Important sampling the union of rare events, with an application to power systems analysis

Art B. Owen

Stanford University

With Yury Maximov and Michael Chertkov

of Los Alamos National Laboratory.

MCQMC 2018

July 1–8, 2018

Rennes, France

MC, QMC, MCMC, RQMC, SMC, MLMC, MIMC, $\cdots$

Call for papers is now online:

`http://mcqmc2018.inria.fr/`

Rare event sampling

Motivation: an electrical grid has N nodes. Power $p_1, p_2, \cdots, p_N$

- Random $p_i > 0$, e.g., wind generation,
- Random $p_i < 0$, e.g. consumption,
- Fixed p_i for controllable nodes,

AC phase angles θ_i

- $(\theta_1, \ldots, \theta_N) = \mathcal{F}(p_1, \ldots, p_N)$
- Constraints: $|\theta_i - \theta_j| < \bar{\theta}$ if $i \sim j$
- Find $\mathbb{P}\Big(\max_{i\sim j} |\theta_i - \theta_j| \geqslant \bar{\theta}\Big)$

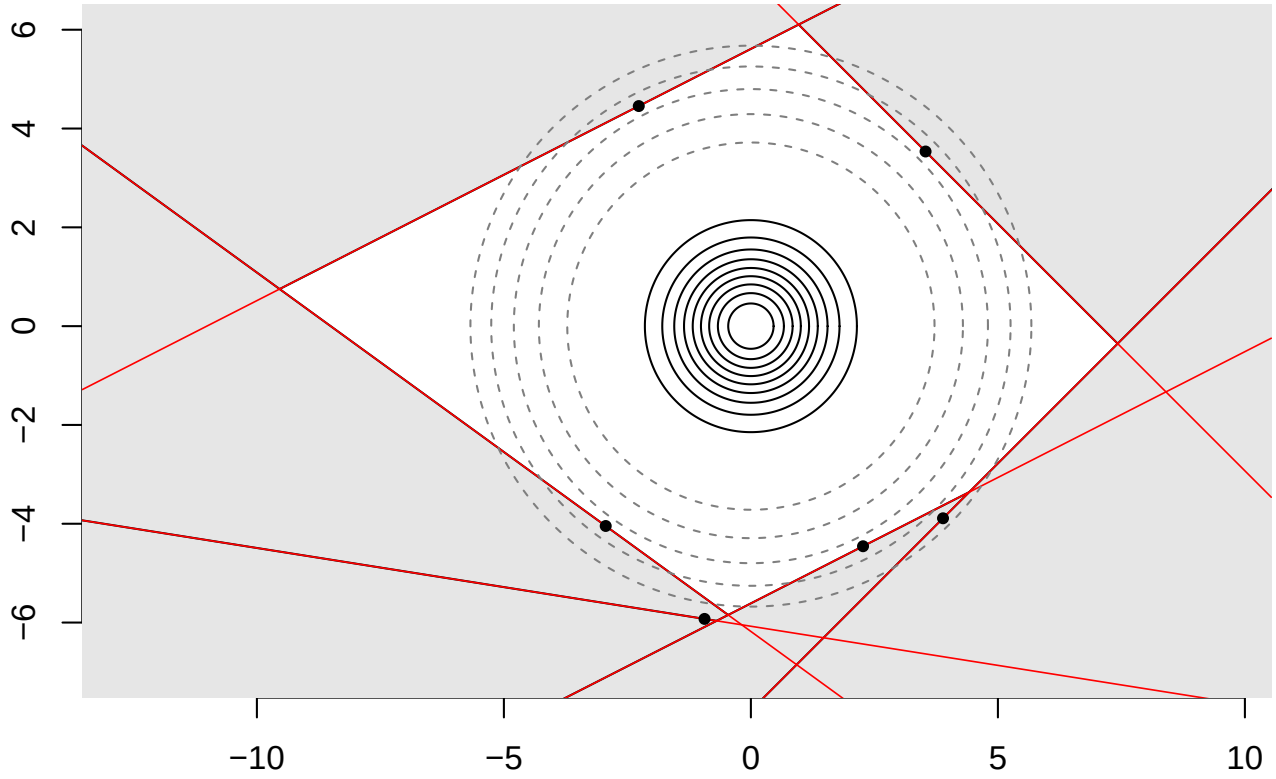
Simplified model

- $(p_1, \ldots, p_N)$ Gaussian
- θ linear in $p_1, \ldots, p_N$

Gaussian setup

For $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$, $H_j = \{\boldsymbol{x} \mid \omega_j^{\mathsf{T}} \boldsymbol{x} \geqslant \tau_j\}$ find $\mu = \mathbb{P}(\boldsymbol{x} \in \cup_{j=1}^{J} H_j)$.

WLOG $\|\omega_j\| = 1$. Ordinarily $\tau_j > 0$. d hundreds. J thousands.



Solid: deciles of $\|\boldsymbol{x}\|$. Dashed: $10^{-3} \cdots 10^{-7}$.

Basic bounds

Let $P_j \equiv \mathbb{P}(\omega_j^{\mathsf{T}} \boldsymbol{x} \geqslant \tau_j) = \Phi(-\tau_j)$

Then $\max_{1 \leqslant j \leqslant J} P_j \equiv \underline{\mu} \leqslant \mu \leqslant \bar{\mu} \equiv \sum_{j=1}^{J} P_j$

Inclusion-exclusion

For $u \subseteq 1{:}J = \{1, 2, \ldots, J\}$, let

$H_u = \cup_{j \in u} H_j \quad H_u(\boldsymbol{x}) \equiv 1\{\boldsymbol{x} \in H_u\} \quad P_u = \mathbb{P}(H_u) = \mathbb{E}(H_u(\boldsymbol{x}))$

$$\mu = \sum_{|u|>0} (-1)^{|u|-1} P_u$$

Plethora of bounds

Survey by Yang, Alajaji & Takahara (2014)

Other uses

- Other engineering reliability
- False discoveries in statistics: J correlated test statistics
- Speculative: inference after model selection

  Taylor, Fithian, Markovic, Tian

Importance sampling

For $\boldsymbol{x} \sim p$, seek $\eta = \mathbb{E}_p(f(\boldsymbol{x})) = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Take

$$\hat{\eta} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q$$

Unbiased if $q(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$.

