Optimising Twin Uniform Distribution for Multiplicative Noise Data Masking

Pauline Ding*, University of Wollongong, Australia

AASC 2024

This is a joint work with H. Rowles, J. Brackenbury and Y.X. Lin

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Multiplicative Noise for Data Masking

Introduction

- The perturbed value for measure X_i of the ith individual is

$$\tilde{X}_i = X_i \times M_i \ .$$

- M_i is a random multiplicate noise, i.i.d. sampled from noise M.
- Multiplicative Noise scheme has uniform protection regardless of original data value.
- Careful selection of M can lead to desirable statistical disclosure control results, i.e.,
  - remaining accurate statistical properties and
  - avoiding value disclosure.

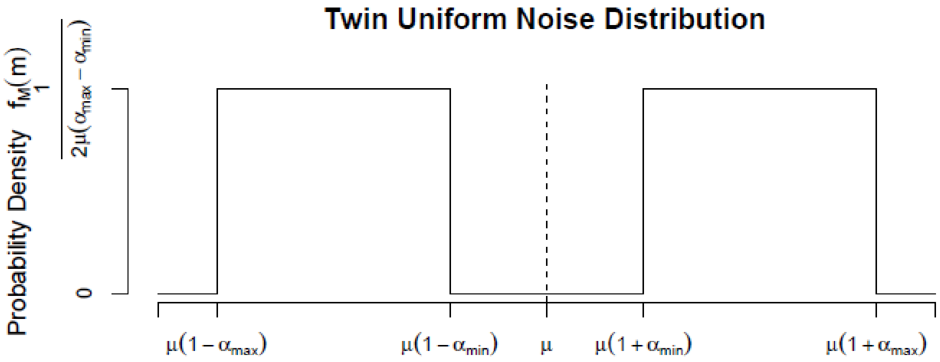
Multiplicative Noise for Data masking

Twin Uniform Noise Distribution

- Brackenbury et al (2020) proposed a *twin uniform* distribution for M

$$f_M(m) = \begin{cases} \frac{1}{2\mu(\alpha_{\max}-\alpha_{\min})}, & m \in A \\ 0, & \text{else}, \end{cases} \tag{1}$$

where $A = [\mu(1-\alpha_{\max}), \mu(1-\alpha_{\min})] \cup [\mu(1+\alpha_{\min}), \mu(1+\alpha_{\max})]$ for $\mu \in \mathbb{R}$ and $\alpha_{min} < \alpha_{max} \in \mathbb{R}^+$. (1) is denoted as $M \sim \text{TwinUnif}(\mu, \alpha_{\min}, \alpha_{\max})$.



Multiplicative Noise for Data masking

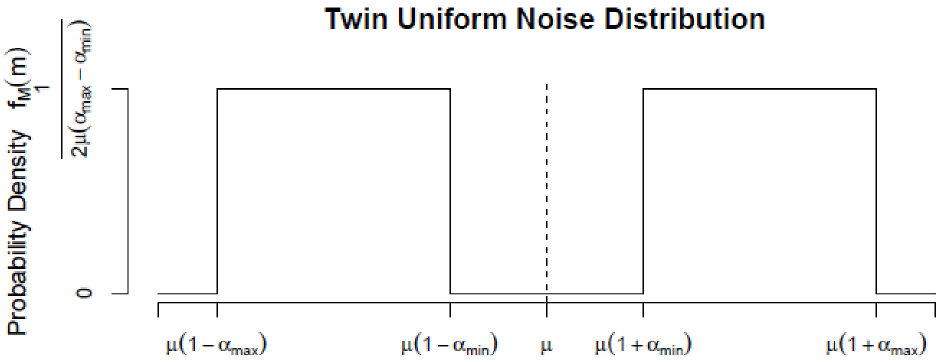
Twin Uniform Noise Distribution

- Brackenbury et al (2020) proposed a *twin uniform* distribution for M

$$f_M(m) = \begin{cases} \frac{1}{2\mu(\alpha_{\max}-\alpha_{\min})}, & m \in A \\ 0, & \text{else}, \end{cases} \qquad (1)$$

where $A = [\mu(1-\alpha_{\max}), \mu(1-\alpha_{\min})] \cup [\mu(1+\alpha_{\min}), \mu(1+\alpha_{\max})]$ for $\mu \in \mathbb{R}$ and $\alpha_{min} < \alpha_{max} \in \mathbb{R}^+$. (1) is denoted as $M \sim \text{TwinUnif}(\mu, \alpha_{\min}, \alpha_{\max})$.

