Disclosure Limitation and Large Gaps in Small Hierarchical Models

Edwin O'Shea

de Brún centre, NUI Galway

http://www.maths.nuigalway.ie/∼edwin

Limiting Disclosure: Censoring

For the benefit of its students University of X releases to the public both:

- average of final grades for each course.
- number of graduates in each degree option.

Only final year statistics students can take "Stochastic Processes" and in 2009 only one person graduated from UofX with a degree in statistics.

Question Should UofX publicly release the average grade for "Stochastic Processes" and the number of graduates in Statistics for 2009 ?

In the interest of privacy for the one statistics graduate, UofX should not disclose (release to the public) the average grade in "Stochastic Processes". i.e. the "Stochastic Processes" average should be censored.

Limiting Disclosure: Lower and Upper bounds

2×2 tables:

u_{11}	u_{12}	60
u_{21}	u_{22}	40
70	30	100

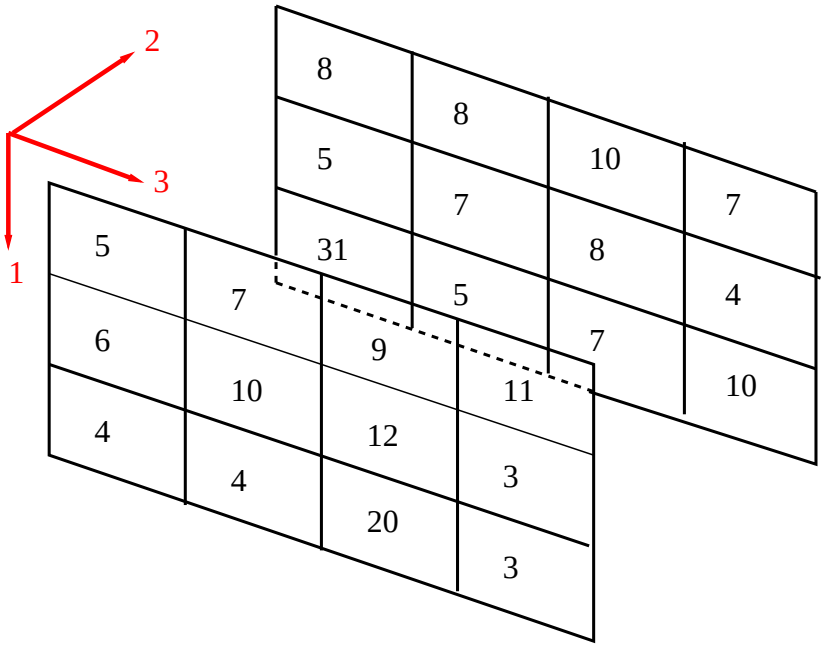
$$u_{11} + u_{12} = 60;\ u_{21} + u_{22} = 40;$$
$$u_{11} + u_{21} = 70;\ u_{12} + u_{22} = 30;$$
$$u_{11}, u_{12}, u_{21}, u_{22} \geq 0 \text{ and integral}$$

Fréchet bounds on 2×2 tables:

$$30 = \max\{0,\ 60 + 70 - 100\} \leq u_{11} \leq \min\{60,\ 70\} = 60$$

Upshot If u_{11} was a person's individual private data then u_{11} is secure.

