{m_0}\ \underbrace{1,1,\dots,1}_{m_1})$$

$$\boldsymbol{x}_0 = (\underbrace{\alpha,\alpha,\dots,\alpha,}_{m_0}\ \underbrace{\beta,\beta,\dots,\beta}_{m_1})$$

$$\alpha = -\sqrt{m_1/nm_0},\quad \beta = \sqrt{m_0/nm_1}$$

Swap distance

$$\mathrm{swap}(\boldsymbol{x}_k,\boldsymbol{x}_\ell) = \#\{i \mid \boldsymbol{x}_{ki} > 0 > \boldsymbol{x}_{\ell i}\}$$

Swap distance

If $\mathrm{swap}(\boldsymbol{x}_k, \boldsymbol{x}_\ell) = r$ then

$$\boldsymbol{x}_k^\mathsf{T}\boldsymbol{x}_\ell \equiv u(r) = 1 - r\Big(\frac{1}{m_0} + \frac{1}{m_1}\Big)$$

Now

$$\mathrm{Var}_1(p(\boldsymbol{y};t)) = \frac{1}{N}\sum_{r=0}^{\min(m_0,m_1)} \binom{m_0}{r}\binom{m_1}{r} V_2(u(r);t,d) - \hat{p}_1(t)^2$$

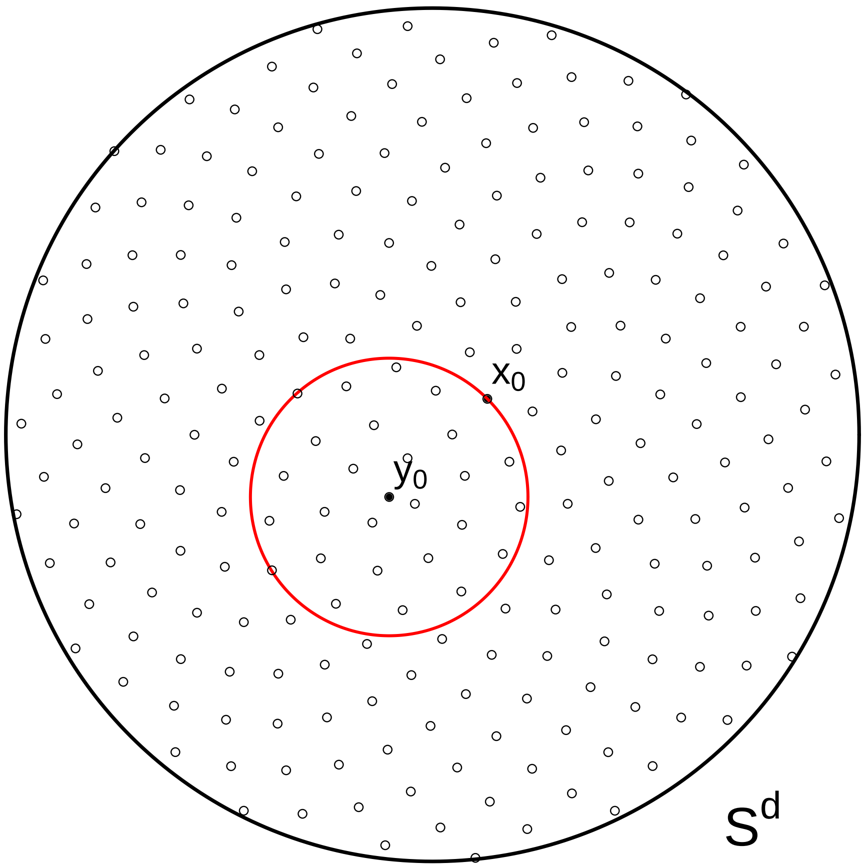
$$V_2(u;t,d) = \sigma_d\big(C_2(\boldsymbol{x},\boldsymbol{y};t)\big) \quad \text{for } \boldsymbol{x}^\mathsf{T}\boldsymbol{y} = u$$

Upshot

$\hat{p}_1 = \mathbb{E}_1(p(\boldsymbol{y};\hat{\rho}))$.

From $\underline{\mathrm{m}} = \min(m_0, m_1)$ integrals we get $\mathrm{Var}_1(p)$.

Reference distribution 1



$\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$

Move red circle over $\mathbb{S}^d$

Get $\hat{p}_1 = \mathbb{E}(\#\text{pts inside})$

and $\mathrm{Var}(\#\text{pts inside})$

What we would prefer

- Letting F approach singleton on $\{\boldsymbol{y}_0\}$
- Replacing RMSE by $\sup$ norm

Finer aproximation

Ref. distn 1 gives accuracy of caps of size exactly $\hat{p}_1$ when $\boldsymbol{y} \sim \mathbf{U}(\mathbb{S}^d)$.

