PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
xxxxx, xxxxx, xxxx

Keywords:
contingency tables, integer matrices, sequential importance sampling

Author for correspondence:
Mark Newman
e-mail: mejn@umich.edu

Improved estimates for the number of non-negative integer matrices with given row and column sums

Maximilian Jerdee,[1] Alec Kirkley[2] and M. E. J. Newman[1,3]

[1]Department of Physics, University of Michigan, Ann Arbor, MI 48109. U.S.A.
[2]Institute of Data Science, University of Hong Kong, Hong Kong
[3]Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109. U.S.A.

The number of non-negative integer matrices with given row and column sums features in a variety of problems in mathematics and statistics but no closed-form expression for it is known, so we rely on approximations. In this paper, we describe a new such approximation, motivated by consideration of the statistics of matrices with non-integer numbers of columns. This estimate can be evaluated in time linear in the size of the matrix and returns results of accuracy as good as or better than existing linear-time approximations across a wide range of settings. We show that the estimate is asymptotically exact in the regime of sparse tables, while empirically performing at least as well as other linear-time estimates in the regime of dense tables. We also use the new estimate as the starting point for an improved numerical method for either counting or sampling matrices with given margins using sequential importance sampling. Code implementing our methods is available.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

Matrices with non-negative integer elements and prescribed row and column sums arise in a range of statistical, physical, and mathematical contexts. For example, they appear in statistics and information theory as contingency tables, whose elements count the number of times a state or event A occurred, contingent on the occurrence of another state or event B.

An important but difficult problem is to compute the number of matrices for given values of the row and column sums, i.e., the number $\Omega(\mathbf{r}, \mathbf{c})$ of $m \times n$ non-negative integer matrices whose rows sum to $\mathbf{r} = (r_1, \ldots, r_m)$ and whose columns sum to $\mathbf{c} = (c_1, \ldots, c_n)$. This number plays a key role for instance in the calculation of mutual information measures for classification and community detection [1] and in sequential importance sampling methods for integer matrices [2, 3, 4]. See Ref. [5] for a review.

No exact expression is known for $\Omega(\mathbf{r}, \mathbf{c})$ for general $\mathbf{r}$ and $\mathbf{c}$, and its numerical computation is #P-hard [6], meaning it is improbable that an algorithm exists with run time polynomial in m and n for general m, n. Workable exact algorithms do exist for small m, n [7] and for cases with bounded row or column sums [8], but outside of these settings the only tractable approach is approximation. In this paper we review approximation methods for this problem and present a new, computationally efficient approximation that is simple to implement and compares favorably with previous approaches in terms of both accuracy and running time.

Previous approximation methods for this problem fall into three broad classes, which we will refer to as *linear-time*, *maximum-entropy*, and *sampling-based* methods. The majority fall into the first category, the linear-time methods, which are characterized by their rapid $\mathrm{O}(m + n)$ computation time, although they typically achieve this efficiency at the expense of accuracy and scope. The linear-time approaches include methods based on combinatoric arguments [9, 10] and moment-matching arguments [11, 12], and methods tailored specifically to the sparse regime [13, 14] in which most elements of the matrix are zero. The method we propose also falls into the linear-time category and consistently performs near the top of this class across a wide array of test cases. We show that it is asymptotically exact in the regime of sparse tables with bounded row and column sums, a property shared by other approximations specifically geared towards this regime. Unlike other sparse estimates, however, our new estimate also performs well in the dense regime, where the typical table entry diverges while the table shape remains constant. Indeed, in the dense regime the new estimate is asymptotically equal to the previous estimate of Diaconis and Efron [11], which has seen use in practical applications and has been observed to work well for dense cases [5].

The second class of methods are maximum-entropy methods, developed in this context by Barvinok and Hartigan [15]. For large m and n these methods outperform the linear-time methods in terms of accuracy outside of the sparse regime but are much slower, requiring the numerical solution of a continuous convex optimization problem followed by evaluation of an $(m + n - 1) \times (m + n - 1)$ matrix determinant, for a time complexity of about $\mathrm{O}((m + n)^3)$. The basic method employs a Gaussian maximum-entropy approximation but the result can be further refined using an "Edgeworth correction," which requires an additional $\mathrm{O}(m^2 n^2)$ computation but substantially improves accuracy.

The third class of approximations are sampling-based methods, including Markov-chain Monte Carlo (MCMC) methods [16] and sequential importance sampling (SIS) [2]. Given sufficient running time these methods will converge to the true answer, although the time taken can be prohibitive. SIS is typically better than MCMC for calculating $\Omega(\mathbf{r}, \mathbf{c})$ in terms of both speed and accuracy [2, 3] and we make use of the SIS method in this paper to establish benchmarks for the evaluation of the other methods. As a bonus, the new linear-time approximation we propose can also be used to improve the convergence of SIS, allowing us to apply the latter method to substantially larger matrices than has previously been possible. SIS also has advantages over other methods in certain parameter regimes. Specifically, it is known that $\Omega(\mathbf{r}, \mathbf{c})$ is non-analytic at certain phase transition points [17, 18] and this behavior cannot be reproduced by, for instance,

the linear-time estimates (including our own), which are all smooth, but the SIS approach should converge to the true answer regardless and so may be used even in the vicinity of such points.