Tables and Marginals



inter $\{1,2\}$ =

32	33
35	24
31	53

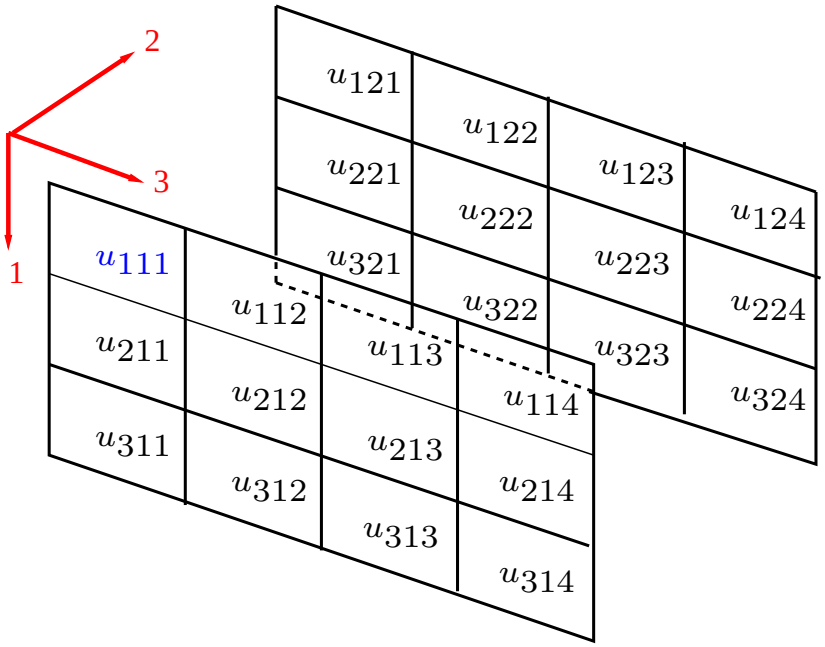
inter $\{2,3\}$ =

15	21	41	21
44	20	25	21

The hierarchical model $\Delta = \{\{1,2\}, \{2,3\}\}$ on $3 \times 2 \times 4$ tables.

Marginal $\mathbf{b} := [\text{inter}\{1,2\}, \text{inter}\{2,3\}] = (A_{\Delta,3\times2\times4})\mathbf{u},\ \mathbf{u} \geq \mathbf{0},\ \mathbf{u}$ integral

Given the publicly released $\mathbf{b}$ how secure is u_{111} ?



$$\min\{u_{111} : A\mathbf{u} = \mathbf{b},\ \mathbf{u} \geq \mathbf{0}\ \textbf{real}\}$$
$$\leq \min\{u_{111} : A\mathbf{u} = \mathbf{b},\ \mathbf{u} \geq \mathbf{0}\ \textbf{integral}\}$$
$$\leq \textbf{true}\ u_{111}\ \textbf{in secret table}$$
$$\leq \max\{u_{111} : A\mathbf{u} = \mathbf{b},\ \mathbf{u} \geq \mathbf{0}\ \textbf{integral}\}$$
$$\leq \max\{u_{111} : A\mathbf{u} = \mathbf{b},\ \mathbf{u} \geq \mathbf{0}\ \textbf{real}\}$$

Problem: If $\mathbf{b}$ is unknown apriori, can security of u_1 be ensured ?

$\max_{\mathbf{b}}(\min\{u_1\ \textbf{integral}\} - \min\{u_1\ \textbf{real}\}) = \text{gap}_{-}(A_{\Delta, d_1 \times \cdots \times d_n})$

Note Smaller gaps imply possibly less protection for individual cell entries.

Detecting Gaps Quickly

[HoştenSturmfels] The gap can be computed precisely via irreducible decompositions of monomial ideals and solving group relaxations.

[HoştenSturmfels] **Quick gap** if $\mathbf{g} = \mathbf{g}^+ - \mathbf{g}^- \in \mathcal{G}_{\mathbf{pert}}(A_\Delta)$ with $g_1 = \alpha$ then

$$\text{gap}_{-}(A_\Delta) \geq \alpha - 1 \qquad \text{(where } \mathbf{pert} = [1, \varepsilon, \varepsilon^2, \varepsilon^3, \ldots,]\text{)}$$