- $E(M) = \mu$
- $var(M) = \mu^2(\alpha_{max} + \alpha_{max}\ \alpha_{min} + \alpha_{min}^2)/3$



Properties of Twin Uniform Noise Distribution

Disclosure risks and Shift

- Primary Disclosure Risk

$$p_\delta(\alpha_{\max}, \alpha_{\min}) = P\left(\left|\frac{\hat{X}_i - X_i}{X_i}\right| < \delta\right) = P\left(\left|\frac{M}{\mu} - 1\right| < \delta\right)$$

$$= P[\mu(1-\delta) < M < \mu(1-\alpha_{\min})] + P[\mu(1+\alpha_{\min}) < M < \mu(1+\delta)]$$

$$= \frac{\delta - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}.$$

Properties of Twin Uniform Noise Distribution

Disclosure risks and Shift

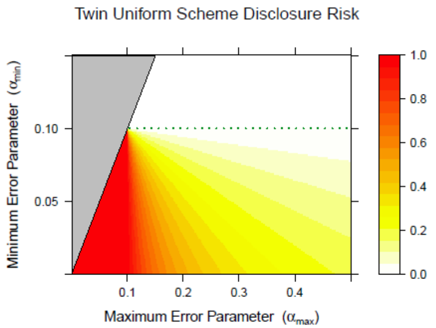
- Primary Disclosure Risk

$$p_\delta(\alpha_{\max}, \alpha_{\min}) = P\left(\left|\frac{\hat{X}_i - X_i}{X_i}\right| < \delta\right) = P\left(\left|\frac{M}{\mu} - 1\right| < \delta\right)$$

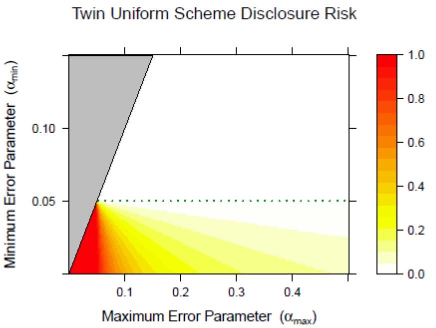
$$= P[\mu(1-\delta) < M < \mu(1-\alpha_{\min})] + P[\mu(1+\alpha_{\min}) < M < \mu(1+\delta)]$$

$$= \frac{\delta - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}.$$

- Primary disclosure risk can be eliminated when setting $\alpha_{min} = \delta$.



Probability of Disclosure below Level $\delta = 0.1$



Probability of Disclosure below Level $\delta = 0.05$

Properties of Twin Uniform Noise Distribution

Disclosure risks and Shift

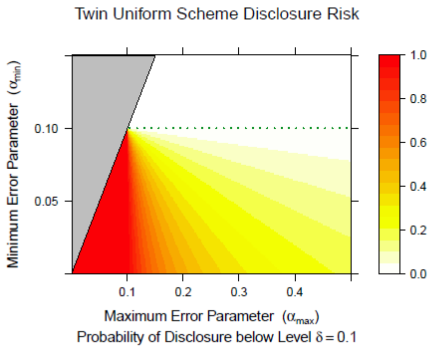
- **Primary Disclosure Risk**

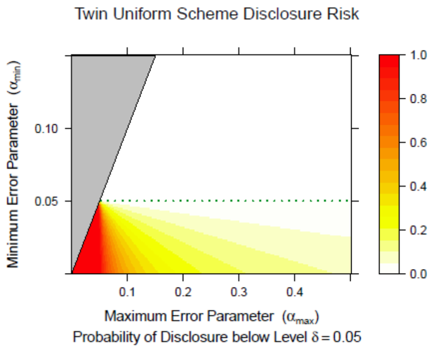
$$p_\delta(\alpha_{\max}, \alpha_{\min}) = P\left(\left|\frac{\hat{X}_i - X_i}{X_i}\right| < \delta\right) = P\left(\left|\frac{M}{\mu} - 1\right| < \delta\right)$$