In addition to the specific problem of counting integer matrices with given row and column sums, a number of related problems have received attention. The problem of sampling such matrices uniformly arises in a variety of contexts [2, 4, 11] and can be tackled efficiently by the same modification of the SIS algorithm we propose in Section 5. Separately, the problems of both counting and sampling matrices whose elements take the values 0 and 1 have seen interest [3]. Although these questions are not our main concern here, our methods can be extended to cover these cases also (with some caveats) and we compare the results with a variety of competing methods in Appendix A. Finally, there are certain special cases of the matrix counting problem, such as cases where the row and column sums are uniform (so-called magic squares), for which one can make progress beyond what is possible in the general case [19, 20]. We will not discuss these cases here however: our focus in this paper is on the general case.

2. Summary of results

We present a number of new results in this paper. First, we derive a new and simple linear-time approximation for the number of non-negative integer matrices with given row and column sums $\mathbf{r}, \mathbf{c}$ thus:

$$\Omega(\mathbf{r},\mathbf{c}) \simeq \binom{N+m\alpha_{\mathbf{c}}-1}{m\alpha_{\mathbf{c}}-1}^{-1} \prod_{i=1}^{m} \binom{r_i+\alpha_{\mathbf{c}}-1}{\alpha_{\mathbf{c}}-1} \prod_{j=1}^{n} \binom{c_j+m-1}{m-1}, \tag{2.1}$$

where

$$\alpha_{\mathbf{c}} = \frac{N^2-N+(N^2-c^2)/m}{c^2-N}, \qquad N=\sum_i r_i = \sum_j c_j, \qquad c^2 = \sum_{j=1}^{n} c_j^2. \tag{2.2}$$

(There are certain trivial cases where these expressions give invalid values for $\alpha_{\mathbf{c}}$ but these are easily dealt with—see Appendix B. For optimal precision we also recommend choosing the rows and columns such that $m \le n$, interchanging $\mathbf{r}$ and $\mathbf{c}$ if necessary to achieve this.)

Second, we have conducted exhaustive tests of this estimate and five other previously published linear-time estimates, comparing them with ground-truth results derived from converged sequential importance sampling. Figure 1 summarizes the results of these tests. We find that most of the differences in performance between estimates can be seen by considering square $m \times m$ matrices of various sizes while varying the sum N of all entries. (Some results from other tests, including tests on non-square matrices, are given in Appendix C, but tell essentially the same story.) In our calculations we generate ten random test cases for each parameter pair (N, m) with margins $\mathbf{r}$ and $\mathbf{c}$ drawn uniformly from the set of m-element positive integer vectors that sum to N. We then perform a lengthy run of sequential importance sampling (SIS) on each sampled test case to establish a ground-truth estimate of the number of matrices. Armed with these SIS estimates, we apply each of the six linear-time estimators to the same test cases and compute the error on each one. We report performance in terms of the fractional error in $\log \Omega(\mathbf{r},\mathbf{c})$, since the logarithm is simpler to deal with numerically and is also often the quantity of most interest [1].

The first panel of Fig. 1, labeled "EC" (for "effective columns"—see Section 3(a)), shows the results for our new estimator, Eq. (2.1). The running time for all of the linear-time estimators is negligible, but as the figure shows their accuracy varies. In particular, we distinguish a sparse regime where $N \ll mn$ so that most matrix elements are zero (up and to the left in the plots) and a dense regime where $N \gg mn$ so that most matrix elements are nonzero (down and to the right). Some estimates, such as those labeled BBK and GMK, perform well in the sparse regime but poorly in the dense regime. Others, such as DE, do the reverse. The EC estimate of this paper, however, is comparable to or better than the others in both the sparse and dense regimes, while still being fast and simple to compute. In the dense regime the fractional error is around 10^{-2}

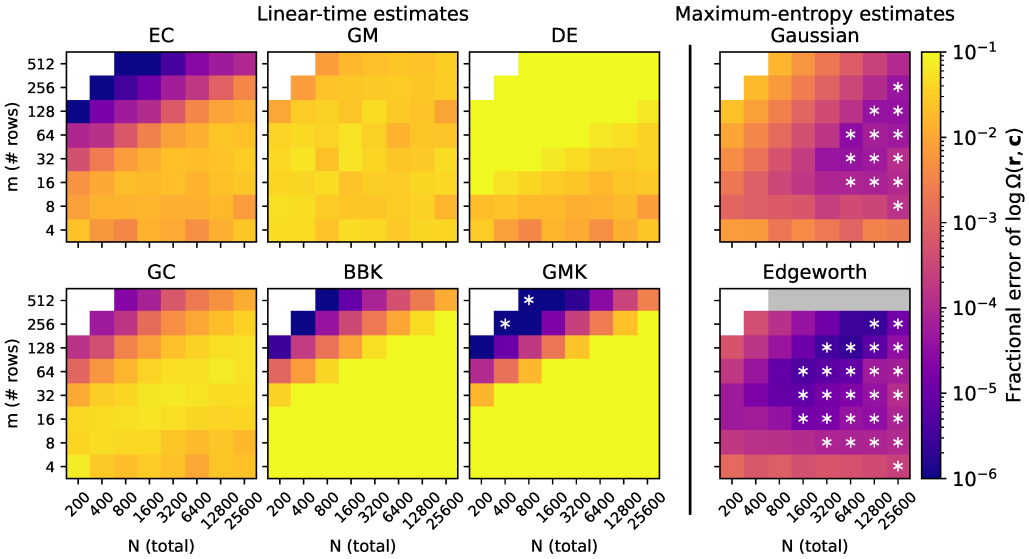

Figure 1. Fractional error in various estimates of $\log \Omega(\mathbf{r}, \mathbf{c})$ for square $m \times m$ matrices with total sum N, relative to ground-truth results computed by sequential importance sampling (see Section 5). Each square represents an average over ten sets of margins $\mathbf{r}, \mathbf{c}$ drawn uniformly at random. Asterisks denote data points for which the error is within five times the estimated error from the sequential importance sampling, and so the true error may be smaller in these cases. White regions indicate invalid parameter combinations where $m > N$ so that margins $\mathbf{r}, \mathbf{c}$ can not be generated without zeros. Gray regions indicate parameter values for which estimates could not be computed in an hour of run time or less. See Appendix C for further details of the benchmarking process.