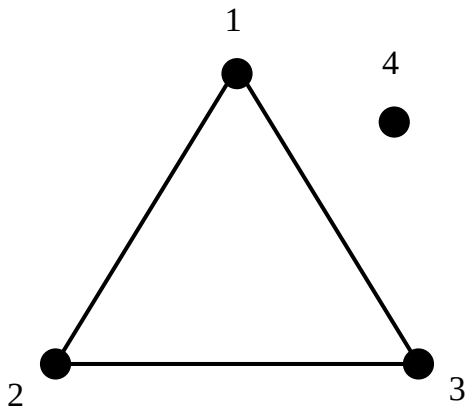
Proof Gröbner basis theory $\Rightarrow \mathbf{g}^+ - \mathbf{e_1}$ is integer optimal (for $[1, 0, 0, \ldots, 0]$)

but $(\mathbf{g}^+ - \mathbf{e_1}) - \frac{(\alpha-1)}{\alpha}(\mathbf{g})$ is optimal for linear relaxation.

Sullivant's construction

[Sullivant] $\exists$ models Δ_n $(\forall n \geq 4)$ on $2 \times 2 \times \cdots \times 2$ tables s.th.

$$\text{gap}_{-}(A_{\Delta_n}) \geq 2^{n-3} - 1$$

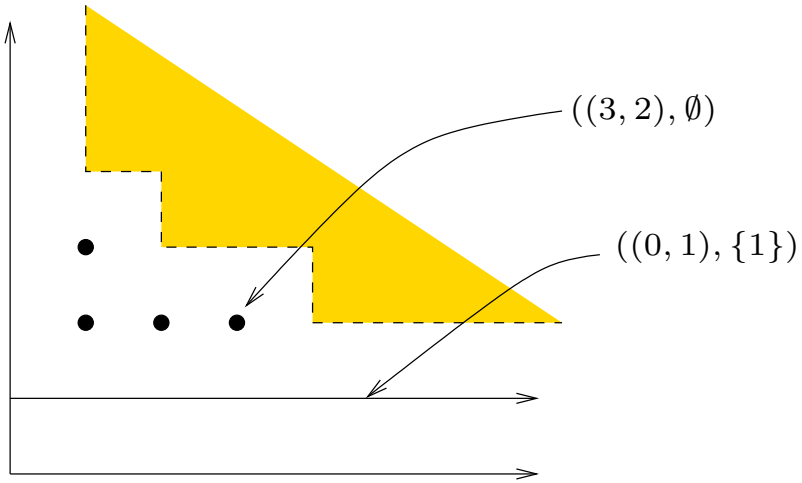


- $\Gamma_n := \partial((n-2)\text{-simplex}) + \text{point}$
- Graver element $\mathbf{g}$ with $g_1 = 2^{n-3}$
- $\Delta_n :=$ Lawrence(Γ_n)
- $\Rightarrow (\mathbf{g}, -\mathbf{g}) \in \mathcal{G}_{\mathbf{pert}}(A_{\Delta_n})$ with **1**-entry equal to 2^{n-3}

Summary There are marginals $\mathbf{b}$ for which Δ_n has large gaps. How common are these marginals ?

