CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 2

29-31 October 2019, the Hague, the Netherlands 16 September 2019

Releasable inner cell frequencies by post-processing protected tabular data

Øyvind Langsrud (Statistics Norway)

Oyvind.Langsrud@ssb.no

Abstract and Paper

The microdata underlying tabular frequency data can be represented by the inner frequency table obtained by crossing all the main dimensional variables. This paper will discuss approaches to postprocessing the results of tabular SDC techniques to obtain releasable inner frequencies, which may not be whole numbers.

As suggested in a recent publication and implemented in the r-package RegSDC, suppressed cells may be replaced by decimal numbers, in a way so that they add correctly up to non-supressed marginals. A similar method could be used in cases without cell suppression, but where only some marginals are considered releasable. This method (Method A) can be viewed as ordinary regression, modified to handle collinearity, with the inner cells as the unknown parameters and with the releasable marginals as response observations. Method A can also be interpreted as linear calibration weighting. An improvement of this approach is to ensure non-negative values by applying nonnegative least squares with ridge penalty (Method B). A more natural choice in this case, would be to obtain inner frequencies as fitted values from log-linear modelling. Considering situations where the marginals are perturbed in a non-additive way, ordinary log-linear modelling is not possible, but Method B is useful. Log-linear fitting may be performed afterwards to obtain results in in accordance with such modelling assumptions. Another situation is when only some of the releasable marginals are perturbed. Then, it is advantageous to modify Method B by incorporating weights (Method C). The perturbed marginals are weighted down, so that the resulting inner frequencies add up almost exactly to the non-perturbed marginals. The methodology is applicable when the marginals are perturbed by Laplace noise to ensure differential privacy. Although, in this case, one may consider least absolute deviation instead of least squares.

The generated frequencies serve several purposes: They can be used internally or published directly. Additivity of marginals can be restored. One can also use the results to generate public microdata. Then, all frequencies must be whole numbers. A possibility is to round up or down by comparing the decimals to draws from the uniform distribution. This is an alternative to synthetic data drawn as a multinomial sample according to cell probabilities, as in the r-package synthpop.

The methodology is tested using example data of various size and complexity. For an efficient implementation, making use of the r-package glmnet seems promising.

Releasable inner cell frequencies by post-processing protected tabular data

Øyvind Langsrud*

* Statistics Norway, Department of methodology and data collection,
Oyvind.Langsrud@ssb.no, github.com/olangsrud

Abstract. The microdata underlying tabular frequency data can be represented by the inner frequency table obtained by crossing all the main dimensional variables. This paper will discuss approaches to post-processing the results of tabular SDC techniques to obtain releasable inner frequencies, which may not be whole numbers. As suggested in a recent publication and implemented in the r-package RegSDC, suppressed cells may be replaced by decimal numbers, in a way so that they add correctly up to non-supressed margins. A similar method could be used in cases without cell suppression, but where only some margins are considered releasable. This method (Method A) can be viewed as ordinary regression, modified to handle collinearity, with the inner cells as the unknown parameters and with the releasable margins as response observations. Method A can also be interpreted as linear calibration weighting. An improvement of this approach is to ensure non-negative values by applying non-negative least squares with ridge penalty (Method B). A more natural choice in this case, would be to obtain inner frequencies as fitted values from log-linear modelling. Considering situations where the margins are perturbed in a non-additive way, ordinary log-linear modelling is not possible, but Method B is useful. Log-linear fitting may be performed afterwards to obtain results in in accordance with such modelling assumptions. Another situation is when only some of the releasable margins are perturbed. Then, it is advantageous to modify Method B by incorporating weights (Method C). The perturbed margins are weighted down, so that the resulting inner frequencies add up almost exactly to the non-perturbed margins. The methodology is applicable when the margins are perturbed by Laplace noise to ensure differential privacy. Although, in this case, one may consider least absolute deviation instead of least squares. The generated frequencies serve several purposes: They can be used internally or published directly. Additivity of margins can be restored. One can also use the results to generate public microdata. Then, all frequencies must be whole numbers. A possibility is to round up or down by comparing the decimals to draws from the uniform distribution. This is an alternative to synthetic data drawn as a multinomial sample according to cell probabilities, as in the r-package synthpop. The methodology is tested using example data of various size and complexity. For an efficient implementation, making use of the r-package glmnet seems promising.