Variance

$\mathrm{Var}(\hat{\eta}) = \sigma_q^2/n$, where

$$\sigma_q^2 = \int \frac{f^2p^2}{q} - \mu^2 = \cdots = \int \frac{(fp - \mu q)^2}{q}$$

Num: seek $q \approx fp/\mu$, i.e., nearly proportional to fp

Den: watch out for small q.

Self-normalized IS

$$\hat{\eta}_{\mathrm{SNIS}} = \frac{\sum_{i=1}^{n} f(\boldsymbol{x}_i) p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)}{\sum_{i=1}^{n} p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)}$$

Available for unnormalized p and / or q.

Good for Bayes, limited effectiveness for rare events.

Optimal SNIS puts $\mathbf{1/2}$ the samples in the rare event.

Optimal plain IS puts **all** samples there.

Best possible asymptotic coefficient of variation is $2/\sqrt{n}$.

Mixture sampling

For $\alpha_j \geqslant 0, \sum_j \alpha_j = 1$

$$\hat{\eta}_\alpha = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{\sum_j \alpha_j q_j(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q_\alpha \equiv \sum_j \alpha_j q_j$$

Defensive mixtures

Take $q_0 = p$, $\alpha_0 > 0$. Get $p/q_\alpha \leqslant 1/\alpha_0$. Hesterberg (1995)

Additional refs

Use $\int q_j = 1$ as control variates. O & Zhou (2000).

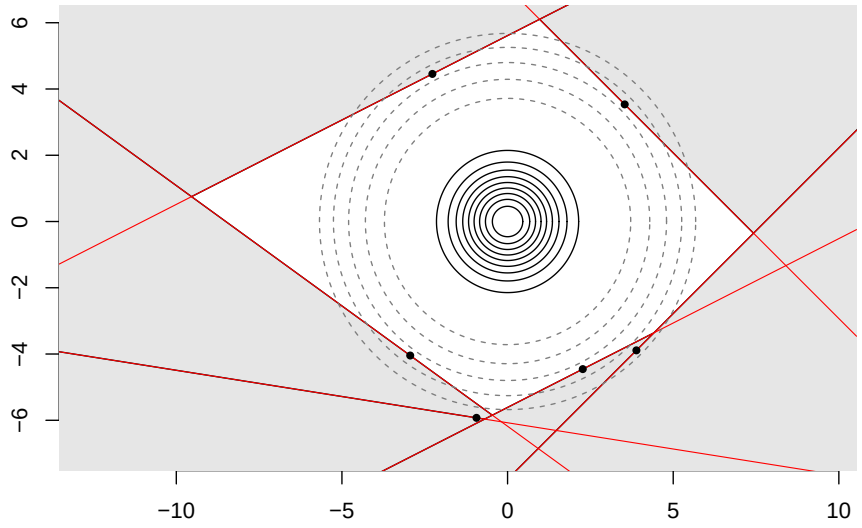
Optimization over α is convex. He & O (2014).

Multiple IS. Veach & Guibas (1994). Veach's Oscar.

Elvira, Martino, Luengo, Bugallo (2015) Generalizations.

Instantons

$\mu_j = \arg\max_{\boldsymbol{x} \in H_j} p(\boldsymbol{x})$ Chertkov, Pan, Stepanov (2011)



- Solid dots μ_j
- Initial thought: $q_j = \mathcal{N}(\mu_j, I)$
- and $q_\alpha = \sum_j \alpha_j q_j$
- I.e., mixture of exponential tilting

Conditional sampling

$$q_j = \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{x} \in H_j) = p(\boldsymbol{x}) H_j(\boldsymbol{x}) / P_j$$

Mixture of conditional sampling

$$\alpha_0, \alpha_1, \ldots, \alpha_J \geqslant 0, \quad \sum_j \alpha_j = 1, \quad q_\alpha = \sum_j \alpha_j q_j, \quad q_0 \equiv p$$

$$\mu = \mathbb{P}(\boldsymbol{x} \in \cup_{j=1}^J H_j) = \mathbb{P}(\boldsymbol{x} \in H_{1:J}) = \mathbb{E}\big(H_{1:J}(\boldsymbol{x})\big)$$

Mixture IS

$$\hat{\mu}_\alpha = \frac{1}{n}\sum_{i=1}^n \frac{H_{1:J}(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{\sum_{j=0}^J \alpha_j q_j(\boldsymbol{x}_i)} \qquad \text{(where } q_j = pH_j/P_j\text{)}$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{H_{1:J}(\boldsymbol{x}_i)}{\sum_{j=0}^J \alpha_j H_j(\boldsymbol{x}_i)/P_j}$$

with $H_0 \equiv 1$ and $P_0 = 1$.

It would be nice to have α_j/P_j constant.