Measuring frequency with standard pairs

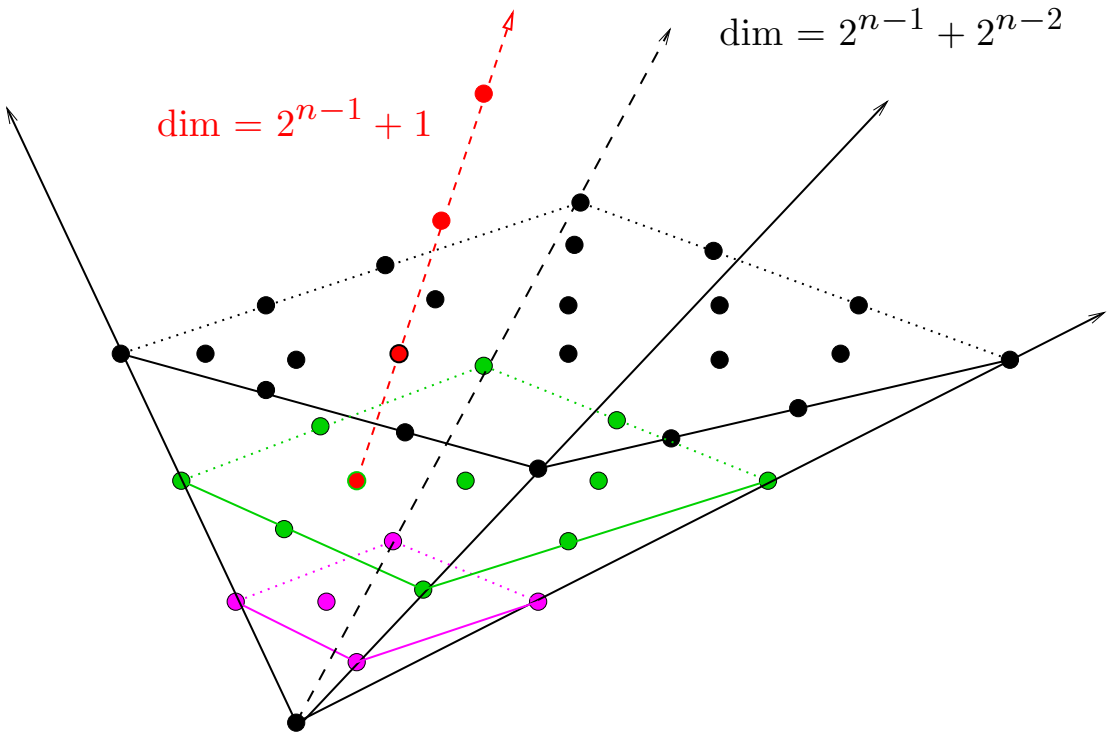
Standard pairs for $\mathrm{IP}_{A,(1,\mathbf{0})}(\mathbf{b}) := \min\{(1,\mathbf{0})\cdot\mathbf{u} \,:\, A\mathbf{u}=\mathbf{b},\, \mathbf{u}\in\mathbb{N}^N\}$



$(\gamma,\tau) := \{\gamma + \sum_{l\in\tau} n_l\mathbf{e}_l \,:\, n_l\in\mathbb{N}\}.$

- $\mathrm{supp}(\gamma)\cap\tau=\emptyset$
- every member of (γ,τ) is optimal
- $\nexists(\gamma',\tau')$ s.th. $(\gamma,\tau)\subset(\gamma',\tau')$

Sullivant's $2^{n-3}-1$ gap occurs rarely



[O'S., 2009] The marginals (●) that create Sullivant's $(2^{n-3}-1)$ gaps all come from a standard pair

$$((2^{n-3}-1)\cdot e_{\mathbf{1}},\ \sigma)$$

where $|\sigma| = 2^{n-1}+1$

This reopens the question of "*whether linear programming is an effective heuristic for detecting disclosures when releasing marginals of multi-way tables.*"

Work in progress

Other large quick gaps on different models were contructed [DevelinSullivant] and [HoştenSturmfels]. Computationally these are also rare.

Question Could it be that for any model all its quick gaps are rare ?

Problem Find a model $\mathcal{S}$ for which gap_$(\mathcal{S}) \geq 1$ and this gap is **not** rare.

Question If $\mathbf{g}$ is a Markov element of a model $\mathcal{S}$ with the support of $\mathbf{g}$ having size greater than the dimension of the model $\mathcal{S}$ then is $\mathbf{g}$ squarefree ?

i.e. $\mathbf{g} \in \text{Markov}(\mathcal{S}) \,\&\, |\text{supp}(\mathbf{g})| > \dim(\mathcal{S}) \implies \mathbf{g}$ has all entries $\{-1, 0, 1\}$. ?

Markov Bases Database `http://mbdb.mis.mpg.de/`