Reference distribution 2

Constrain cap centers too: $\boldsymbol{y}^\mathsf{T}\boldsymbol{x}' = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}'$ for a special point $\boldsymbol{x}'$.

Our favorite $\boldsymbol{x}'$ is $\boldsymbol{x}_0$, then:

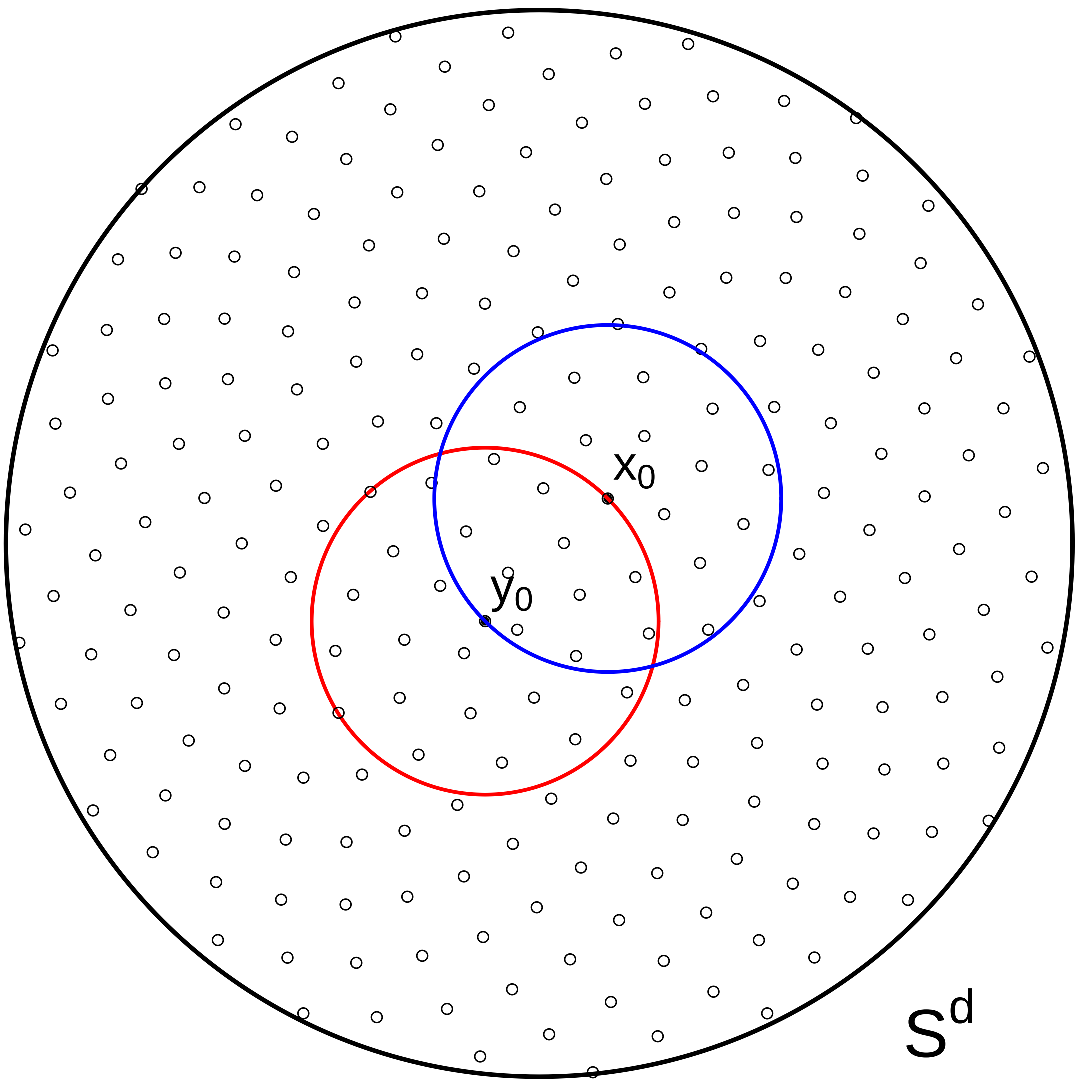
$$\boldsymbol{y} \sim \mathbf{U}\bigl\{\boldsymbol{z} \in \mathbb{S}^d \mid \boldsymbol{z}^\mathsf{T}\boldsymbol{x}_0 = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_0\bigr\}$$

Projection

$$\begin{aligned}\boldsymbol{y} &= \hat{\rho}\boldsymbol{x}_0 + \sqrt{1-\hat{\rho}^2}\boldsymbol{y}^* \\ \boldsymbol{y}^* &\sim \mathbf{U}\bigl\{\boldsymbol{y} \in \mathbb{S}^d \mid \boldsymbol{y}^\mathsf{T}\boldsymbol{x}_0 = 0\bigr\} \equiv \mathbb{S}^{d-1}\end{aligned}$$

Geometrical analysis proceeds via this projection.