ALORE

At Least One Rare Event

Like AMIS of Cornuet, Marin, Mira, Robert (2012)

Put $\alpha_0^* = 0$, and take $\alpha_j^* \propto P_j$. That is

$$\alpha_j^* = \frac{P_j}{\sum_{j'=1}^J P_{j'}} = \frac{P_j}{\bar{\mu}}, \qquad (\bar{\mu} \text{ is the union bound})$$

Then

$$\hat{\mu}_{\alpha^*} = \frac{1}{n}\sum_{i=1}^n \frac{H_{1:J}(\boldsymbol{x}_i)}{\sum_{j=1}^J \alpha_j H_j(\boldsymbol{x}_i)/P_j} = \bar{\mu} \times \frac{1}{n}\sum_{i=1}^n \frac{H_{1:J}(\boldsymbol{x}_i)}{\sum_{j=1}^J H_j(\boldsymbol{x}_i)}$$

If $\boldsymbol{x} \sim q_{\alpha^*}$ then $H_{1:J}(\boldsymbol{x}) = 1$. So

$$\hat{\mu}_{\alpha^*} = \bar{\mu} \times \frac{1}{n}\sum_{i=1}^n \frac{1}{S(\boldsymbol{x}_i)}, \quad S(\boldsymbol{x}) \equiv \sum_{j=1}^J H_j(\boldsymbol{x})$$

Prior work

Adler, Blanchet, Liu (2008, 2012) estimate

$$w(b) = \mathbb{P}\Big(\max_{t\in T} f(t) > b\Big)$$

for a Gaussian random field $f(t)$ over $T \subset \mathbb{R}^d$.

They also consider a finite set $T = \{t_1, \ldots, t_N\}$.

Comparisons

- We use the same mixture estimate as them for finite T and Gaussian data.
- Their analysis is for Gaussian distributions.

  We consider arbitrary sets of J events.
- They take limits as $b \to \infty$.

  We have non-asymptotic bounds and more general limits.
- Our analysis is limited to finite sets.

  They handle extrema over a continuum.

Theorem

O, Maximov & Cherkov (2017)

Let $H_1, \ldots, H_J$ be events defined by $\boldsymbol{x} \sim p$.
Let $q_j(\boldsymbol{x}) = p(\boldsymbol{x})H_j(\boldsymbol{x})/P_j$ for $P_j = \mathbb{P}(\boldsymbol{x} \in H_j)$.
Let $\boldsymbol{x}_i \overset{\text{iid}}{\sim} q_{\alpha^*} = \sum_{j=1}^{J} \alpha_j^* q_j$ for $\alpha_j^* = P_j/\bar{\mu}$.

Take

$$\hat{\mu} = \bar{\mu} \times \frac{1}{n} \sum_{i=1}^{n} \frac{1}{S(\boldsymbol{x}_i)}, \quad S(\boldsymbol{x}_i) = \sum_{j=1}^{J} H_j(\boldsymbol{x}_i)$$

Then $\mathbb{E}(\hat{\mu}) = \mu$ and

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\left(\bar{\mu} \sum_{s=1}^{J} \frac{T_s}{s} - \mu^2\right) \leqslant \frac{\mu(\bar{\mu} - \mu)}{n}$$

where $T_s \equiv \mathbb{P}(S(\boldsymbol{x}) = s)$.

The RHS follows because

$$\sum_{s=1}^{J} \frac{T_s}{s} \leqslant \sum_{s=1}^{J} T_s = \mu.$$

Remarks

With $S(\boldsymbol{x})$ equal to the number of rare events at $\boldsymbol{x}$,

$$\hat{\mu} = \bar{\mu} \times \frac{1}{n} \sum_{i=1}^{n} \frac{1}{S(\boldsymbol{x}_i)}$$

$$1 \leqslant S(\boldsymbol{x}) \leqslant J \implies \frac{\bar{\mu}}{J} \leqslant \hat{\mu} \leqslant \bar{\mu}$$

Robustness

The usual problem with rare event estimation is getting no rare events in n tries.

Then $\hat{\mu} = 0$.

Here the corresponding failure is never seeing $S \geqslant 2$ rare events.

Then $\hat{\mu} = \bar{\mu}$, an upper bound and probably fairly accurate if all $S(\boldsymbol{x}_i) = 1$

Conditioning vs sampling from $\mathcal{N}(\mu_j, I)$

Avoids wasting samples outside the failure zone.

Avoids awkward likelihood ratios.

General Gaussians

For $\boldsymbol{y} \sim \mathcal{N}(\eta, \Sigma)$ let the event be $\gamma_j^{\mathsf{T}} \boldsymbol{y} \geqslant \kappa_j$.

Translate into $\boldsymbol{x}^{\mathsf{T}} \omega_j \geqslant \tau_j$ for

$$\omega_j = \frac{\gamma_j^{\mathsf{T}} \Sigma^{1/2}}{\gamma_j^{\mathsf{T}} \Sigma \gamma_j} \quad \text{and} \quad \tau_j = \frac{\kappa_j - \gamma_j^{\mathsf{T}} \eta}{\gamma_j^{\mathsf{T}} \Sigma \gamma_j}$$

Adler et al.

Their context has all $\kappa_j = b$

Sampling

Get $\boldsymbol{x} \sim \mathcal{N}(0, I)$, such that $\boldsymbol{x}^{\mathsf{T}}\omega \geqslant \tau$.

y will be $\boldsymbol{x}^{\mathsf{T}}\omega$

1) Sample $u \sim \mathbf{U}(0, 1)$.
2) Let $y = \Phi^{-1}(\Phi(\tau) + u(1 - \Phi(\tau)))$. May easily get $y = \infty$.
3) Sample $\boldsymbol{z} \sim \mathcal{N}(0, I)$.
4) Deliver $\boldsymbol{x} = \omega y + (I - \omega\omega^{\mathsf{T}})\boldsymbol{z}$.