or 10^{-3}, becoming as good as 10^{-6} in the sparse regime. These numerical results agree with our analytic findings: in the sparse regime our EC estimate is equal to the BBK estimate and asymptotically exact, as shown in Appendix B. In the dense regime, our EC estimate is equal to the DE estimate and matches its error, as also shown in Appendix B. The estimate denoted GC also gives acceptable performance in both sparse and dense regimes, but performs roughly an order of magnitude worse than the EC estimate in our tests.

We have also performed tests using the two maximum-entropy estimates of [15] for a portion of the same test cases and the results are also shown in Fig. 1. As the figure shows, these estimates generally outperform the linear-time estimates, including our own, outside of the sparse regime, but they do so at the expense of much greater computational effort. As mentioned in the introduction these estimates have time complexities of $\mathrm{O}((m+n)^3)$ for the Gaussian approximation and $\mathrm{O}(m^2 n^2)$ for the Edgeworth version. For a typical case with $m = 128$ and $N = 3200$, our implementations of the linear-time estimates run in under 2 ms each (on commodity hardware, *circa* 2022), the Gaussian maximum-entropy method takes 3 seconds, and the Edgeworth-corrected version takes 22 seconds. For our largest test cases with $m = 512$ the calculation of the Edgeworth correction becomes so demanding as to be impractical, so results for these cases are omitted from Fig. 1.

Apart from their substantial computational demands, the maximum-entropy methods work very well in the regime of intermediate-to-high density and indeed do so well in this region that their accuracy becomes comparable to the accuracy of the sequential importance sampling that we use to compute the ground truth. The SIS calculation, like all sampling methods, displays some statistical error, as shown in Fig. 2. Although this error is usually negligible, it is a limiting factor for evaluating the maximum-entropy estimates in some cases. These cases are denoted by asterisks in Fig. 1.

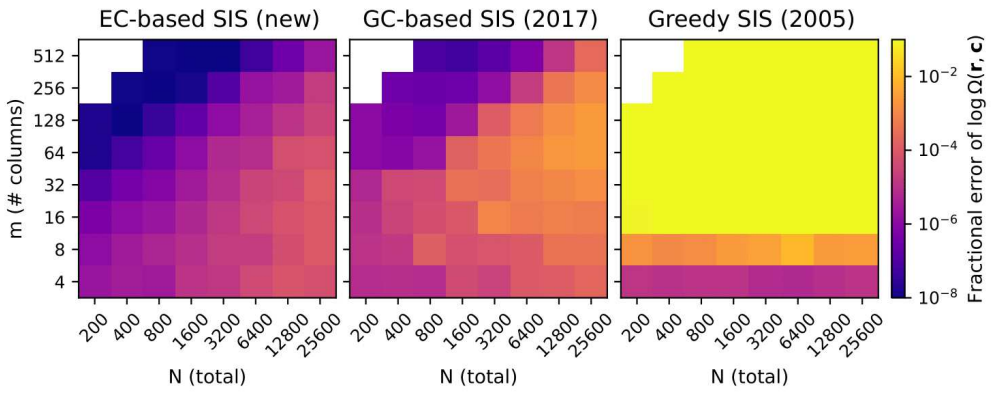

Figure 2. Estimated fractional error of sequential importance sampling estimates of $\log \Omega(\mathbf{r}, \mathbf{c})$ for three different variants of the SIS approach, with a fixed amount of computer time (one hour) spent on each choice of parameter values. The EC-based SIS, which serves as the benchmark for the results in Fig. 1, improves upon existing SIS methods by roughly two orders of magnitude, as discussed in Section 5. Note that the color scale in this figure differs from that in Fig. 1.

Software implementations of the various estimates and SIS methods described in this paper can be found at `https://github.com/maxjerdee/contingency_count`.

3. Linear-time estimates

Turning now to the details, in this section we discuss the linear-time methods for estimating $\Omega(\mathbf{r}, \mathbf{c})$. We first present our new estimate, which is based on the concept of "effective columns." We also describe three related approaches due to Gail and Mantel [12], Diaconis and Efron [11], and Good and Crook [9, 10], and two somewhat different approaches tuned to the sparse case and due to Békéssy, Békéssy, and Komlós [13] and Greenhill and McKay [14].

(a) A new estimate for matrix counts

In this section we derive the approximation for $\Omega(\mathbf{r}, \mathbf{c})$ given in Eq. (2.1). Let $A(\mathbf{c})$ be the set of all non-negative $m \times n$ integer matrices $X = (x_{ij})$ with fixed column sums $\mathbf{c}$ but unconstrained row sums:

$$A(\mathbf{c}) = \left\{ \{x_{ij}\} \in \mathbb{N}^{m \times n} \middle| \sum_i x_{ij} = c_j, j = 1, \ldots, n \right\}. \tag{3.1}$$