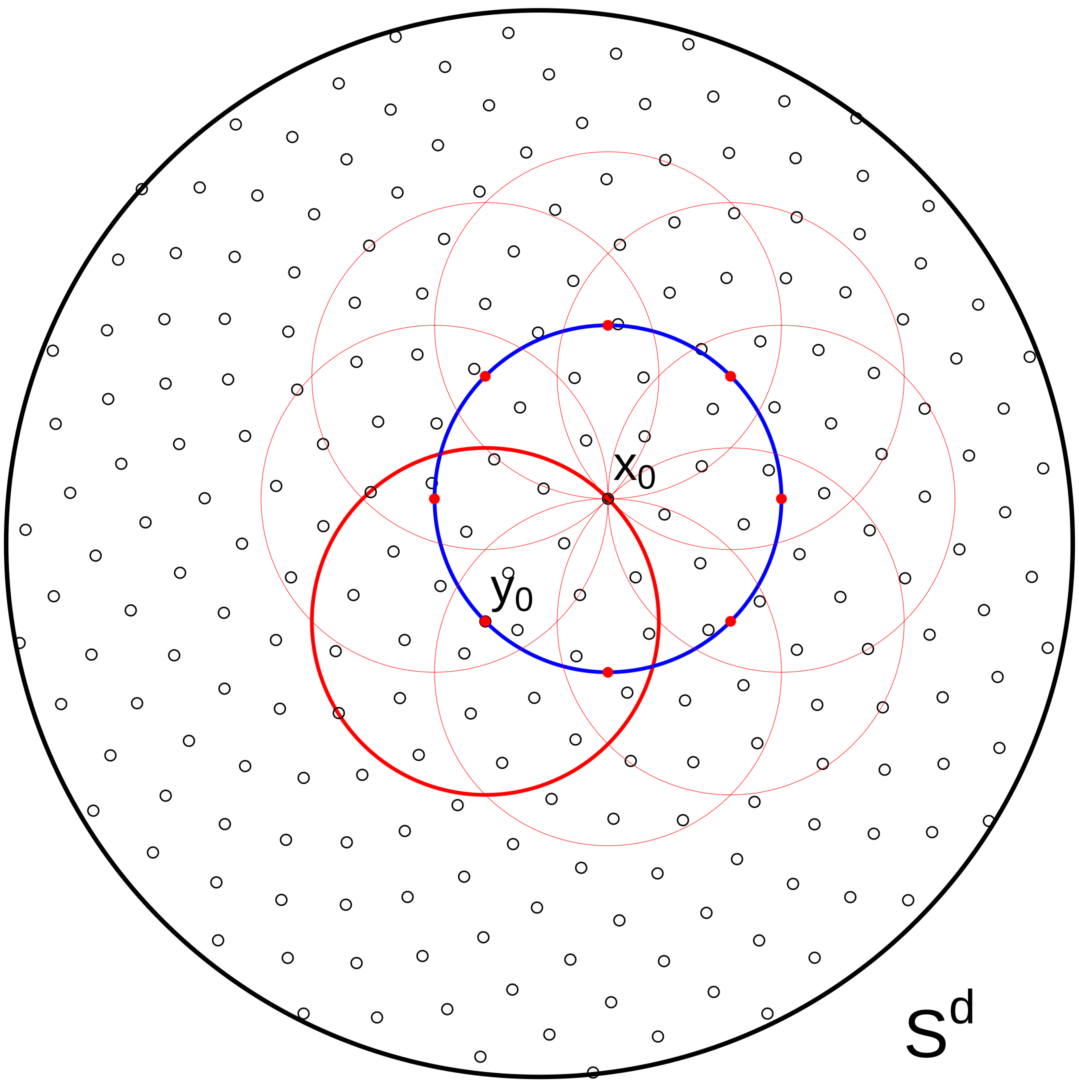
Reference distribution 2



$\boldsymbol{y}$ uniform on blue circle

$$\boldsymbol{y}^{\mathsf{T}}\boldsymbol{x}_0 = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_0$$

Ref distn 2 ctd



$\boldsymbol{y}$ uniform on blue circle

$\boldsymbol{y}^\mathsf{T}\boldsymbol{x}_0 = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_0$

Red caps have same volume as orig.