Better numerics from

1) Sample $u \sim \mathbf{U}(0, 1)$.
2) Let $y = \Phi^{-1}(u\Phi(-\tau))$.
3) Sample $\boldsymbol{z} \sim \mathcal{N}(0, I)$.
4) Let $\boldsymbol{x} = \omega y + (I - \omega\omega^{\mathsf{T}})\boldsymbol{z}$.
5) Deliver $\boldsymbol{x} = -\boldsymbol{x}$.

I.E., sample $\boldsymbol{x} \sim \mathcal{N}(0, I)$ subject to $\boldsymbol{x}^{\mathsf{T}}\omega \leqslant -\tau$ and deliver $-\boldsymbol{x}$.

Step 4 via $\omega y + \boldsymbol{z} - \omega(\omega^{\mathsf{T}}\boldsymbol{z})$.

More bounds

Recall that $T_s = \mathbb{P}(S(\boldsymbol{x}) = s)$. Therefore

$$\bar{\mu} = \sum_{j=1}^{J} P_j = \mathbb{E}\left(\sum_{j=1}^{J} H_j(\boldsymbol{x}) \right) = \mathbb{E}(S) = \sum_{s=1}^{J} sT_s$$

$$\mu = \mathbb{E}\Big(\max_{1 \leqslant j \leqslant J} H_j(\boldsymbol{x}) \Big) = \mathbb{P}(S > 0), \quad \text{so}$$

$$\bar{\mu} = \mathbb{E}(S \mid S > 0) \times \mathbb{P}(S > 0) = \mu \times \mathbb{E}(S \mid S > 0)$$

Therefore

$$\begin{aligned} n \times \mathrm{Var}(\hat{\mu}) &= \bar{\mu} \sum_{s=1}^{J} \frac{T_s}{s} - \mu^2 \\ &= \big(\mu \mathbb{E}(S \mid S > 0)\big)\big(\mu \mathbb{E}(S^{-1} \mid S > 0)\big) - \mu^2 \\ &= \mu^2 \big(\mathbb{E}(S \mid S > 0)\mathbb{E}(S^{-1} \mid S > 0) - 1\big) \end{aligned}$$

Bounds continued

$$\mathrm{Var}(\hat{\mu}) \leqslant \frac{1}{n}\left(\mathbb{E}(S \mid S > 0) \times \mathbb{E}(S^{-1} \mid S > 0) - 1\right)$$

Lemma

Let S be a random variable in $\{1, 2, \ldots, J\}$ for $J \in \mathbb{N}$. Then

$$\mathbb{E}(S)\mathbb{E}(S^{-1}) \leqslant \frac{J + J^{-1} + 2}{4}$$

with equality if and only if $S \sim \mathbf{U}\{1, J\}$.

Corollary

$$\mathrm{Var}(\hat{\mu}) \leqslant \frac{\mu^2}{n}\frac{J + J^{-1} - 2}{4}.$$

Thanks to Yanbo Tang and Jeffrey Negrea for an improved proof of the lemma.

Numerical comparison

R function `mvtnorm` of Genz & Bretz (2009) gets

$$\mathbb{P}\big(\boldsymbol{a} \leqslant \boldsymbol{y} \leqslant \boldsymbol{b}\big) = \mathbb{P}\big(\cap_j \{a_j \leqslant y_j \leqslant b_j\}\big) \quad \text{for} \quad \boldsymbol{y} \sim \mathcal{N}(\eta, \Sigma)$$

Their code makes sophisticated use of quasi-Monte Carlo.

Adaptive, up to 25,000 evals in FORTRAN.

It was not designed for rare events.

It computes an intersection.

Usage

Pack ω_j into Ω and τ_j into $\mathcal{T}$

$$1 - \mu = \mathbb{P}(\Omega^{\mathsf{T}} \boldsymbol{x} \leqslant \mathcal{T}) = \mathbb{P}(\boldsymbol{y} \leqslant \mathcal{T}), \quad \boldsymbol{y} \sim \mathcal{N}(0, \Omega^{\mathsf{T}}\Omega)$$

It can handle up to 1000 inputs. IE $J \leqslant 1000$

Upshot

ALORE works (much) better for rare events.

`mvtnorm` works better for non-rare events.

Polygon example

$\mathcal{P}(J, \tau)$ regular J sided polygon in $\mathbb{R}^2$ outside circle of radius $\tau > 0$.

$$\mathcal{P} = \left\{ \boldsymbol{x} \;\middle|\; \begin{pmatrix} \sin(2\pi j/J) \\ \cos(2\pi j/J) \end{pmatrix}^{\mathsf{T}} \boldsymbol{x} \leqslant \tau, \quad j = 1, \ldots, J \right\}$$

a priori bounds

$$\mu = \mathbb{P}(\boldsymbol{x} \in \mathcal{P}^c) \leqslant \mathbb{P}(\chi^2_{(2)} \geqslant \tau^2) = \exp(-\tau^2/2)$$

This is pretty tight. A trigonometric argument gives

$$1 \geqslant \frac{\mu}{\exp(-\tau^2/2)} \geqslant 1 - \frac{\left(J \tan\left(\frac{\pi}{J}\right) - \pi\right)\tau^2}{2\pi} \doteq 1 - \frac{\pi^2 \tau^2}{6J^2}$$

Let's use $J = 360$

So $\mu \doteq \exp(-\tau^2/2)$ for reasonable τ.