By standard arguments the number of ways to choose the entries of column j so that they sum to c_j is $\binom{c_j+m-1}{m-1}$ and the columns are independent so the number of matrices in the set $A(\mathbf{c})$ is

$$|A(\mathbf{c})| = \prod_{j=1}^{n} \binom{c_j + m - 1}{m - 1}. \tag{3.2}$$

Now we further restrict to the subset of matrices $A(\mathbf{r}, \mathbf{c}) \subseteq A(\mathbf{c})$ with both row and column sums fixed:

$$A(\mathbf{r}, \mathbf{c}) = \left\{ \{x_{ij}\} \in \mathbb{N}^{m \times n} \middle| \sum_{i=1}^{m} x_{ij} = c_j, \sum_{j=1}^{n} x_{ij} = r_i \right\}. \tag{3.3}$$

Our quantity of interest is the size of this set $\Omega(\mathbf{r},\mathbf{c}) = |A(\mathbf{r},\mathbf{c})|$. Under a uniform distribution over $A(\mathbf{c})$, the conditional probability $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ of observing a particular row sum $\mathbf{r}$ is

$$\mathbf{Pr}(\mathbf{r}|\mathbf{c}) = \frac{|A(\mathbf{r},\mathbf{c})|}{|A(\mathbf{c})|} = \frac{\Omega(\mathbf{r},\mathbf{c})}{|A(\mathbf{c})|}. \tag{3.4}$$

Since we have an exact expression for $|A(\mathbf{c})|$ in Eq. (3.2), the problem of calculating $\Omega(\mathbf{r},\mathbf{c}) = |A(\mathbf{c})|\,\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ is thus reduced to one of finding $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$. This is still a difficult problem and requires making an approximation. Inspired by work of Gail and Mantel [12] we take a variational approach and propose a family of candidate approximant distributions for $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$, then choose the best member of this family using a moment-matching argument.

Motivated by Diaconis and Efron [11], our family of candidate distributions is based on the form of the unconditional distribution on the row sums $\mathbf{r}$ without any constraint on the column sums. By the same argument that led to Eq. (3.2), the number of matrices with row sums $\mathbf{r}$ is

$$|A(\mathbf{r})| = \prod_{i=1}^{m} \binom{r_i + n - 1}{n - 1}. \tag{3.5}$$

At the same time the set A of all non-negative $m \times n$ integer matrices that sum to N has size

$$|A| = \binom{N + mn - 1}{mn - 1}, \tag{3.6}$$

and hence, under a uniform distribution over A, the probability of observing row sum $\mathbf{r}$ is

$$\mathbf{Pr}(\mathbf{r}) = \frac{|A(\mathbf{r})|}{|A|} = \binom{N + mn - 1}{mn - 1}^{-1} \prod_{i=1}^{m} \binom{r_i + n - 1}{n - 1}. \tag{3.7}$$

The key to our argument is to approximate $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ by this unconditional distribution, but with the number of columns n replaced with a free parameter $\alpha_{\mathbf{c}}$, which we call the number of *effective columns*:

$$\mathbf{Pr}(\mathbf{r}|\mathbf{c}) \simeq \mathbf{Pr}(\mathbf{r}|\alpha_{\mathbf{c}}) = \binom{N + m\alpha_{\mathbf{c}} - 1}{m\alpha_{\mathbf{c}} - 1}^{-1} \prod_{i=1}^{m} \binom{r_i + \alpha_{\mathbf{c}} - 1}{\alpha_{\mathbf{c}} - 1}. \tag{3.8}$$

The resulting distribution over $\mathbf{r}$ is the one that would be observed under the uniform distribution over $m \times \alpha_{\mathbf{c}}$ matrices whose elements sum to N.

Now we relax the constraint that $\alpha_{\mathbf{c}}$ be an integer, defining the obvious generalization of the binomial coefficient

$$\binom{n}{k} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k)}. \tag{3.9}$$

We are merely using Eq. (3.8) as a trial distribution for our variational approximation, so the physical interpretation of $\alpha_{\mathbf{c}}$ as an integer number of columns is not important. So long as $\alpha_{\mathbf{c}} > 0$ the distribution is well-defined, normalized, and non-negative for every possible $\mathbf{r}$.

In other contexts the distribution Eq. (3.8) is known as the (symmetric) Dirichlet-multinomial distribution. When $\alpha_{\mathbf{c}} = 1$ the possible row sums $\mathbf{r}$ are uniformly distributed among the possible non-negative integer choices of r_i that sum to N. As $\alpha_{\mathbf{c}} \to \infty$ the distribution of $\mathbf{r}$ approaches a multinomial distribution where $\mathbf{r}$ is formed by taking N samples from a uniform probability vector $(1/m, \ldots, 1/m)$, the generalization of a symmetric binomial distribution. For $0 < \alpha_{\mathbf{c}} < 1$ the distribution of $\mathbf{r}$ will favor more extreme values of the coordinates r_i, analogous to the behavior of the symmetric Dirichlet distribution.

Our approximation involves replacing the true distribution $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ by $\mathbf{Pr}(\mathbf{r}|\alpha_{\mathbf{c}})$, with the value of $\alpha_{\mathbf{c}}$ chosen to make the approximation as good as possible in a certain sense. To do this we use a moment-matching approach in which the value of $\alpha_{\mathbf{c}}$ is chosen such that $\mathbf{Pr}(\mathbf{r}|\alpha_{\mathbf{c}})$ has the same mean and covariances as the true distribution $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$. Such a value always exists and it has a simple expression, as we now show.