$$= P[\mu(1-\delta) < M < \mu(1-\alpha_{\min})] + P[\mu(1+\alpha_{\min}) < M < \mu(1+\delta)]$$

$$= \frac{\delta - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}.$$

- **Primary disclosure risk can be eliminated when setting $\alpha_{min} = \delta$.**





- **Secondary Disclosure Risk**
  - An issue of regression risk $corr(\tilde{X}_i, X_i)$
  - Ma et al (2019) showed that a correlation below 0.8 gives large enough estimation uncertainty.
  - α_{max} needs to be large enough.

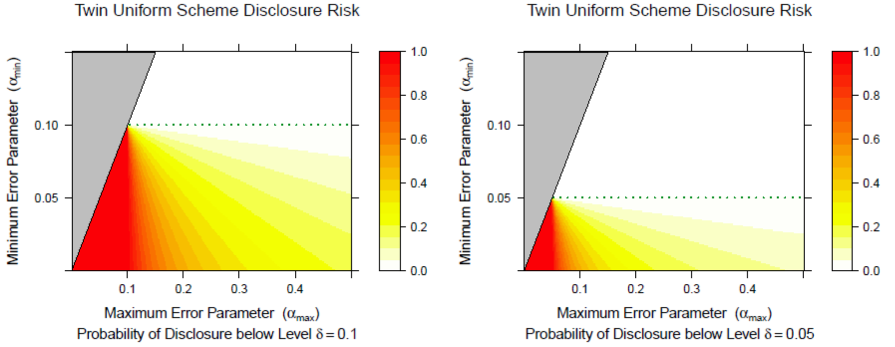
Properties of Twin Uniform Noise Distribution

Disclosure risks and Shift

- Primary Disclosure Risk

$$p_\delta(\alpha_{\max}, \alpha_{\min}) = P\left(\left|\frac{\hat{X}_i - X_i}{X_i}\right| < \delta\right) = P\left(\left|\frac{M}{\mu} - 1\right| < \delta\right)$$
$$= P[\mu(1-\delta) < M < \mu(1-\alpha_{\min})] + P[\mu(1+\alpha_{\min}) < M < \mu(1+\delta)]$$
$$= \frac{\delta - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}.$$

- Primary disclosure risk can be eliminated when setting $\alpha_{min} = \delta$.



- Secondary Disclosure Risk
  - An issue of regression risk $corr(\tilde{X}_i, X_i)$
  - Ma et al (2019) showed that a correlation below 0.8 gives large enough estimation uncertainty.
  - α_{max} needs to be large enough.

- Shifting
  - To avoid no perturbation to $X_i = 0$, shifted multiplicative masked data is

$$\hat{Y}_i = (X_i + a)M_i,$$

  - $\hat{Y}_i$ is an unbiased estimator of X_i
  - Estimator of the sum is unbiased.

Parameters for Optimisation

2 Parameters: $\alpha_{\max}, a$ (shift)

- Regression Risk

$$corr_{M;\hat{Y}_i,Y_i}(\alpha_{max}, a) = \frac{cov(\hat{Y}_i, Y_i)}{\sqrt{Var(\hat{Y}_i)Var(Y_i)}}$$

$$= \frac{Var(X_i)}{\sqrt{Var(X_i)^2 + \frac{\alpha_{\max}^2 + \alpha_{\max}\alpha_{\min} + \alpha_{\min}^2}{3}[E(X_i) + a]^2}}$$

- Utility (of the sum)
  - Conditional standard error (CSE) given observed $\{X_i\}$ is

$$CSE_M(\alpha_{max}, a) = \sqrt{\frac{\alpha_{max}^2 + \alpha_{max}\alpha_{min} + \alpha_{min}^2}{3} \sum_{i=1}^{N} (X_i + a)^2}.$$