And are at same distance from $\boldsymbol{x}_0$

We will get $\mathbb{E}_2(p)$ and $\mathrm{Var}_2(p)$

Every red circle contains $\boldsymbol{x}_0$

So $\mathbb{E}_2(p) \geqslant 1/N$ (granularity)

Using ref distn 2

We will find $\hat{p}_2(\boldsymbol{y}_0; t) = \mathbb{E}_2(p(\boldsymbol{y}; t))$

Average true p value for $\boldsymbol{y}$ on the circle

and given cap volume

We really want $\sup\{p(\boldsymbol{y}, t) \mid \boldsymbol{y}^\mathsf{T}\boldsymbol{x}_0 = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_0\}$

We will get $\mathrm{Var}_2(p(\boldsymbol{y}, t))$.

Alternative constraint

Let $c = \arg\max_k \boldsymbol{x}_k^\mathsf{T}\boldsymbol{y}_0 = \arg\min_k \|\boldsymbol{x}_k - \boldsymbol{y}_0\|$

Closest permutation to $\boldsymbol{y}_0$

$\hat{p}_3 = \mathbb{E}(p(\boldsymbol{y}, t) \mid \boldsymbol{y}^\mathsf{T}\boldsymbol{x}_c = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_c)$

Even more conditioning

If we could constrain (condition on) all $\boldsymbol{y}^\mathsf{T}\boldsymbol{x}_i$ we would have the exact p.

If we could constrain $\|\boldsymbol{y} - \boldsymbol{y}_0\| = \epsilon \to 0$ we could approach the exact p.

(But we can't)

Two ways to do it

Find all the single and double point inclusion probabilities under distn 2.

Handle all swap distances $r = 0, 1, \dots, \min(m_0, m_1) \equiv \underline{\mathrm{m}}$ among $\boldsymbol{x}_0, \boldsymbol{x}_k, \boldsymbol{x}_\ell$.

It takes $O(\underline{\mathrm{m}}^3)$ low dimensional integrals.

See He, Basu, Zhao, O (2016)

Stolarsky

Instead of above probabilistic approach, we can also get there via Stolarsky, further generalizing Brauchart & Dick (2013).

Doubly generalized Stolarsky

$$\int_{-1}^{1} v(t) \int_{\mathbb{S}^d} h(\boldsymbol{y}^{\mathsf{T}}\boldsymbol{x}_0) \left|\sigma_d(C(\boldsymbol{y};t)) - p(\boldsymbol{y},t)\right|^2 \mathrm{d}\sigma_d(\boldsymbol{y})\,\mathrm{d}t$$

$=$ fairly long expression

with a new reproducing kernel

Take limits

$h \rightarrow$ point mass at $\boldsymbol{y}^{\mathsf{T}}\boldsymbol{x}_0 = \boldsymbol{y}_0^{\mathsf{T}}\boldsymbol{x}_0$

$v \rightarrow$ point mass at $\hat{\rho}$

Get the same answer as by probability/geometry.

He, Basu, Zhao, O (2016).