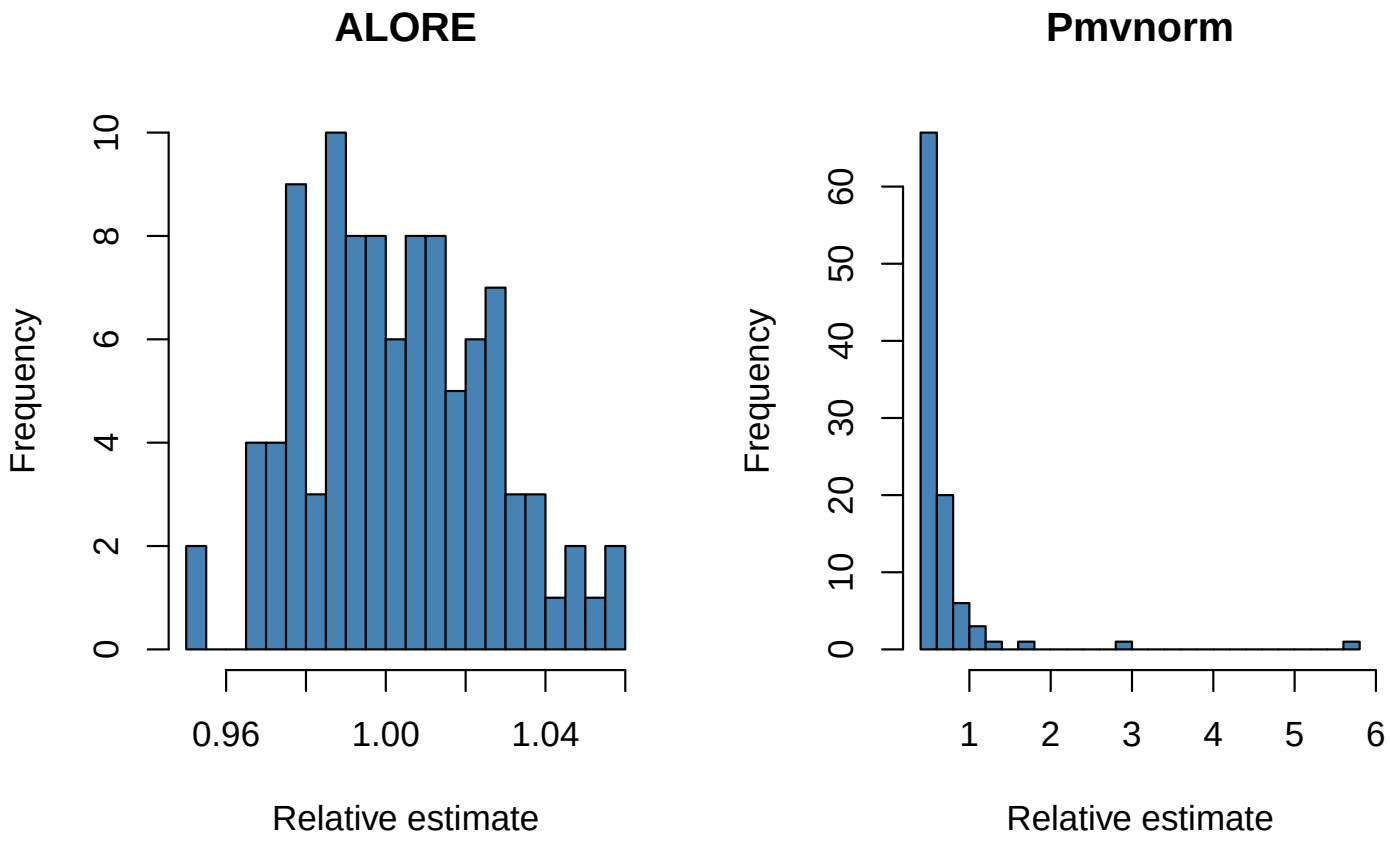
IS vs MVN

ALORE had $n = 1000$. MVN had up to $25{,}000$. 100 random repeats.

τ	μ	$\mathbb{E}((\hat{\mu}_{\mathrm{ALORE}}/\mu - 1)^2)$	$\mathbb{E}((\hat{\mu}_{\mathrm{MVN}}/\mu - 1)^2)$
2	1.35×10^{-01}	0.000399	9.42×10^{-08}
3	1.11×10^{-02}	0.000451	9.24×10^{-07}
4	3.35×10^{-04}	0.000549	2.37×10^{-02}
5	3.73×10^{-06}	0.000600	$1.81\times10^{+00}$
6	1.52×10^{-08}	0.000543	4.39×10^{-01}
7	2.29×10^{-11}	0.000559	3.62×10^{-01}
8	1.27×10^{-14}	0.000540	1.34×10^{-01}

For $\tau = 5$ MVN had a few outliers.

Polygon again



Results of 100 estimates of the $\mathbb{P}(\boldsymbol{x} \notin \mathcal{P}(360, 6))$, divided by $\exp(-6^2/2)$.

Left panel: ALORE. Right panel: pmvnorm.

Symmetry

Maybe the circumscribed polygon is too easy due to symmetry.

Redo it with just $\omega_j = (\cos(2\pi j/J), \sin(2\pi j/J))^{\mathsf{T}}$ for the 72 prime numbers $j \in \{1, 2, \ldots, 360\}$.

For $\tau = 6$ variance of $\hat{\mu}/\exp(-18)$ is 0.00077 for ALORE ($n = 1000$), and 8.5 for `pmvnorm`.

High dimensional half spaces

Take random $\omega_j \in \mathbb{R}^d$ for large d.