Parameters for Optimisation

2 Parameters: $\alpha_{\max}$, a (shift)

- Regression Risk

$$corr_{M;\hat{Y}_i,Y_i}(\alpha_{max}, a) = \frac{cov(\hat{Y}_i, Y_i)}{\sqrt{Var(\hat{Y}_i)Var(Y_i)}}$$

$$= \frac{Var(X_i)}{\sqrt{Var(X_i)^2 + \frac{\alpha_{\max}^2 + \alpha_{\max}\alpha_{\min} + \alpha_{\min}^2}{3}[E(X_i) + a]^2}}$$

- Utility (of the sum)
  - Conditional standard error (CSE) given observed $\{X_i\}$ is

$$CSE_M(\alpha_{max}, a) = \sqrt{\frac{\alpha_{max}^2 + \alpha_{max}\alpha_{min} + \alpha_{min}^2}{3}\sum_{i=1}^{N}(X_i + a)^2}$$

		Effects on Measures	
Parameter	Change	Regression Risk $\mathrm{corr}_M(\hat{Y}_i, Y_i)$	Estimation Error CSE_M
Shift Value a	↑	↓	↑
	↓	↑	↓
Upper bound $\alpha_{\max}$	↑	↓	↑
	↓	↑	↓

Multi-objective Optimisation (Pareto Optimality)

Goal: minimize regression risk $corr$ and minimize estimation error CSE

Normalisation of the measures

$$g_{norm}(x) = \frac{g(x) - \min[g(x)]}{\max[g(x)] - \min[g(x)]}.$$

Multi-objective Optimisation (Pareto Optimality)

Goal: minimize regression risk $corr$ and minimize estimation error CSE

Normalisation of the measures

$$g_{norm}(x) = \frac{g(x) - \min[g(x)]}{\max[g(x)] - \min[g(x)]}.$$

- $$\mathrm{corr}_{norm}(\alpha_{max}, a) = \frac{\mathrm{corr}_M(\alpha_{max}, a) - \min[\mathrm{corr}_M(\alpha_{max}, a)]}{\max[\mathrm{corr}_M(\alpha_{max}, a)] - \min[\mathrm{corr}_M(\alpha_{max}, a)]}$$

  and

  $$CSE_{norm}(\alpha_{max}, a) = \frac{CSE_M(\alpha_{max}, a) - \min[CSE_M(\alpha_{max}, a)]}{\max[CSE_M(\alpha_{max}, a)] - \min[CSE_M(\alpha_{max}, a)]}.$$

Multi-objective Optimisation (Pareto Optimality)

Goal: minimize regression risk *corr* and minimize estimation error *CSE*

Normalisation of the measures

$$g_{norm}(x) = \frac{g(x) - \min[g(x)]}{\max[g(x)] - \min[g(x)]}.$$

$$\mathrm{corr}_{norm}(\alpha_{max}, a) = \frac{\mathrm{corr}_M(\alpha_{max}, a) - \min[\mathrm{corr}_M(\alpha_{max}, a)]}{\max[\mathrm{corr}_M(\alpha_{max}, a)] - \min[\mathrm{corr}_M(\alpha_{max}, a)]}$$

and

$$CSE_{norm}(\alpha_{max}, a) = \frac{CSE_M(\alpha_{max}, a) - \min[CSE_M(\alpha_{max}, a)]}{\max[CSE_M(\alpha_{max}, a)] - \min[CSE_M(\alpha_{max}, a)]}.$$

Linear weighted sum optimisation method

$$\begin{aligned} g_{combinded}(\alpha_{max}, a) &= W_1 \cdot \mathrm{corr}_{norm}(\alpha_{max}, a) + W_2 \cdot CSE_{norm}(\alpha_{max}, a) \\ &= 0.5\,\mathrm{corr}_{norm}(\alpha_{max}, a) + 0.5\,CSE_{norm}(\alpha_{max}, a). \end{aligned}$$

Multi-objective Optimisation (Pareto Optimality)