The expectation and covariances of the r_i under $\mathbf{Pr}(\mathbf{r}|\alpha_\mathbf{c})$ are straightforward to compute:

$$\mathbf{E}(r_i) = \frac{N}{m}, \qquad \mathbf{cov}(r_i, r_k) = (N/m)\frac{m\alpha_\mathbf{c} + N}{m\alpha_\mathbf{c} + 1}\left(\delta_{ik} - m^{-1}\right). \tag{3.10}$$

For $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ the calculation is only a little more involved. In the uniform distribution over $A(\mathbf{c})$ each column j is independently uniformly distributed over the possible choices of matrix elements x_{ij} that satisfy $\sum_{i=1}^{m} x_{ij} = c_j$. The expectations and covariances of these column entries alone are then

$$\mathbf{E}(x_{ij}) = \frac{c_j}{m}, \qquad \mathbf{cov}(x_{ij}, x_{kj}) = \frac{c_j(m + c_j)}{m(m+1)}(\delta_{ik} - m^{-1}). \tag{3.11}$$

Since the row margins are the sums of these independent column entries, the expectations and covariances add so that

$$\mathbf{E}(r_i) = \sum_{j=1}^{n} \mathbf{E}(x_{ij}) = \frac{N}{m}, \quad \mathbf{cov}(r_i, r_k) = \sum_{j=1}^{n} \mathbf{cov}(x_{ij}, x_{kj}) = \frac{Nm + c^2}{m(m+1)}(\delta_{ik} - m^{-1}), \tag{3.12}$$

where we have introduced the shorthand $c^2 = \sum_{j=1}^{n} c_j^2$.

Thus the expectations of $\mathbf{Pr}(\mathbf{r}|\alpha_\mathbf{c})$ and $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ already match and, equating the covariances in Eqs. (3.10) and (3.12) and solving for $\alpha_\mathbf{c}$, we get

$$\alpha_\mathbf{c} = \frac{N^2 - N + (N^2 - c^2)/m}{c^2 - N}. \tag{3.13}$$

Finally, we assemble Eqs. (3.2) and (3.8) into our "effective columns" estimate of $\Omega(\mathbf{r}, \mathbf{c})$ thus:

$$\begin{aligned}\Omega^{\mathrm{EC}}(\mathbf{r}, \mathbf{c}) &= \mathrm{Pr}(\mathbf{r}|\alpha_\mathbf{c})\,|A(\mathbf{c})| \\ &= \binom{N + m\alpha_\mathbf{c} - 1}{m\alpha_\mathbf{c} - 1}^{-1} \prod_{i=1}^{m} \binom{r_i + \alpha_\mathbf{c} - 1}{\alpha_\mathbf{c} - 1} \prod_{j=1}^{n} \binom{c_j + m - 1}{m - 1},\end{aligned} \tag{3.14}$$

where $\alpha_\mathbf{c}$ is given by Eq. (3.13).

Note that, although $\Omega(\mathbf{r}, \mathbf{c})$ is trivially symmetric under the interchange of rows and columns, our estimate of it is not. (The same is true of the DE and GM estimates also.) The symmetry is broken when we choose to approximate $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ and not $\mathbf{Pr}(\mathbf{c}|\mathbf{r})$. In practice our estimate appears to perform better for matrices with more columns than rows $m < n$, as can be seen in Fig. 5, so it may improve performance to swap the definitions of $\mathbf{r}$ and $\mathbf{c}$ when $m > n$. This prescription also has the corollary effect of rendering the estimate symmetric.

Although our derivation of the EC estimate does not make specific reference to the sparse limit, it performs well in that regime. In fact, as we show in Appendix B, it gives an asymptotically exact result in the sparse limit where $N \to \infty$ with bounded row and column sums. This feature is not unprecedented, although among the estimates we consider it is shared by only the BBK and GMK estimates. Unlike those two estimates, however, the EC estimate exhibits good performance in the dense regime as well, where it is asymptotically equal to the DE estimate as also shown in Appendix B.

(b) The estimate of Gail and Mantel

In the following sections we review some of the other approaches for estimating $\Omega(\mathbf{r}, \mathbf{c})$, outlining the motivation and derivations of the other linear-time approximations discussed in Section 2. We start with an approach due to Gail and Mantel (GM) [12], who propose the following approximation for $\Omega(\mathbf{r}, \mathbf{c})$:

$$\Omega^{\mathrm{GM}}(\mathbf{r}, \mathbf{c}) = \left(\frac{m-1}{2\pi m\sigma^2}\right)^{(m-1)/2} m^{1/2} e^{-Q/2} \prod_{j=1}^{n} \binom{c_j + m - 1}{m - 1}, \tag{3.15}$$

where

$$\sigma^2 = \frac{(c^2 + mN)(m-1)}{(m+1)m^2}, \quad Q = \frac{m-1}{\sigma^2 m}\left(r^2 - \frac{N^2}{m}\right), \quad r^2 = \sum_i r_i^2, \quad c^2 = \sum_j c_j^2. \tag{3.16}$$

The derivation of this estimate follows the same general logic as our own, with the true distribution $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ approximated by a family of simpler distributions that is fitted to the true distribution with a moment-matching argument. The difference lies in the particular family used: Gail and Mantel use a multinormal distribution, by contrast with the Dirichlet-multinomial distribution in our derivation.

Numerical results for the approximation of Gail and Mantel were given in Fig. 1. The method is typically outperformed by the other linear-time estimators, indicating that there is some art to picking an appropriate family of distributions for the moment matching argument. We note that in this case the multinormal distribution is not justified (as one might imagine) by the central limit theorem, despite $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ being a mixture of independent columns, because the probability density is typically evaluated away from the expected value of $\mathbf{r}$, $\mathbf{E}(\mathbf{r}) = (N/m, \ldots, N/m)$, in a regime where local limit arguments do not apply.