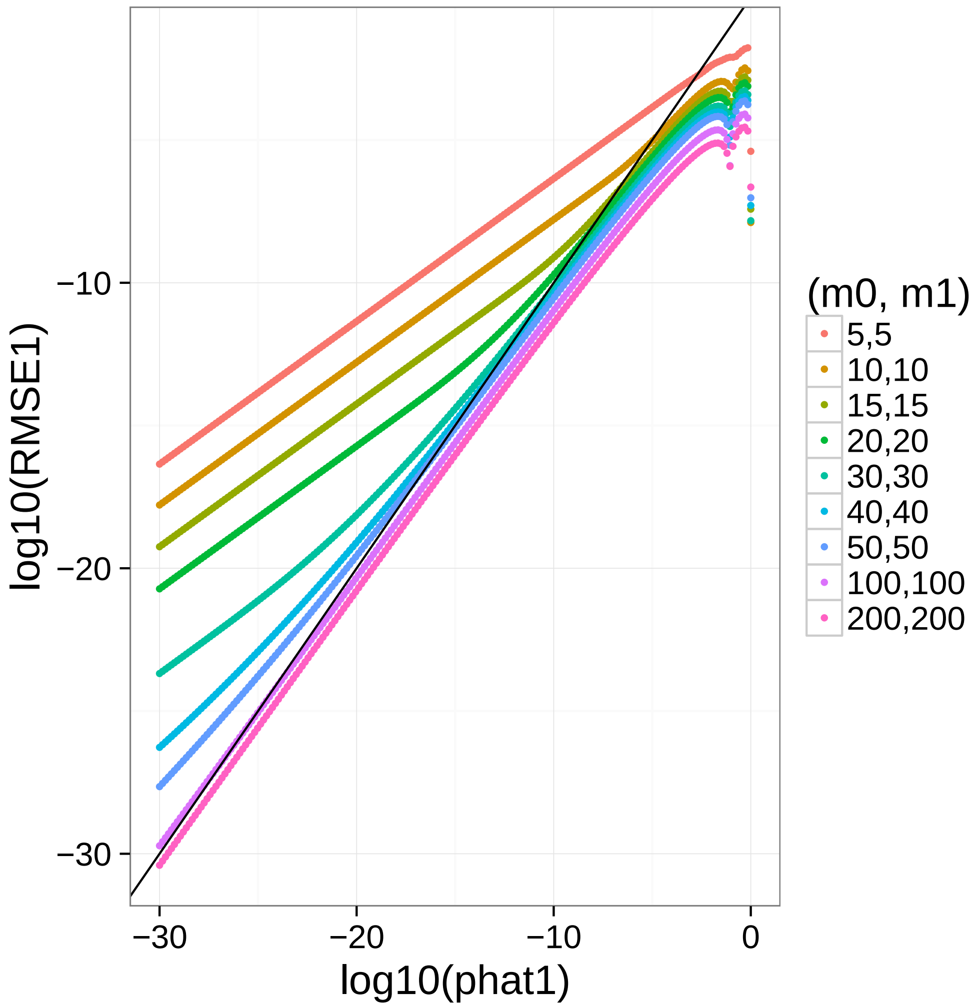
Numerical comparisons

We have estimators

- $\hat{p}_1$: average of p over caps $C(\boldsymbol{y}; \hat{\rho})$
- $\hat{p}_2$: average of p over caps $C(\boldsymbol{y}; \hat{\rho})$ with $\boldsymbol{y}^\mathsf{T}\boldsymbol{x}_0 = \boldsymbol{y}_0^\mathsf{T}\boldsymbol{x}_0$
- $\hat{p}_3$: like $\hat{p}_2$ but using closest $\boldsymbol{x}_k$ to $\boldsymbol{y}_0$

We can also compute $\mathbb{E}((p - \hat{p}_j)^2)$ under ref distns 1 and 2.