Goal: minimize regression risk *corr* and minimize estimation error *CSE*

Normalisation of the measures

$$g_{norm}(x) = \frac{g(x) - \min[g(x)]}{\max[g(x)] - \min[g(x)]}.$$

$$\text{corr}_{norm}(\alpha_{max}, a) = \frac{\text{corr}_M(\alpha_{max}, a) - \min[\text{corr}_M(\alpha_{max}, a)]}{\max[\text{corr}_M(\alpha_{max}, a)] - \min[\text{corr}_M(\alpha_{max}, a)]}$$

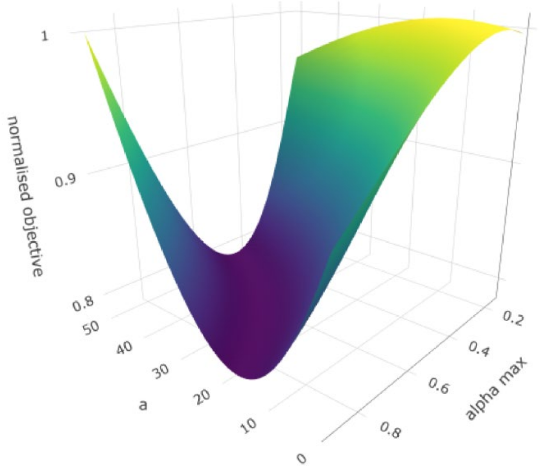
and

$$CSE_{norm}(\alpha_{max}, a) = \frac{CSE_M(\alpha_{max}, a) - \min[CSE_M(\alpha_{max}, a)]}{\max[CSE_M(\alpha_{max}, a)] - \min[CSE_M(\alpha_{max}, a)]}.$$

Linear weighted sum optimisation method

$$\begin{aligned} g_{combinded}(\alpha_{max}, a) &= W_1 \cdot \text{corr}_{norm}(\alpha_{max}, a) + W_2 \cdot CSE_{norm}(\alpha_{max}, a) \\ &= 0.5\,\text{corr}_{norm}(\alpha_{max}, a) + 0.5\,CSE_{norm}(\alpha_{max}, a). \end{aligned}$$

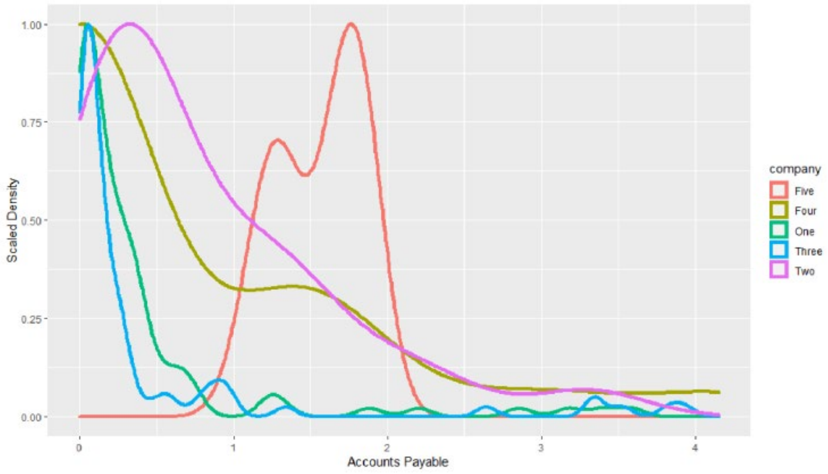
The solution for minizine the combined objective is not unique.



Real Data Illustration

2019/20 US Public Companies Accounts Payable Data of 5 companies

- Aim: to protect individual transaction but allow total yearly debt (sum) accurately estimated.



Real Data Illustration

2019/20 US Public Companies Accounts Payable Data

- Aim: to protect individual transaction but allow total yearly debt (sum) accurately estimated.