1 Introduction

Based on tabular frequency data, synthetic data can be drawn from a fitted log-linear model. As implemented in the r-package synthpop (Nowok et al., 2016), the cell probabilities can be estimated by iterative proportional fitting. Synthetic data is obtained by generating a multinomial sample with these probabilities. Raab (2019) has investigated how this type of synthetic data can be made differential private (Dwork and Roth, 2014). One possibility is to perturb the counts of the inner table, which is obtained by crossing all the main dimensional variables (cover table). According to Raab (2019), a better strategy is to perturb the margins that are sufficient for the chosen model. The estimation involving iterative proportional fitting is carried out after the perturbation. Raab (2019) mentions that the estimation process is tricky and that more work is needed. A problem is that, after perturbation, the data are not additive. The present paper suggests alternative ways of performing the estimation.

The above perturbation means that random noise form the Laplace distribution is added. Under ϵ-differential privacy, and when perturbing inner cells, noise from the $Lap(1/\epsilon)$ distribution is added. The inner cells are then releasable. Margins can be recalculated from the perturbed inner cells. If instead k margins are perturbed, more noise must be added since each individual contributes to k cells. Then, noise from the $Lap(k/\epsilon)$ distribution is added (straightforward privacy budget). Rinott et al. (2018) have investigated differential privacy for frequency tables in more detail.

An alternative to drawing synthetic data is to consider the fitted cell frequencies (cell probabilities times the overall total) as final output. Such frequencies are not whole numbers, but can still be considered valuable. In particular, any aggregate of interest can be computed. The aim of the present paper is to explore linear modelling techniques for estimating/fitting inner cell frequencies, within the context of perturbed (or suppressed) margins (or cells).

Generally, we let the vector $\boldsymbol{y}$ be consisting of all the elements of the inner frequency table. Furthermore, we let $\boldsymbol{z}$ be a general vector of frequencies which can be computed from $\boldsymbol{y}$ via a dummy matrix $\boldsymbol{X}$:

$$\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{y} \tag{1}$$

Each element of $\boldsymbol{z}$ is either an inner cell or a sum of several inner cells. In the first case this means that the corresponding column of $\boldsymbol{X}$ has only one element that is one (others are zero). We will use this setup for different purposes. For instance, $\boldsymbol{z}$ can be selected margins that are considered releasable. Furthermore, $\boldsymbol{z}$ can be margins that are subjected to perturbation. With $\boldsymbol{z}$ known and $\boldsymbol{y}$ unknown, $\boldsymbol{y}$ can be estimated or synthetically generated. Another possibility is to let $\boldsymbol{z}$ be all cells that are found to be safe by means of a cell suppression method. Then, as described in Langsrud (2019a), suppressed cells may be replaced by decimal numbers.

Section 2 outlines how estimation can be performed using linear modelling tools. Thereafter Section 3 illustrates various methods by considering a small example dataset. Section 4 includes a larger dataset and differential privacy is studied further. Section 5 concludes with some final remarks.

2 Linear modelling estimation

To estimate $\boldsymbol{y}$ from $\boldsymbol{z}$ one possibility is

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{X}^T\right)^\dagger \boldsymbol{z} \tag{2}$$

where † denotes a Moore-Penrose generalized inverse. This method can be viewed as ordinary least squares regression, modified to handle collinearity, with $\boldsymbol{y}$ as the unknown parameters and with $\boldsymbol{z}$ as the dependent observations ($\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{y} + \mathsf{error}$). Such regression estimation is reasonable when $\boldsymbol{z}$ is a perturbed version of $\boldsymbol{X}^T\boldsymbol{y}$ and when the added noise is the only stochastic element. When the linear least squares system has several solutions (collinearity), the generalized inverse leads to the solution with minimal length (L2 norm) of the parameter vector. When (1) holds (no perturbation), equation (2) can equivalently be expressed as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{X}^\dagger\boldsymbol{y} \tag{3}$$

This time, $\hat{\boldsymbol{y}}$ can be interpreted as fitted values from ordinary regression (modified to handle collinearity) with $\boldsymbol{y}$ as dependent observations ($\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \mathsf{error}$). That is, the underlying assumption is that $\boldsymbol{y}$ is stochastically generated through the regression model. Note that, when $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, the generalized inverse can be expressed as $\boldsymbol{X}^\dagger = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. The standard regression equation is recognised when plugging this into (2) and (3).