RMSE of $\hat{p}_1$ under ref 1



Recall $\hat{p}_1 = \sigma_d(C(\cdot\,;\hat{\rho}))$

$m_0 = m_1$ from 5 to 200

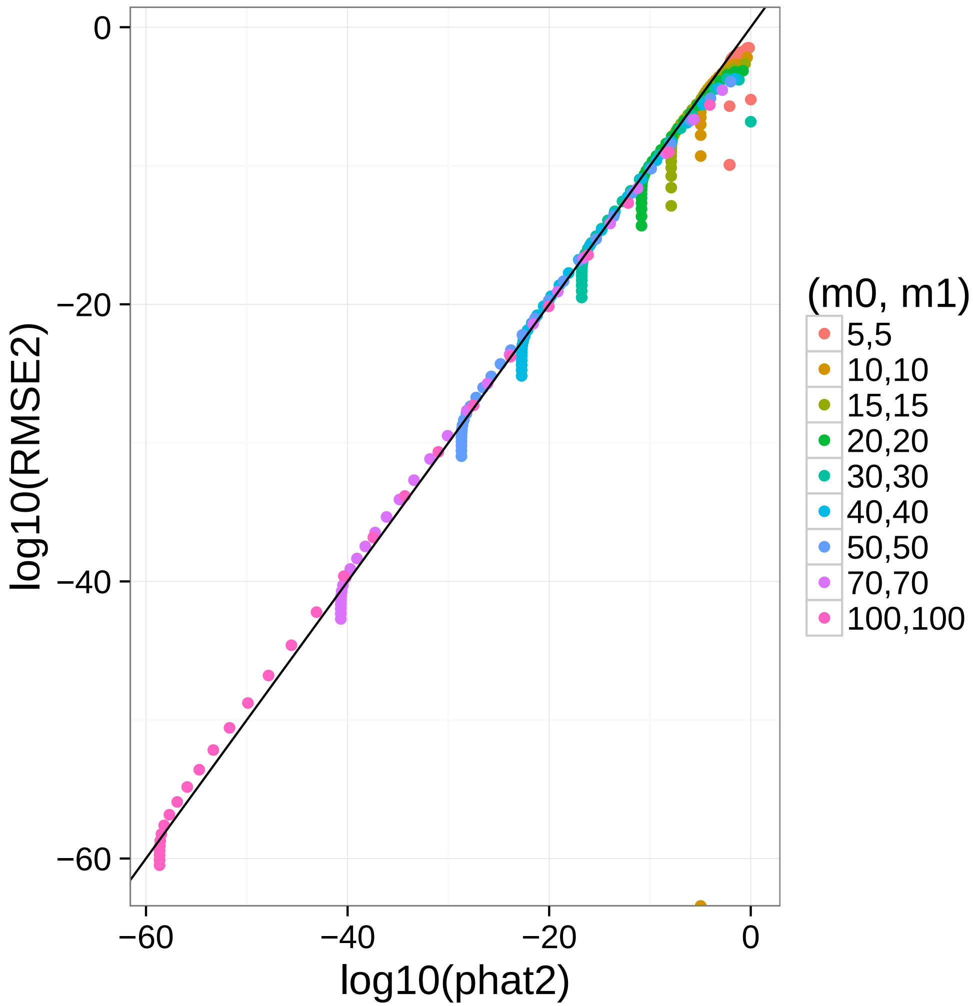
$m_0 \neq m_1$ was similar

$\text{RMSE}_1(\hat{p}_1) \to 0$ as $\hat{p}_1 \to 0$

$\text{RMSE}_1(\hat{p}_1)/\hat{p}_1$ grows

Granularity problem

RMSE of $\hat{p}_2$ under ref 2



Recall $\hat{p}_2 = \mathbb{E}_2(p(\boldsymbol{y}; \hat{\rho}))$