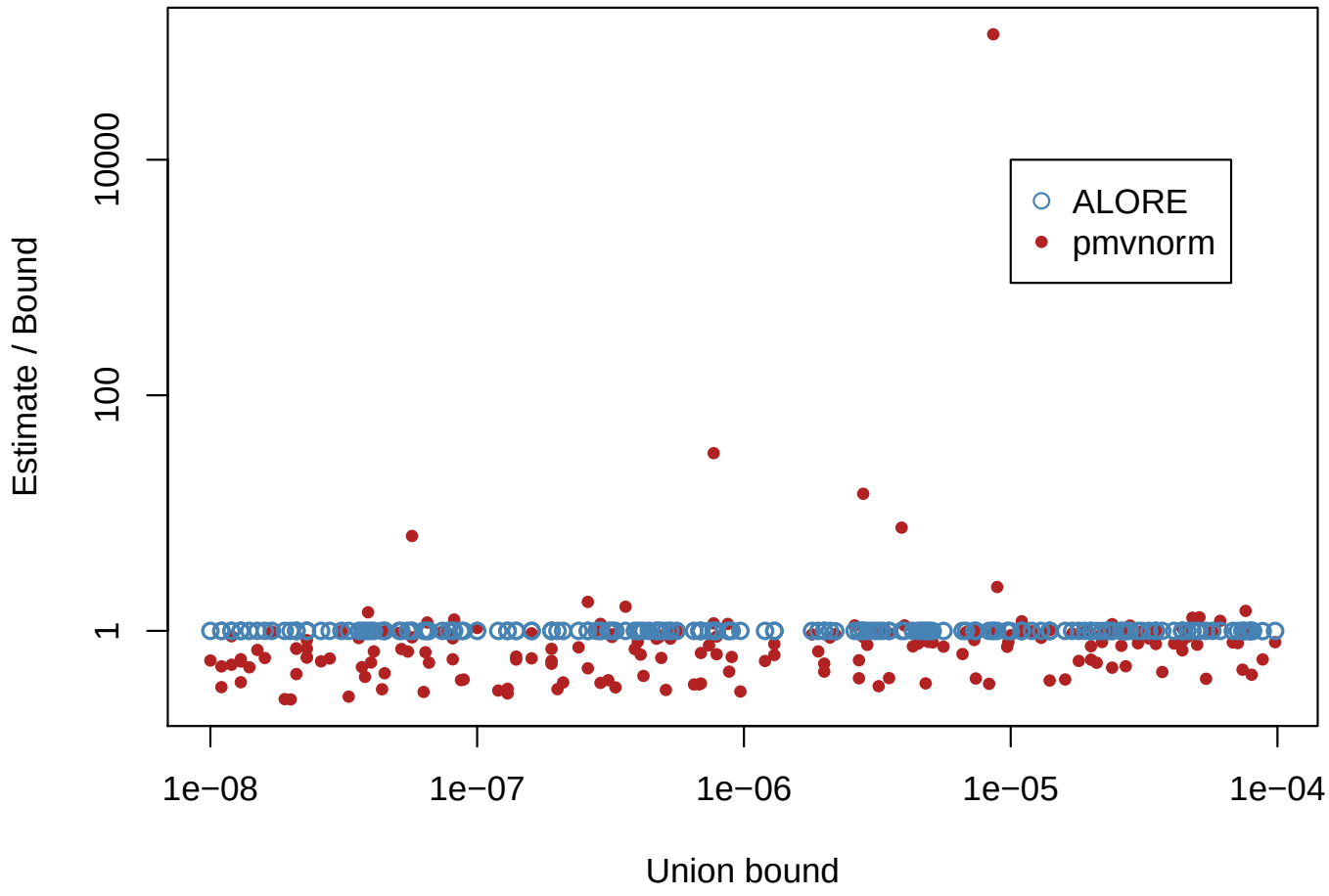
By concentration of measure they're nearly orthogonal.

$$\mu \doteq 1 - \prod_{j=1}^{J}(1 - P_j) \doteq \sum_j P_j = \bar{\mu} \quad \text{(for rare events)}$$

Simulation

- Dimension $d \sim \mathbf{U}\{20, 50, 100, 200, 500\}$
- Constraints $J \sim \mathbf{U}\{d/2, d, 2d\}$
- τ such that $-\log_{10}(\bar{\mu}) \sim \mathbf{U}[4, 8]$ (then rounded to 2 places).

High dimensional results



Results of 200 estimates of the μ for varying high dimensional problems with nearly independent events.

Power systems

Network with N nodes called busses and M edges.

Typically M/N is not very large.

Power p_i at bus i. Each bus is Fixed or Random or Slack:

slack bus S has $p_S = -\sum_{i \neq S} p_i$.

$$p = (p_F^{\mathsf{T}}, p_R^{\mathsf{T}}, p_S)^{\mathsf{T}} \in \mathbb{R}^N$$

Randomness driven by $p_R \sim \mathcal{N}(\eta_R, \Sigma_{RR})$

Inductances B_{ij}

A Laplacian $B_{ij} \neq 0$ if $i \sim j$ in the network. $B_{ii} = -\sum_{j \neq i} B_{ij}$.

Taylor approximation

$$\theta = B^+ p \quad \text{(pseudo inverse)}$$

Therefore θ is Gaussian as are all $\theta_i - \theta_j$ for $i \sim j$.

Examples

Our examples come from MATPOWER (a Matlab toolbox).

Zimmernan, Murillo-Sánchez & Thomas (2011)

We used $n = 10{,}000$.

Polish winter peak grid

2383 busses, $d = 326$ random busses, $J = 5772$ phase constraints.