In this paper we will consider both these approaches. The methodology can be modified to ensure non-negative estimates. Then, the use of a generalized inverse may be replaced by incorporating ridge penalty (Hoerl and Kennard, 1970) in the estimation process. Note that, with Laplace distributed errors, maximum likelihood estimates are obtained by least absolute deviation regression. With $\boldsymbol{y}$ as dependent counts, log-linear modelling can be done instead of ordinary regression. The actual model assumptions may be that $\boldsymbol{y}$ is stochastically generated through a log-linear model involving one x-matrix and that noise (Laplace) is added after applying another x-matrix. In this paper we will not try to obtain theoretical/optimal estimates under these assumptions. Instead, we will focus on how estimates can be obtained in practise using available tools. As shown below, the estimation can be done in two steps: first by least squares and thereafter by log-linear modelling. Another example of a practical solution is to weight up exact

	col1	col2	col3	Total
row1	3	6	2	11
row2	1	4	7	12
row3	5	8	27	40
Total	9	18	36	63

Table 1: The example frequency table

margins rather than introduce restrictions. It may be important to keep the x-matrix sparse (mostly zeros) and therefore avoid pre-processing this matrix. Below we describe briefly how the estimation process is carried out using various r packages.

The dummy matrices were generated as sparse matrices and functionality in the package SSBtools (Langsrud and Mevik, 2019) was utilized. Moore-Penrose generalized inverse can be obtained by the function ginv in the standard package MASS. The generalized inverse solution can be approximated by ridge regression (Hoerl and Kennard, 1970) using a very small shrinkage parameter. With very large datasets, the computations can be done efficiently by the package glmnet (Friedman et al., 2010). Then, non-negative estimation is also possible. The data in Section 4 was handled this way. With smaller datasets, more accurate non-negative estimates can be achieved using the package nnls (Mullen and van Stokkum, 2012). Then a trick is used to incorporate ridge shrinkage. The x-matrix was vertically bounded with the identity matrix multiplied by a small shrinkage parameter. The corresponding dependent observations were set to zero. Computations in the next section were done in this manner. When $\boldsymbol{z}$ consists of margins, $\boldsymbol{y}$ may be re-estimated from $\boldsymbol{X}^T\hat{\boldsymbol{y}}$ using iterative proportional fitting. An alternative is to apply glm using the poisson family with estimated $\hat{\boldsymbol{y}}$ as input and with a model formula corresponding to $\boldsymbol{X}$. Below, fits according to log-linear modelling were obtained this way. Least absolute deviation regression is implemented in the package L1pack (Osorio and Wolodzko, 2017). However, we will not apply this package below, since collinearity, non-negative estimation and very large datasets are not handled.

3 Exemplification of various methods

We consider the example data given in Table 1. The vector, $\boldsymbol{y}$, of inner cells has nine elements in this case.

When frequencies below four are sensitive, they may be protected by cell suppression. Including secondary suppression this means that four cells need to be

	col1	col2	col3	Total
row1	1.250 (0.633)	6	3.750 (4.367)	11
row2	2.750 (3.367)	4	5.250 (4.633)	12
row3	5	8	27	40
Total	9	18	36	63

Table 2: Suppressed cells replaced by fitted values (generalized inverse). In the parentheses scaled residuals have been added.

	col1	col2	col3	Total
row1	-0.333 (0.000)	2.667 (2.500)	8.667 (8.500)	11
row2	0.000 (0.000)	3.000 (3.000)	9.000 (9.000)	12
row3	9.333 (9.000)	12.333 (12.500)	18.333 (18.500)	40
Total	9	18	36	63

Table 3: Inner cells as fitted values from linear modelling (ordinary regression, generalized inverse or ridge penalized). Corresponding non-negative estimates are included in the parentheses.

suppressed. By letting $\boldsymbol{z}$ be composed of the 12 safe frequencies, we can estimate the suppressed frequencies by equation (2). These estimates are given in Table 2. Residuals that preserve the margins can be added (Langsrud, 2019a), and one reason is to ensure that the suppressed frequencies are decimal numbers. Then, only the inner cells need to be stored. Various totals can be calculated when needed. When the result is a whole number, this means that this cell is not suppressed. Addition of residuals are included in Table 2. The residuals are scaled so that root mean square error is 1.234. The possibility of this type of scaling is included in the r package RegSDC (Langsrud, 2019b).