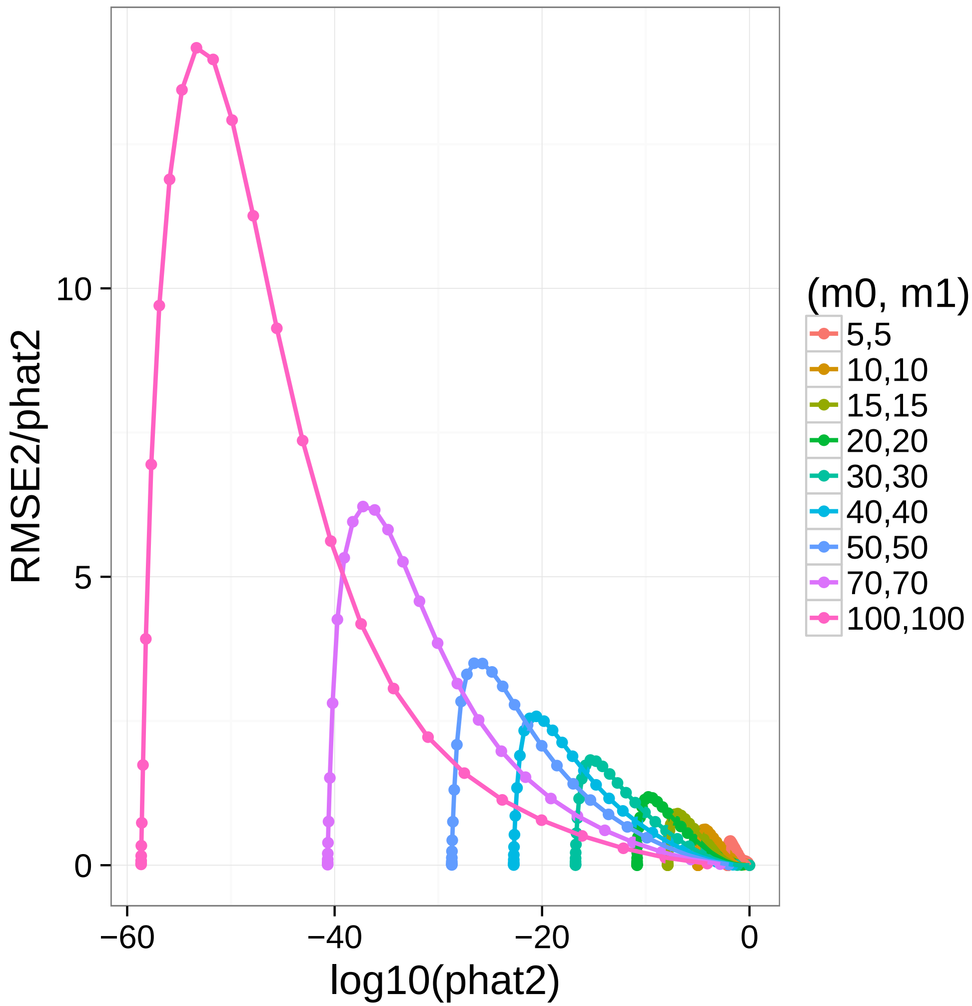
Note 45 degree line

Relative error proportional to mean

$\hat{p}_2 \to 1/N$

RMSE $= 0$ for $1/N \leqslant \hat{p}_2 < 2/N$

Coefficient of variaton of $\hat{p}_2$



Supremum grows with n

But does not get very large

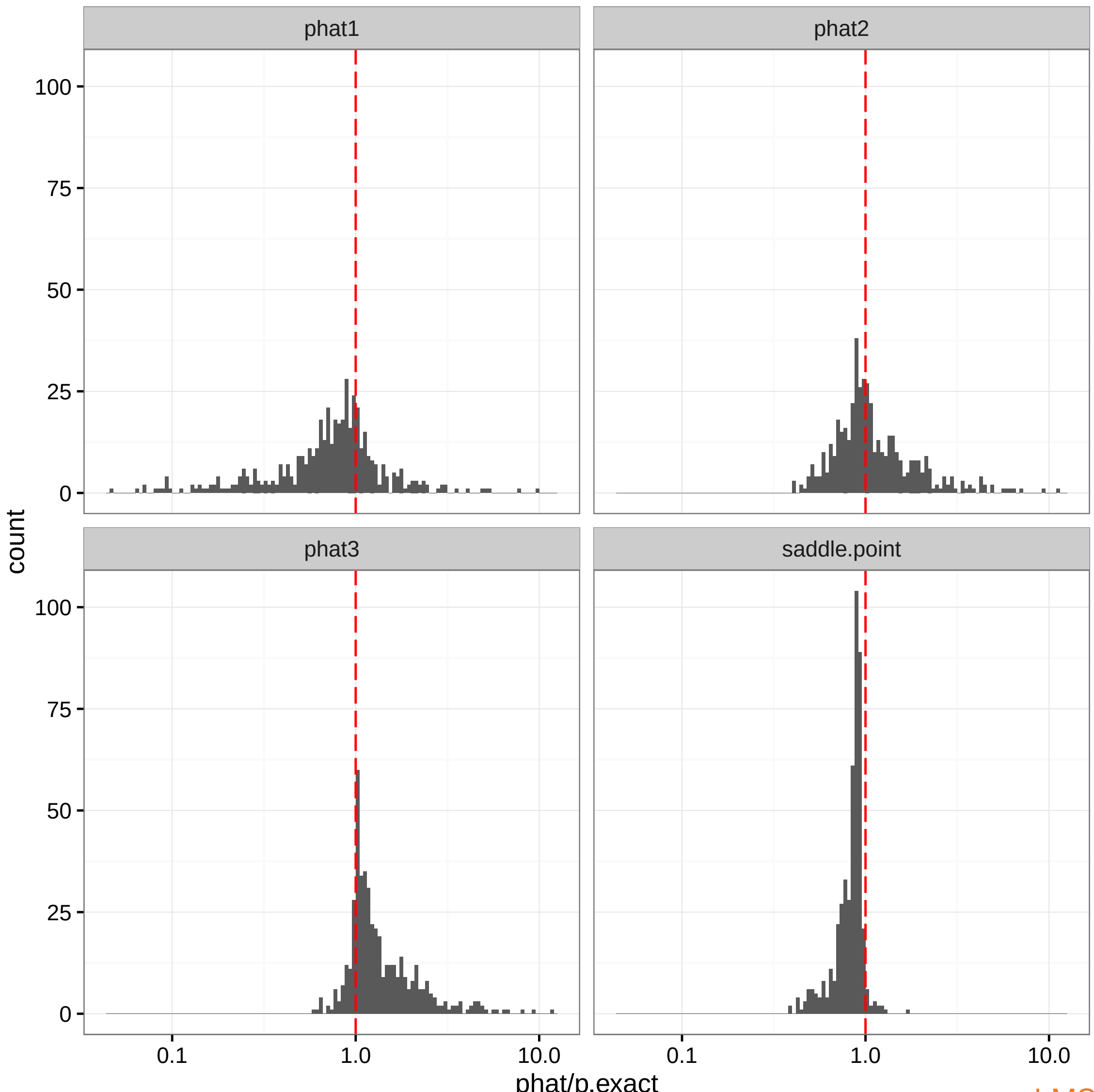
$\hat{p}_2 = 10^{-30}$ RMSE = 5×10^{-30}