(c) The estimate of Diaconis and Efron

A related approximation has been proposed by Diaconis and Efron (DE) [11]:

$$\Omega^{\mathrm{DE}}(\mathbf{r},\mathbf{c}) = \left(N + \frac{mn}{2}\right)^{(m-1)(n-1)} \left(\prod_{i=1}^{m} \bar{r}_i\right)^{K_{\mathbf{c}}-1} \left(\prod_{j=1}^{n} \bar{c}_j\right)^{m-1} \frac{\Gamma(mK_{\mathbf{c}})}{\Gamma(m)^n \Gamma(K_{\mathbf{c}})^m}, \tag{3.17}$$

where

$$\begin{aligned} w &= \frac{N}{N + \frac{1}{2}mn}, \qquad \bar{r}_i = \frac{1-w}{m} + \frac{wr_i}{N}, \qquad \bar{c}_j = \frac{1-w}{n} + \frac{wc_j}{N}, \\ K_{\mathbf{c}} &= \frac{m+1}{m\bar{c}^2} - \frac{1}{m}, \qquad \bar{c}^2 = \sum_j \bar{c}_j^2. \end{aligned} \tag{3.18}$$

The derivation of this estimate uses a moment-matching argument, like the estimates of this paper and of Gail and Mantel, Eqs. (2.1) and (3.15), but with some crucial differences. Instead of considering the set $A(\mathbf{r},\mathbf{c})$ of integer matrices with the required row and column sums, Diaconis and Efron consider the space (polytope) $P(\mathbf{r},\mathbf{c})$ of all $m \times n$ matrices of non-negative reals (not necessarily integers):

$$P(\mathbf{r},\mathbf{c}) = \left\{ \{x_{ij}\} \in \mathbb{R}^{m\times n} \middle| x_{ij} \geq 0, \sum_{i=1}^{m} x_{ij} = c_j, \sum_{j=1}^{n} x_{ij} = r_i \right\}. \tag{3.19}$$

The count $\Omega(\mathbf{r},\mathbf{c})$ of integer matrices can be thought of as the volume of the intersection of this polytope with the lattice formed by the (unconstrained) set A of non-negative integer matrices that sum to N:

$$\Omega(\mathbf{r},\mathbf{c}) = |A(\mathbf{r},\mathbf{c})| = |P(\mathbf{r},\mathbf{c}) \cap A|. \tag{3.20}$$

Diaconis and Efron use moment-matching not to estimate $|A(\mathbf{r},\mathbf{c})|$ but instead to estimate the volume of the polytope $P(\mathbf{r},\mathbf{c})$, then compute the size of the intersection Eq. (3.20) from it. Since the polytope is a continuous region, the distribution $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ on it is also continuous and is represented with a continuous approximant distribution $\mathbf{Pr}(\mathbf{r}|K_{\mathbf{c}})$, chosen to be N times the symmetric Dirichlet distribution $\mathrm{Dir}(K_{\mathbf{c}})$, with the Dirichlet parameter $K_{\mathbf{c}}$ chosen to match the mean and covariances of the true distribution $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ over the polytope. Armed with the resulting approximation for the volume of the polytope, $\Omega(\mathbf{r},\mathbf{c})$ is then estimated as the number of lattice points within it as in Eq. (3.20), calculated from the volume of the polytope times the density of lattice points. Finally, an "edge-effects" correction is applied to better reflect the number of lattice points contained.

For dense matrices the performance of this estimate is similar to that of our own estimate—see Fig. 1. Indeed, in the dense limit where $N \to \infty$ while the ratios of column and row sums are kept fixed, we can show that the two estimates are asymptotically equivalent (Appendix B). For sparse matrices, on the other hand, the approximation of the number of lattice points using the volume of the continuous polytope fails and the DE estimate breaks down.

(d) The estimate of Good and Crook

Good and Crook (GC) [9] proposed the following estimate for $\Omega^{\mathrm{GC}}(\mathbf{r}, \mathbf{c})$:

$$\Omega^{\mathrm{GC}}(\mathbf{r}, \mathbf{c}) = \binom{N + mn - 1}{mn - 1}^{-1} \prod_{i=1}^{m} \binom{r_i + n - 1}{n - 1} \prod_{j=1}^{n} \binom{c_j + m - 1}{m - 1}, \tag{3.21}$$

which is equivalent to our own estimate if one does not apply moment matching but instead simply assumes that the number of effective columns is equal to the number of actual columns: $\alpha_{\mathbf{c}} = n$. Under the circumstances, it seems likely that this estimate would not perform as well as our effective columns estimate, as indeed can be seen in the numerical results of Fig. 1.

(e) Estimates for the sparse regime

The remaining two linear-time approximations presented in Fig. 1 are closely related and both aimed at approximating $\Omega(\mathbf{r}, \mathbf{c})$ in the sparse regime where $N \ll mn$ and most matrix elements are zero. In this regime Békéssy, Békéssy, and Komlós (BBK) [13] give the following approximation (also proposed independently by O'Neil [21]):

$$\Omega^{\mathrm{BBK}}(\mathbf{r}, \mathbf{c}) = \frac{N!}{\prod_{i=1}^{m} r_i! \prod_{j=1}^{m} c_j!} \exp\left[\frac{2}{N^2} \sum_{i=1}^{m} \binom{r_i}{2} \sum_{j=1}^{n} \binom{c_j}{2}\right] \left[1 + \mathrm{O}\left(N^{-1}\right)\right], \tag{3.22}$$