By considering all inner cells as suppressed, they can be estimated from the totals by equation (2) similarly. The results are in Table 3 (residuals are not added). Ordinary regression interpretation with $\boldsymbol{y}$ as dependent observations is natural in this case. By treating $\boldsymbol{y}$ as parameters instead, non-negative estimates can be obtained. The results are included in Table 3. However, log-linear modelling with $\boldsymbol{y}$ as dependent counts is most natural in this case. Then, negative values will never occur. The results are given in Table 4. Note that the results can also be interpreted as calibrated weights after linear calibration (Table 3) and raking (Table 4) with equal sampling weights. In particular, non-negative estimation corresponds to linear calibration with a lower bound.

	col1	col2	col3	Total
row1	1.571	3.143	6.286	11
row2	1.714	3.429	6.857	12
row3	5.714	11.429	22.857	40
Total	9	18	36	63

Table 4: Inner cells as fitted values from log-linear modelling.

	col1	col2	col3	Total
row1	0.000 (1.443)	1.300 (2.459)	7.967 (5.364)	9.267 [11.000]
row2	0.000 (1.754)	2.300 (2.990)	8.967 (6.522)	11.267 [13.000]
row3	9.933 (6.736)	13.333 (11.484)	20.000 (25.047)	43.267 [45.000]
Total	9.933 [11.000]	16.933 [18.000]	36.933 [38.000]	63.800 [61.000]

Table 5: Inner cells as non-negative estimates (ridge penalized) from perturbed totals (in brackets). Corresponding recomputed (additivity restored) totals are included. Subsequently, the inner cells have been re-estimated by log-linear modelling (in parentheses).

	col1	col2	col3	Total
row1	0.000 (0.106)	6.856 (7.178)	10.056 (9.629)	16.913 [19.757]
row2	0.000 (0.073)	4.249 (4.965)	7.449 (6.660)	11.698 [14.542]
row3	0.415 (0.236)	17.005 (15.968)	20.205 (21.422)	37.625 [40.470]
Total	0.415 [-2.429]	28.110 [25.266]	37.711 [34.867]	66.236

Table 6: Inner cells as non-negative estimates (ridge penalized) from totals with Laplace noise (in brackets). Corresponding recomputed (additivity restored) totals are included. Subsequently, the inner cells have been re-estimated by log-linear modelling (in parentheses).

	col1	col2	col3	Total
row1	0.000 (0.728)	3.200 (3.931)	7.800 (6.341)	11
row2	0.000 (0.794)	3.700 (4.289)	8.300 (6.917)	12
row3	4.168 (2.646)	15.616 (14.296)	20.216 (23.058)	40
Total	4.168 [3.286]	22.516 [21.633]	36.316 [35.433]	63

Table 7: Inner cells as non-negative estimates (ridge penalized) from column totals with Laplace noise (in brackets) and from the exact row totals (highly weighted). Corresponding recomputed (additivity restored) totals are included. Subsequently, the inner cells have been re-estimated by log-linear modelling (in parentheses).

The totals may not be considered safe without perturbation. In Table 5 the seven totals have been perturbed (in brackets) according to a perturbation table generated by the r package ptable (Enderle and Giessing, 2019) with parameter setting $D = 5$, $V = 3$ and $js = 2$. Non-negative estimates of the inner cells were obtained in a way corresponding to a modified/generalized (ridge penalized) variant of equation (2). The totals are recomputed from these estimates. Furthermore, the inner cells have been re-estimated by log-linear modelling. After non-negative estimation, the only nonzero value in column 1 is 9.933 (row 3). The additivity restored total is therefore 9.933. After log-linear modelling the values in column 1 have become 1.443, 1.754, and 6.736. The sum is still 9.933.

In Table 6, Laplace noise is added to the row and column totals ($k = 2$). We use $\epsilon = 0.5$ and thus $Lap(4)$ noise is added. In addition, one could also perturb the overall total similarly, but then $Lap(6)$ had to be used ($k = 3$). Negative values may occur and in Table 6, the column 1 total is negative after perturbation. Apart from the fact that the overall total is not perturbed and not included in $\boldsymbol{z}$, the computations are similar to those for Table 5.