For $m_0 = m_1 = 70$

$\hat{p}_2 = 10^{-50}$ RMSE = 10^{-49}

For $m_0 = m_1 = 100$

Comparison



$$Y_0 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$

$$Y_1 \overset{\text{iid}}{\sim} \mathcal{N}(2,1)$$

$$m_0 = m_1 = 10$$

$$\hat{p}/p$$

Findings for normal data

1) $\hat{p}_1$ (t test) not very accurate
2) Saddlepoint accurate, but usually underestimates
3) $\hat{p}_2$ better than $\hat{p}_1$ and underestimates less
4) $\hat{p}_3$ more conservative than $\hat{p}_2$

Similar results for Exponential, $t_{(5)}$ and $\mathbf{U}(0, 1)$

Dissertation of Hera He (2016): $\hat{p}_2$ comes out best on some real gene sets on Parkinson's disease.

Computation time

Data Set	Saddle	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
Zhang	0.0631	0.0024	0.0031	0.0032
Moran	0.0894	0.0029	0.0037	0.0038
Scherzer	0.1394	0.0034	0.0045	0.0047

Averaged over 6180 gene sets.

$\hat{p}$ vs true p

6180 gene sets, $5 \leqslant |G| \leqslant 2131$, avg size 93.08

True p from big Monte Carlo.

No small p values in Zhang data

Data	Condition	Corr.	# sets	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_{\text{saddle}}$
Moran	$p < 0.05$	Pearson	3594	0.9997	0.9997	0.9997	0.9934
Moran	$p < 0.05$	Kendall	3594	0.9857	0.9857	0.9866	0.9397
Moran	$p < 10^{-4}$	Pearson	253	0.9684	0.9688	0.9787	0.7930
Moran	$p < 10^{-4}$	Kendall	253	0.8820	0.8820	0.9033	0.6863
Scherzer	$p < 0.05$	Pearson	504	0.9997	0.9997	0.9997	0.9836
Scherzer	$p < 0.05$	Kendall	504	0.9871	0.9871	0.9871	0.8965
Scherzer	$p < 10^{-3}$	Pearson	16	0.9950	0.9950	0.9956	0.8794
Scherzer	$p < 10^{-3}$	Kendall	16	0.9500	0.9500	0.9500	0.7833

Choices

Method	Strength	Weakness
All permutations	Exact	Too expensive
Monte Carlo	Near exact	Cannot attain small p
Saddlepoint	Relative error	often too small, no error estimate
$\hat{p}_1$	Simple. RMS error	Inaccurate near granularity
$\hat{p}_2$	Relative error, RMS error	No prob. statement
$\hat{p}_3$	Relative error, biased up, RMS error	No prob. statement

Last 4 methods estimate p. But have no all-encompasing probability statement.

Connection and directions

- Quadratic statistics: dissertation of Hera He (2016)
- Ref 1 is similar to rotation tests

  Langsrud (2005), Wu, Lim, Vaillant, Asselin-Labat, Visvader, Smyth (2010)
- There are often covariates
- GWAS needs $p \leqslant 5 \times 10^{-8}$ for single SNPs

  permutations not popular

  approximations may facilitate SNP set analysis

Challenges

1) We really want an L_∞ analog of Stolarsky.
2) Stolarsky matches energy distance on the sphere. Does the connection go deeper?
3) The quadratic case involves 'quadratic caps'; some geometric challenges.

$$\{\boldsymbol{x} \in \mathbb{S}^d \mid \boldsymbol{x}^\mathsf{T} Q \boldsymbol{x} \geqslant q\}$$

Co-authors and other help

- Hera He & Kinjal Basu & Qingyuan Zhao
- John Robinson, comments on saddlepoints
- Neil J. A. Sloane, comments on geometry of numbers
- Jessica Larson and Genentech folks

Thanks

- Lecturers: Nicolas Chopin, Mark Huber, Jeffrey Rosenthal
- Guest speakers: Michael Giles, Gareth Roberts
- The London Mathematical Society: Elizabeth Fisher, Iain Stewart
- CRISM & The University of Warwick, Statistics
- Sponsors: Amazon, Google
- Partners: ISBA, MCQMC, BAYSM
- Poster: Talissa Gasser, Hidamari Design
- NSF: DMS-1407397 & DMS-1521145
- Planners: Murray Pollock, Christian Robert, Gareth Roberts
- Support: Paula Matthews, Murray Pollock, Shahin Tavakoli