$\bar{\omega}$	$\hat{\mu}$	se$/\hat{\mu}$	$\underline{\mu}$	$\bar{\mu}$
$\pi/4$	3.7×10^{-23}	0.0024	3.6×10^{-23}	4.2×10^{-23}
$\pi/5$	2.6×10^{-12}	0.0022	2.6×10^{-12}	2.9×10^{-12}
$\pi/6$	3.9×10^{-07}	0.0024	3.9×10^{-07}	4.4×10^{-07}
$\pi/7$	2.0×10^{-03}	0.0027	2.0×10^{-03}	2.4×10^{-03}

$\bar{\omega}$ is the phase constraint, $\hat{\mu}$ is the ALORE estimate, se is the estimated standard error, $\underline{\mu}$ is the largest single event probability and $\bar{\mu}$ is the union bound.

Pegase 2869

Fliscounakis et al. (2013) "large part of the European system"

$N = 2869$ busses. $d = 509$ random busses. $J = 7936$ phase constraints.

Results

$\bar{\omega}$	$\hat{\mu}$	se$/\hat{\mu}$	$\underline{\mu}$	$\bar{\mu}$
$\pi/2$	3.5×10^{-20}	0^*	3.3×10^{-20}	3.5×10^{-20}
$\pi/3$	8.9×10^{-10}	5.0×10^{-5}	7.7×10^{-10}	8.9×10^{-10}
$\pi/4$	4.3×10^{-06}	1.8×10^{-3}	3.5×10^{-06}	4.6×10^{-06}
$\pi/5$	2.9×10^{-03}	3.5×10^{-3}	1.8×10^{-03}	4.1×10^{-03}

Notes

$\bar{\theta} = \pi/2$ is unrealistically large. We got $\hat{\mu} = \bar{\mu}$. All $10{,}000$ samples had $S = 1$.

Some sort of Wilson interval or Bayesian approach could help.

One half space was sampled 9408 times, another 592 times.

Other models

IEEE case 14 and IEEE case 300 and Pegase 1354 were all dominated by one failure mode so $\underline{\mu} \doteq \bar{\mu}$ and no sampling is needed.

Another model had random power corresponding to wind generators but phase failures were not rare events in that model.

The Pegase 13659 model was too large for our computer. The Laplacian had $37{,}250$ rows and columns.

The Pegase 9241 model was large and slow and phase failure was not a rare event.

Caveats

We used a DC approximation to AC power flow (which is common) and the phase estimates were based on a Taylor approximation.

Next steps

1) Nonlinear boundaries

2) Non-Gaussian models

3) Optimizing cost subject to a constraint on μ

Sampling half-spaces will work if we have convex failure regions and a log-concave nominal density.

Thanks

- Michael Chertkov and Yury Maximov, co-authors
- Center for Nonlinear Studies at LANL, for hospitality
- Alan Genz for help on `mvtnorm`
- Bert Zwart, pointing out Adler et al.
- Yanbo Tang and Jeffrey Negrea, improved Lemma proof
- NSF DMS-1407397 & DMS-1521145
- DOE/GMLC 2.0 Project: Emergency monitoring and controls through new technologies and analytics
- Jianfeng Lu, Ilse Ipsen
- Sue McDonald, Kerem Jackson, Thomas Gehrmann

Bonus topic

Thinning MCMC output:

It really can improve *statistical* efficiency.

Short story

If it costs 1 unit to advance $\boldsymbol{x}_{i-1} \to \boldsymbol{x}_i$

and $\theta > 0$ units to compute $y_i = f(\boldsymbol{x}_i)$

then thinning to every k'th value lets us get larger n

and less variance if ACF(y_i) decays slowly.

MCMC (notation for)

A simple MCMC generates:

$$\boldsymbol{x}_i = \varphi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i) \in \mathbb{R}^d, \quad \boldsymbol{u}_i \sim \mathbf{U}(0,1)^m$$

More general ones use $\boldsymbol{u}_i \in \mathbb{R}^{m_i}$ where m_i may be random.

E.g., step i consume m_i uniform random variables.

The function φ is constructed so $\boldsymbol{x}_i$ has desired stationary distribution π.

We approximate

$$\mu = \int f(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{by} \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)$$

For simplicity, ignore burn-in / warmup.

Thinning

Get $\boldsymbol{x}_{ki}$ and $f(\boldsymbol{x}_{ki})$ for $i = 1, \ldots, n$ and $k \geqslant 1$.

- Thinning by a factor k usually reduces autocorrelations.
- Then $f(\boldsymbol{x}_{ki})$ are "more nearly IID".
- Thinning can also save storage $\boldsymbol{x}_i$.

Statistical efficiency

Let $y_i = f(\boldsymbol{x}_i) \in \mathbb{R}$. Geyer (1992) shows that for $k > 1$

$$\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} y_i\Big) \leqslant \mathrm{Var}\Big(\frac{1}{\lfloor n/k \rfloor}\sum_{i=1}^{\lfloor n/k \rfloor} y_{ki}\Big)$$

Quotes

Link & Eaton (2011):

> "Thinning is often unnecessary and always inefficient."

MacEachern & Berliner (1994):

> "This article provides a justification of the ban against sub-sampling the output of a stationary Markov chain that is suitable for presentation in undergraduate and beginning graduate-level courses."

Gamerman & Lopes (2006)' on thinning the Gibbs sampler:

> "There is no gain in efficiency, however, by this approach and estimation is shown below to be always less precise than retaining all chain values."

Not so fast

The analysis assumes that we compute $f(\boldsymbol{x}_i)$ and only use every k'th one.

Suppose that it costs

- 1 unit to advance the chain: $\boldsymbol{x}_i \to \boldsymbol{x}_{i+1}$.
- θ units to compute $y_i = f(\boldsymbol{x})$.

If we thin the chain we compute f less often and can use larger n.