In Table 7, only the column totals are perturbed. It is assumed that the true row totals can be released. Now, $Lap(2)$ noise is added ($k = 1$, $\epsilon = 0.5$). In the estimation, $\boldsymbol{z}$ is composed of the 7 totals (overall included) and three of them are perturbed. The non-perturbed totals are given high weight (1000) and therefore those totals remain almost unchanged.

4 A differential privacy example

To illustrate the methodology one a larger dataset we use public microdata for the German labour force survey for year 2013 published by Eurostat (2018). We consider the five-way frequency table obtained by crossing the variables AGE(7), HOURREAS(19), ISCO1D(12), REFWEEK(52) and SIZEFIRM(6). The number of categories is in parenthesis and missing is treated as a category. The total number of inner cells is 497952, but only 46689 of them have a nonzero count. The overall total is 478173. Below, we will focus on two-way and three-way margins. The total number of two-way cells is 2961 (2727 nonzero) and the number three-way cells is 39262 (24224 nonzero). When $\boldsymbol{z}$ is composed of all three-way cells, this means that $\boldsymbol{X}$ in equation (1) is a 497952×39262 matrix. Utilizing sparse matrices in R and the package glmnet, the computations described above are manageable. Although, the re-estimation by log-linear modelling was not possible. Below we will consider nine ways of perturbing/releasing data. Whenever inner cells are non-negatively estimated, the two-way and three-way margins will be recomputed from these estimates. The nine ways are:

inner cells: Laplace noise is added to the inner cells and all the margins are recomputed from these inner cells. When $\epsilon = 0.5$, $Lap(2)$ noise is added.

3-way (no fitting): Laplace noise is added to the three-way margins and those three-way margins are released. We have a total of ten three-way margins. Therefore, when $\epsilon = 0.5$, $Lap(20)$ noise is added.

3-way (zero replace): As above, but negative values are replaced by zeros.

3-way: As above, Laplace noise is added to the three-way margins. The inner cells are non-negatively estimated from the perturbed three-way margins. The two-way and three-way margins are recomputed from the non-negative estimates.

3-way (2-way exact): Laplace noise is added to the three-way margins. The inner cells are non-negatively estimated from the perturbed three-way margins and upweighted (weight is 1000) non-perturbed two-way margins.

nothing (2-way exact): The inner cells are non-negatively estimated from non-perturbed two-way margins.

2-way: Laplace noise noise is added to the two-way margins. We have a total of ten two-way margins. Therefore, when $\epsilon = 0.5$, $Lap(20)$ noise is added. The inner cells are non-negatively estimated from the perturbed two-way margins. Thereafter, the margins are recomputed.

2-way (no fitting): As above, Laplace noise is added to the two-way margins and those two-way margins are released.

2-way (zero replace): As above, but negative values are replaced by zeros.

Table 8 illustrates "3-way (2-way exact)" by showing five inner cells, five two-way cells and five three-way cells. The recomputed (fitted) two-way cells are almost identical to the original ones. In Table 8 we see that the two inner cell zeros are kept after estimation. Looking at all the data, 443880 of the 451263 inner cell zeros are kept.

Structural zeros, if any, should not be perturbed. How to treat nonstructural zeros in general, is an important discussion. When applying perturbation tables, it is common to never perturbe zeros. However, Rinott et al. (2018) conclude that nonstructural zeros should be perturbed. Within the above regression setup, choosing to not perturb zeros can be implemented very easily. We simply omit the rows of $\boldsymbol{X}$ and $\boldsymbol{y}$ corresponding to zeros. Afterwards, columns of $\boldsymbol{X}$ with only zeros can also be omitted so that the zeros of $\boldsymbol{z}$ are omitted. Below, we will look at the data without zeros in this way. Another possibility is to only remove the zeros in $\boldsymbol{y}$ corresponding to the zeros in $\boldsymbol{z}$.