where the error term describes the asymptotic growth of the error with N in the sparse limit where $r_{\max} = \max(\mathbf{r})$ and $c_{\max} = \max(\mathbf{c})$ are held fixed as $N \to \infty$, so that $m, n \to \infty$. Greenhill and McKay (GMK) [14] improved on this estimate with correction terms thus:

$$\begin{aligned} \Omega^{\mathrm{GMK}}(\mathbf{r}, \mathbf{c}) = \frac{N!}{\prod_{i=1}^{m} r_i! \prod_{j=1}^{n} c_j!} \exp\Bigg[&\frac{R_2 C_2}{2N^2} + \frac{R_2 C_2}{2N^3} + \frac{R_3 C_3}{3N^3} - \frac{R_2 C_2 \left(R_2 + C_2\right)}{4N^4} \\ &- \frac{R_2^2 C_3 + R_3 C_2^2}{2N^4} + \frac{R_2^2 C_2^2}{2N^5} + \mathrm{O}\left(\frac{r_{\max}^3 c_{\max}^3}{N^2}\right)\Bigg], \end{aligned} \tag{3.23}$$

where

$$R_k = \sum_{i=1}^{m} [r_i]_k \,, \qquad C_k = \sum_{j=1}^{n} [c_j]_k \,, \tag{3.24}$$

and $[x]_k$ is the falling factorial

$$[x]_k = x(x-1)\ldots(x-k+1). \tag{3.25}$$

Given the $\mathrm{O}\left(r_{\max}^3 c_{\max}^3 / N^2\right)$ form of the error term in (3.23), this estimate is asymptotically correct for $\log \Omega(\mathbf{r}, \mathbf{c})$ as $n, m, N \to \infty$ for sufficiently sparse matrices with $r_{\max} c_{\max} \sim \mathrm{o}(N^{2/3})$, and not just in the regime of constant $r_{\max}$ and $c_{\max}$ where the BBK estimate converges. Note that if we keep only the first term in the exponent of Eq. (3.23) we recover the BBK estimate, Eq. (3.22), so the GMK estimate can be viewed as a correction to the BBK estimate better tailored to the sparse limit.

The numerical performance of the BBK and GMK estimates is shown in Fig. 1. Both perform well in the sparse limit, as one might expect, but are poor in denser regimes. In the sparse limit with constant $r_{\max}$ and $c_{\max}$, our EC estimate converges asymptotically to the BBK estimate and hence also converges to the true $\Omega(\mathbf{r}, \mathbf{c})$, as explained in Appendix B. But unlike the other sparse estimates the EC estimate also performs well far from the sparse regime.

4. Maximum-entropy estimates

Barvinok and Hartigan have developed maximum entropy techniques for approximating a variety of counting problems, including two approximations for counts of contingency tables, a simpler and faster Gaussian approximation and a more refined approximation that incorporates a so-called Edgeworth correction [22].

(a) Gaussian maximum-entropy estimate

Barvinok and Hartigan [22] give a Gaussian approximation for $\Omega(\mathbf{r}, \mathbf{c})$, which under quite general conditions on $\mathbf{r}$ and $\mathbf{c}$ can be shown to return an asymptotically correct value of $\log \Omega(\mathbf{r}, \mathbf{c})$ as $N \to \infty$. The approximation takes the form

$$\Omega^{\mathrm{G}}(\mathbf{r}, \mathbf{c}) = \frac{e^{g(Z)}}{(2\pi)^{(m+n-1)/2}\sqrt{\det Q}}. \tag{4.1}$$

Here $g(Z)$ is a scalar function of a matrix $Z = (z_{ij})$ defined thus:

$$g(Z) = \sum_{ij} \left[(z_{ij}+1)\log(z_{ij}+1) - z_{ij}\log z_{ij}\right]. \tag{4.2}$$

The value of Z is chosen to maximize this function over the same polytope $P(\mathbf{r}, \mathbf{c})$ introduced in Eq. (3.19), the space of all matrices with non-negative real entries (not necessarily integers) that marginalize to $\mathbf{r}, \mathbf{c}$. Note that, since $g(Z)$ is concave, there is a unique Z that maximizes it within the polytope. In practice, this maximum is found numerically with one of the many standard methods for convex optimization.

The Q in Eq. (4.1) is an $(m+n-1) \times (m+n-1)$ symmetric matrix $Q = (q_{ij})$ whose nonzero elements are

$$\begin{aligned} q_{i,j+m} = q_{j+m,i} &= z_{ij}^2 + z_{ij} && \text{for } i = 1 \ldots m,\ j = 1 \ldots n-1, \\ q_{ii} &= r_i + \sum_{j=1}^{n} z_{ij}^2 && \text{for } i = 1 \ldots m, \\ q_{j+m,j+m} &= c_j + \sum_{i=1}^{m} z_{ij}^2 && \text{for } j = 1 \ldots n-1. \end{aligned} \tag{4.3}$$

The computation of the determinant of this matrix in Eq. (4.1) has time complexity roughly $\mathrm{O}((m+n)^3)$ and hence the evaluation of the entire estimate takes at least this long, which makes this method substantially more demanding for large matrices than the linear-time methods of Section 3.