When it pays

If θ is large and $\mathrm{Corr}(y_i, y_{i+k})$ decays slowly, then thinning can be much more efficienty.

When can that happen?

E.g., $\boldsymbol{x}$ describes a set of particles and f computes interpoint distances.
Updating f will cost proportionally to updating $\boldsymbol{x}$, maybe much more.

Thinned estimate

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} f(x_{ik})$$

kn_k advances $\boldsymbol{x}_i \to \boldsymbol{x}_{i+1}$ and

n_k computations $\boldsymbol{x}_i \to y_i = f(\boldsymbol{x}_i)$.

Budget: $kn_k + \theta n_k \leqslant B$

$$n_k = \left\lfloor \frac{B}{k+\theta} \right\rfloor \approx \frac{B}{k+\theta}.$$

Variance

$$\mathrm{Var}(\hat{\mu}_k) \doteq \frac{\sigma^2}{n_k} \left(1 + 2 \sum_{\ell=1}^{\infty} \rho_{k\ell} \right), \quad \rho_\ell = \mathrm{Corr}(y_t, y_{t+\ell}).$$

Efficiency

$$\mathrm{eff}(k) = \frac{\mathrm{Var}(\hat{\mu}_1)}{\mathrm{Var}(\hat{\mu}_k)} \doteq \frac{\frac{\sigma^2}{n_1}\left(1 + 2\sum_{\ell=1}^{\infty}\rho_\ell\right)}{\frac{\sigma^2}{n_k}\left(1 + 2\sum_{\ell=1}^{\infty}\rho_{k\ell}\right)} = \frac{n_k\left(1 + 2\sum_{\ell=1}^{\infty}\rho_\ell\right)}{n_1\left(1 + 2\sum_{\ell=1}^{\infty}\rho_{k\ell}\right)}$$

NB: $n_k < n_1$ but ordinarily $\sum_{\ell=1}^{\infty}\rho_\ell > \sum \rho_{k\ell}$.

Sample size ratio

$$\frac{n_k}{n_1} \approx \frac{B/(k+\theta)}{B/(1+\theta)} = \frac{1+\theta}{k+\theta}$$

Therefore

$$\mathrm{eff}(k) = \frac{(1+\theta)\left(1 + 2\sum_{\ell=1}^{\infty}\rho_\ell\right)}{(k+\theta)\left(1 + 2\sum_{\ell=1}^{\infty}\rho_{k\ell}\right)}$$

Autocorrelation models

ACF plots often resemble AR(1):

$$\rho_\ell = \rho^{|\ell|}, \quad 0 < \rho < 1$$

E.g., figures in Jackman (2009). Newman & Barkema (1999):

> "the autocorrelation is expected to fall off exponentially at long times".

Geyer (1991) notes exponential upper bound under ρ-mixing.

Monotone non-negative autocorrelations

$$\rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant \rho_\ell \geqslant \rho_{\ell+1} \geqslant \cdots \geqslant 0$$

Under the AR(1) model

$$\begin{aligned}
\mathrm{eff}(k) = \mathrm{eff}_{\mathrm{AR}}(k) &= \frac{(1+\theta)\left(1+2\sum_{\ell=1}^{\infty}\rho^{\ell}\right)}{(k+\theta)\left(1+2\sum_{\ell=1}^{\infty}\rho^{k\ell}\right)} \\
&= \frac{(1+\theta)\left(1+2\rho/(1-\rho)\right)}{(k+\theta)\left(1+2\rho^{k}/(1-\rho^{k})\right)} \\
&\vdots \\
&= \frac{1+\theta}{k+\theta}\,\frac{1+\rho}{1-\rho}\,\frac{1-\rho^{k}}{1+\rho^{k}}
\end{aligned}$$

For large θ and ρ near 1, thinning will be very efficient.

Optimal thinning factor k

$\theta \setminus \rho$	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999	0.999999
0.001	1	1	1	4	18	84	391	1817
0.01	1	1	2	8	39	182	843	3915
0.1	1	1	4	18	84	391	1817	8434
1	1	2	8	39	182	843	3915	18171
10	2	4	17	83	390	1816	8433	39148
100	3	7	32	172	833	3905	18161	84333
1000	4	10	51	327	1729	8337	39049	181612

θ is the cost to compute $y = f(\boldsymbol{x})$.

ρ is the autocorrelation.

Efficiency of optimal k

$\theta \setminus \rho$	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999	0.999999
0.001	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01
0.1	1.00	1.00	1.06	1.09	1.10	1.10	1.10	1.10
1	1.00	1.20	1.68	1.93	1.98	2.00	2.00	2.00
10	1.10	2.08	5.53	9.29	10.59	10.91	10.98	11.00
100	1.20	2.79	13.57	51.61	85.29	97.25	100.17	100.82
1000	1.22	2.97	17.93	139.29	512.38	845.38	963.79	992.79

Versus $k = 1$.

Least k for 95% efficiency

$\theta \setminus \rho$	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999	0.999999
0.001	1	1	1	1	1	1	1	1
0.01	1	1	1	1	1	1	1	1
0.1	1	1	2	2	2	2	2	2
1	1	2	5	11	17	19	19	19
10	2	4	12	45	109	164	184	189
100	2	5	22	118	442	1085	1632	1835
1000	2	6	31	228	1182	4415	10846	16311

Table 1: Smallest k to give at least 95% of the efficiency of the most efficient k, as a function of θ and the autoregression parameter ρ.

Additional

- Thinning can pay when $\rho_1 > 0$ and $\rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant 0$.
- $\theta > 0$ reduces the optimal acceptance rate from 0.234

  Jeffrey Rosenthal has a quantitative version.