We make use of a modified version of the Hellinger distance so that negative values are allowed:

AGE	HOUR-REAS	ISCO1D	REF-WEEK	SIZE-FIRM	Original	Perturbed	Fitted
20	97	300	24	12	11		6.10
32	97	0	40	12	2		5.58
65	99	500	11	13	3		18.54
20	3	100	35	99	0		0.00
32	13	200	15	99	0		0.00
	15		13		92		92.01
7			9		1158		1157.97
			51	10	320		319.97
	5		31		0		0.00
75	6				0		0.05
20	10		7		2	32.10	0.00
	97	500	43		356	325.62	347.09
47	8		8		30	20.37	32.92
	5	0	3		0	-32.96	0.00
75	14		1		0	-7.29	0.00

Table 8: Example cells when the perturbation type is "3-way (2-way exact)" under 0.5-differential privacy using data with zeros.

$$\mathrm{HD}(f,g) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{f_i} - \mathrm{sign}(g_i)\sqrt{|g_i|}\right)^2} \tag{4}$$

Here, f is a vector of n original counts (elements in $\boldsymbol{z}$) and g is the corresponding vector of perturbed counts. According to Shlomo et al. (2015) we calculate the measure of utility as

$$\mathrm{Utility}(f,g) = 1 - \mathrm{HD}(f,g)/\sqrt{\textstyle\sum_{i=1}^{n} f_i} \tag{5}$$

The utility measure is made to be bounded between 0 and 1. However, we have allowed negative values, and the Laplace noise can lead to negative utility. The utility values from our study is given in Table 9 ($\epsilon = 0.5$) and Table 10 ($\epsilon = 1.5$). In addition, in Table 11 ($\epsilon = 0.5$) and Table 12 ($\epsilon = 1.5$), we present the mean absolute deviation: $\sum_{i=1}^{n} |f_i - g_i|/n$.

These tables show that the fitting process leads to improvements. For example, in Table 9 with zeros, three-way without fitting results in 0.744 as utility. After fitting, the utility has increased to 0.870. As seen, this increase cannot be obtained by only removing zeros. We have this phenomenon in all four tables for data with

Perturbation type	Data with zeros			Data without zeros		
	inner	2-way	3-way	inner	2-way	3-way
inner cells	-0.008	0.898	0.731	0.733	0.981	0.932
3-way (no fitting)			0.744			0.817
3-way (zero replace)			0.832			0.887
3-way	0.659	0.928	0.870	0.690	0.978	0.914
3-way (2-way exact)	0.691	0.999	0.920	0.716	1.000	0.926
nothing (2-way exact)	0.746	1.000	0.926	0.796	1.000	0.934
2-way	0.742	0.974	0.916	0.785	0.977	0.923
2-way (no fitting)		0.955			0.962	
2-way (zero replace)		0.971			0.975	

Table 9: Utility (Hellinger) under 0.5-differential privacy calculated from inner cells and two-way and three-way margins.

Perturbation type	Data with zeros			Data without zeros		
	inner	2-way	3-way	inner	2-way	3-way
inner cells	0.428	0.947	0.857	0.889	0.994	0.978
3-way (no fitting)			0.863			0.909
3-way (zero replace)			0.908			0.942
3-way	0.755	0.963	0.927	0.800	0.990	0.952
3-way (2-way exact)	0.745	0.999	0.952	0.793	1.000	0.959
nothing (2-way exact)	0.746	1.000	0.926	0.796	1.000	0.934
2-way	0.745	0.986	0.924	0.794	0.988	0.931
2-way (no fitting)		0.977			0.982	
2-way (zero replace)		0.985			0.988	

Table 10: Utility (Hellinger) under 1.5-differential privacy calculated from inner cells and two-way and three-way margins.

	Data with zeros			Data without zeros		
Perturbation type	inner	2-way	3-way	inner	2-way	3-way
inner cells	2.00	82.21	23.13	1.98	23.11	7.86
3-way (no fitting)			19.92			19.98
3-way (zero replace)			13.36			15.65
3-way	0.35	55.68	9.95	3.64	26.06	11.84
3-way (2-way exact)	0.32	0.02	6.38	3.28	0.02	9.99
nothing (2-way exact)	0.29	0.00	9.15	2.72	0.00	13.65
2-way	0.29	16.39	10.04	2.85	17.13	15.44
2-way (no fitting)		19.81			19.83	
2-way (zero replace)		17.45			18.37	

Table 11: Mean absolute deviation under 0.5-differential privacy calculated from inner cells and two-way and three-way margins.