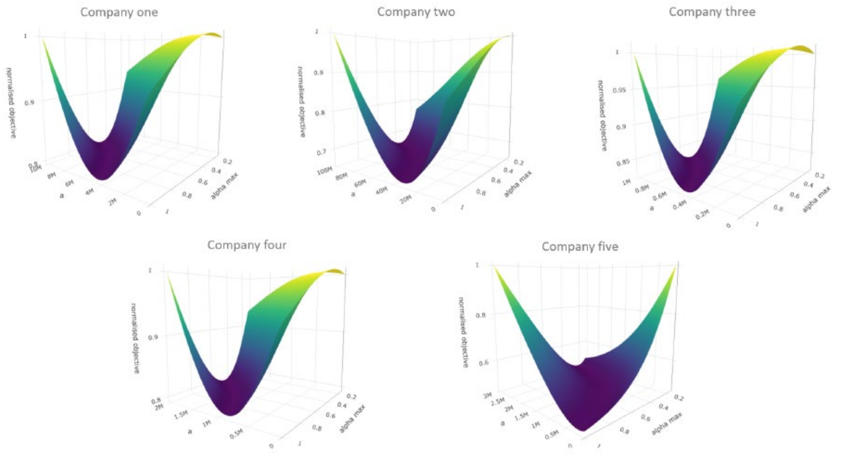
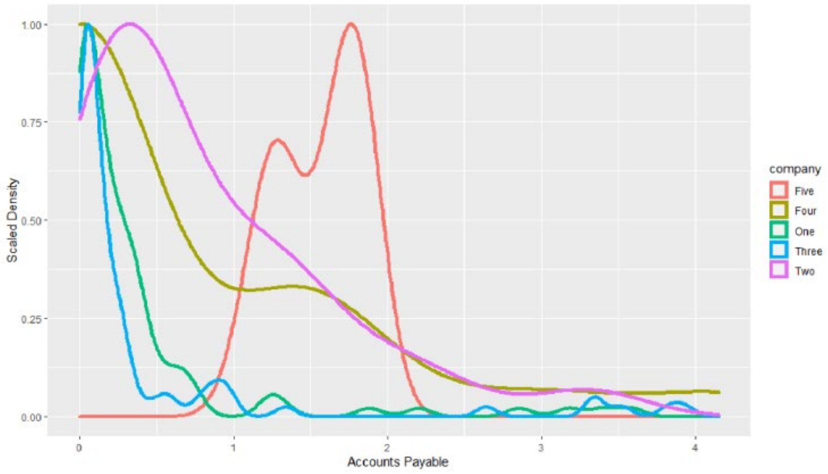


Fig. 5: 3D plot of $g_{combinded,c}$ (10) for five companies against α_{max} and a.



- Equally weighted sum optimisation with normalisations

$$g_{combinded,c}(\alpha_{max}, a) = 0.5\,\mathrm{corr}_{norm,c}(\alpha_{max}, a) + 0.5\,CSE_{norm,c}(\alpha_{max}, a)$$

Results and Closing Remarks

- We selected optimal values for α_{max} and a , showed good performance in the empirical measure of disclosure (sample correlation) and utility (relative error of sum estimates).

Company	Optimal (α_{max}, a)	Sample Correlation	Relative Error of Sum
1	(0.914, 4.54)	0.463	0.052
2	(0.571, 65.66)	0.204	0.064
3	(0.657, 0.78)	0.623	0.018
4	(0.519, 1.96)	0.586	0.01
5	(0.373, 0.55)	0.523	0.023

Results and Closing Remarks

- We selected optimal values for α_{max} and a , showed good performance in the empirical measure of disclosure (sample correlation) and utility (relative error of sum estimates).

Company	Optimal (α_{max}, a)	Sample Correlation	Relative Error of Sum
1	(0.914, 4.54)	0.463	0.052
2	(0.571, 65.66)	0.204	0.064
3	(0.657, 0.78)	0.623	0.018
4	(0.519, 1.96)	0.586	0.01
5	(0.373, 0.55)	0.523	0.023

- Closing remarks
    - Twin uniform distribution can be optimisation in term of disclosure and utility simultaneously.
    - Multi-objective optimisation with normalisation allows derivation of optimal solutions, which is case dependent and not unique.

Thanks for your time.

poshaugh@uow.edu.au

UNIVERSITY
OF WOLLONGONG
AUSTRALIA