	Data with zeros			Data without zeros		
Perturbation type	inner	2-way	3-way	inner	2-way	3-way
inner cells	0.67	27.40	7.71	0.66	7.70	2.62
3-way (no fitting)			6.64			6.66
3-way (zero replace)			4.80			5.75
3-way	0.22	17.43	3.85	2.09	8.85	4.82
3-way (2-way exact)	0.24	0.01	3.03	2.31	0.01	4.71
nothing (2-way exact)	0.29	0.00	9.15	2.72	0.00	13.65
2-way	0.29	5.79	9.36	2.75	6.04	14.07
2-way (no fitting)		6.60			6.61	
2-way (zero replace)		6.03			6.32	

Table 12: Mean absolute deviation under 1.5-differential privacy calculated from inner cells and two-way and three-way margins.

Perturbation type	Data with zeros $\epsilon = 0.5$	Data with zeros $\epsilon = 1.5$	Data without zeros $\epsilon = 0.5$	Data without zeros $\epsilon = 1.5$
3-way	2.97%	0.84%	0.68%	0.13%
3-way (2-way exact)	0.00%	0.00%	0.00%	0.00%
2-way	0.06%	0.01%	0.04%	0.01%

Table 13: The positive bias of the overall total caused by non-negative estimation.

	inner			2-way			3-way		
Weight	10^{-7}	10^{-10}	10^{-14}	10^{-7}	10^{-10}	10^{-14}	10^{-7}	10^{-10}	10^{-14}
0	2.25	2.14	2.09	8.84	8.84	8.85	4.81	4.81	4.82
1	2.26	2.13	2.06	3.72	3.03	3.02	4.58	4.46	4.47
3	2.40	2.22	2.05	3.52	1.48	1.45	5.06	4.44	4.39
10	2.74	2.33	2.05	3.62	0.67	0.55	6.26	4.77	4.35
100	3.45	2.76	2.06	4.20	0.53	0.07	11.79	8.42	4.30
1000	3.51	2.80	2.31	4.50	0.37	0.01	15.28	12.60	4.71
10000	3.58	2.93	2.86	4.59	0.38	0.00	16.31	14.68	9.21
∞	3.50	2.87	2.72	1.89	0.20	0.00	16.29	14.45	13.65

Table 14: Effect of varying the convergence threshold (10^{-7}, 10^{-10} and 10^{-14}) and the weight of exact two-way margins: Mean absolute deviation under 1.5-differential privacy calculated from data without zeros and when the perturbation type is "3-way (2-way exact)".

and without zeros and when looking at the two-way and the three-way perturbation types. Perturbing the inner cells straightforwardly tends to be the best method when considering the data without zeros. With zeros, it is no doubt that perturbing margins is better. In Table 9, it is interesting that the three-way utility increases when going from three-way to two-way perturbation. We do not have this phenomenon in the other tables. When making use of highly weighted exact two-way margins, these margins are almost preserved. The results are also improved in general. As shown in Table 13, bias problems caused by non-negative estimation are eliminated. However, to correct the bias, it would be sufficient to include the overall total rather than ten two-way margins.

Note that, in this case, the numerical calculations are challenging due to the heavily weighted margins. Above, the convergence threshold parameter in glmnet, whose default value is 10^{-7}, was set to 10^{-14}. The effect of changing this parameter is illustrated in Table 14. The weights, 0 and ∞, represents "3-way" and "nothing

(2-way exact)". Mean absolute deviation, $\epsilon = 1.5$ and data without zeros were chosen because the changes are most visible in this case. The high precision was needed in this example. Using 1000 as weight gave good results, but a higher weight must be avoided.

5 Concluding remarks

The fitted inner frequencies may be published directly as a sort of perturbed counts. They are not whole numbers and they do not follow a realistic distribution. However, this is also the case for ordinary Laplace perturbed counts. A nice property is that margins can be calculated when needed and non-additivity is not a problem. An alternative is to use the inner frequencies internally to calculate output on the fly. Hierarchical tables fit well within the general setup (1). Although, this has not been studied above.

Synthetic data can be drawn as a multinomial sample with cell probabilities calculated from the fitted inner frequencies (divide by overall total). Then, re-estimation by log linear modelling is preferable. If distributional properties are not important, public microdata can be generated in other ways to improve utility. When inner frequencies are whole numbers, this is a representation of a microdataset. A subdataset can therefore be obtained by removing the decimals. The remaining part of the microdata can be generated by comparing the decimals to draws from the uniform distribution. To preserve the overall total, this can be modified by multinomial sampling. One can imagine further improvements, that try to preserve margins. An even better strategy would be to introduce restrictions so that all estimates become whole numbers. However, this is not possible using the linear modelling techniques studied in this paper. Although, using package glmnet it is possible to limit the number of nonzero elements in $\hat{\boldsymbol{y}}$. Using glmnet, standard gaussian error was assumed, even if Laplace noise was added. If it is possible to take the Laplace distribution into account, this means that it is possible to improve the results.

The methodology studied here is closely related to what is called database reconstruction (Abowd, 2018). Then, $\boldsymbol{z}$ represents the published statistics and $\boldsymbol{y}$ represents the database. The database reconstruction theorem (Dinur and Nissim, 2003) is formulated by assuming bounded noise and reconstruction by a linear-programming algorithm. The above estimation technique is another way of performing the reconstruction. Some elements of $\hat{\boldsymbol{y}}$ can be perfectly fitted when this information is hidden in $\boldsymbol{z}$. Cell suppression is often done to avoid perfect fit or perfect reconstruction of small frequencies. The methodology in the present paper is also closely related to the matrix mechanism (Li et al., 2015) and some of

the work here can be viewed as a special application of it. Using the matrix mechanism to protect statistics according to a particular x-matrix (1), the safe intermediate y-vector (2) is first created by using another x-matrix. The challenge is to optimally select the latter x-matrix. Li et al. (2015) establish a theoretical framework using the Moore-Penrose generalized inverse. In addition, the problem of using the non-negativity constraint is discussed in detail.

Section 4 illustrates that the linear estimation techniques can be useful. Although an implementation provides inaccurate results, privacy is not violated because the input is protected. In practise, several choices must be made. With a lot of zeros, the question of perturbing/estimating (all or some of) the zeros is extremely important. Choosing margins to be perturbed and possible margins to be preserved are other important questions. If some margins are to be preserved, one may still do something to protect small frequencies. A model (set of margins) could be built in a traditional manner or one could try to optimize a measure of utility. The illustrations above were made simple by selecting all two-way and/or three-way margins. Last but not least, under differential privacy, a value of ϵ must be chosen.

References

Abowd, J. M. (2018). Staring Down the Database Reconstruction Theorem. Joint Statistical Meetings, Vancouver, BC, Canada July 30, 2018.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA. ACM.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Enderle, T. and Giessing, S. (2019). *ptable: Generation of perturbation tables.* R package on github.com/tenderle/ptable.

Eurostat (2018). EU Labour Force Survey Database User Guide, version April 2018. Technical report, European Commission, Eurostat, Directorate F, Unit F-3.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Langsrud, Ø. (2019a). Information preserving regression-based tools for statistical disclosure control. *Statistics and Computing*, 29(5):965–976.

Langsrud, Ø. (2019b). *RegSDC: Information Preserving Regression-Based Tools for Statistical Disclosure Control*. R package on CRAN.

Langsrud, Ø. and Mevik, B.-H. (2019). *SSBtools: Statistics Norway's Miscellaneous Tools*. R package on CRAN.

Li, C., Miklau, G., Hay, M., Mcgregor, A., and Rastogi, V. (2015). The matrix mechanism: Optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24(6):757–781.

Mullen, K. M. and van Stokkum, I. H. M. (2012). *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*. R package on CRAN.

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74(11):1–26.

Osorio, F. and Wolodzko, T. (2017). *L1pack: Routines for L1 estimation*. R package on CRAN.

Raab, G. (2019). Practical Experience with Making Synthetic Data Differentially Private. Simons Workshop on "Data Privacy: From Foundations to Applications", March 6th, 2019, Berkeley, USA.

Rinott, Y., O'Keefe, C. M., Shlomo, N., and Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statist. Sci.*, 33(3):358–385.

Shlomo, N., Antal, L., and Elliot, M. (2015). Measuring disclosure risk and data utility for flexible table generators. *Journal of Official Statistics*, 31(2):305–324.