To see where the Gaussian estimate comes from, consider a probability distribution $P(X|Z)$ over unrestricted non-negative integer matrices $X = (x_{ij})$ given a matrix $Z = (z_{ij})$ of real parameters. Each integer matrix element x_{ij} is independently drawn from a geometric distribution with expectation z_{ij}, which means the full distribution is

$$P(X|Z) = \prod_{ij} \left(\frac{1}{1+z_{ij}}\right)\left(\frac{z_{ij}}{1+z_{ij}}\right)^{x_{ij}}. \tag{4.4}$$

The entropy of this probability distribution is equal to the function $g(Z)$ defined in Eq. (4.2) and the value of Z is chosen to maximize this entropy over the polytope $Z \in P(\mathbf{r}, \mathbf{c})$. This means that, for this specific choice of Z, $P(X|Z)$ depends on X through the values of its row and column sums only, so that the distribution becomes uniform over $X \in P(\mathbf{r}, \mathbf{c})$, taking a constant value

equal to

$$P(X|Z) = e^{-g(Z)}. \tag{4.5}$$

Given that there are, by definition, $\Omega(\mathbf{r}, \mathbf{c})$ values of X inside the polytope, the total probability that X lies in the polytope, and hence that it has margins $\mathbf{r}$ and $\mathbf{c}$, is given by $P\{X \in P(\mathbf{r}, \mathbf{c})|Z\} = e^{-g(Z)}\Omega(\mathbf{r}, \mathbf{c})$, and hence

$$\Omega(\mathbf{r}, \mathbf{c}) = e^{g(Z)} P\{X \in P(\mathbf{r}, \mathbf{c})|Z\}. \tag{4.6}$$

Thus, if we can calculate the probability that X has the correct margins under the distribution Eq. (4.4) we can calculate $\Omega(\mathbf{r}, \mathbf{c})$. To do this, we observe that the polytope $P(\mathbf{r}, \mathbf{c})$ is defined by a set of linear constraints with the general form $AX = b$, where A is an $(m + n - 1) \times mn$ matrix, b is an $(m + n - 1)$-vector, and X is now represented in "unrolled" form as an mn-element vector rather than an $m \times n$ matrix. We then consider the $(m + n - 1)$-dimensional random variable $Y = AX$. This transformed variable satisfies $Y = AX = b$ on $P(\mathbf{r}, \mathbf{c})$ and hence $P\{X \in P(\mathbf{r}, \mathbf{c})|Z\} = P\{Y = b|Z\}$.

Finally, since the entries of X are independent random variables, we expect the distribution of Y to be asymptotically Gaussian by the local central limit theorem. This allows us to approximate the distribution with a Gaussian, and, matching the covariances of this Gaussian with the true covariances of Y, which are captured in the matrix Q of Eq. (4.3), we can estimate $P\{Y = b|Z\}$ and hence the value of $\Omega(\mathbf{r}, \mathbf{c})$.

(b) Edgeworth correction

Building on the Gaussian approximation, Barvinok and Hartigan [15] have given a further improved approximation for $\Omega(\mathbf{r}, \mathbf{c})$ by employing a so-called Edgeworth correction. This takes the form

$$\Omega^{\mathrm{E}}(\mathbf{r}, \mathbf{c}) = \frac{e^{g(Z)}}{(2\pi)^{(m+n-1)/2}\sqrt{\det Q}} \exp\left(-\frac{\mu}{2} + \nu\right), \tag{4.7}$$

where μ and ν are defined below. Barvinok and Hartigan show that under some mild conditions on the growth of the margins $\mathbf{r}$ and $\mathbf{c}$, this gives an asymptotically correct estimate of $\Omega(\mathbf{r}, \mathbf{c})$ as $N \to \infty$.

To specify the values of μ and ν in Eq. (4.7) a few more definitions are needed. First, we define a quadratic form $q : \mathbb{R}^{m+n-1} \to \mathbb{R}$ by

$$q(x) = \tfrac{1}{2} x^T Q x, \tag{4.8}$$

where Q is the matrix defined in Eq. (4.3). We also define two functions $f, h : \mathbb{R}^{m+n-1} \to \mathbb{R}$ on the variables $(u_1, \ldots, u_m, t_1, \ldots, t_{n-1}) \in \mathbb{R}^{m+n-1}$ thus (with $t_n = 0$):

$$f(u, t) = \frac{1}{6} \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} z_{ij}\left(z_{ij} + 1\right)\left(2z_{ij} + 1\right)\left(u_i + t_j\right)^3, \tag{4.9}$$

$$h(u, t) = \frac{1}{24} \sum_{\substack{1 \le i \le m \\ 1 \le j \le n}} z_{ij}\left(z_{ij} + 1\right)\left(6z_{ij}^2 + 6z_{ij} + 1\right)\left(u_i + t_j\right)^4. \tag{4.10}$$

The Edgeworth correction terms are then given by

$$\mu = \mathbf{E}\left(f^2\right), \qquad \nu = \mathbf{E}(h), \tag{4.11}$$

where the expectations are taken over the Gaussian probability density on $\mathbb{R}^{m+n-1}$ proportional to e^{-q}.

Barvinok and Hartigan note that, given the definition Eq. (4.8), Q^{-1} is the covariance matrix of the u_i and t_j under the distribution e^{-q}. This distribution e^{-q} is symmetric under $(u_i, t_j) \to$

$(-u_i, -t_j)$ so that $\mathbf{E}(u_i) = \mathbf{E}(t_j) = 0$ and hence

$$\mathbf{E}(u_i t_j) = (Q^{-1})_{i(j+m)}\,. \tag{4.12}$$

The values of μ and ν can then be evaluated using Wick contractions for correlators of Gaussian random variables to express the desired expectations in terms of covariances given by Eq. (4.12). Specifically, one uses

$$\mathbf{E}\big[(u_i + t_j)^4\big] = 3\big[\mathbf{E}(u_i^2) + 2\mathbf{E}(u_i t_j) + \mathbf{E}(t_j^2)\big]^2 \tag{4.13}$$

and

$$\begin{aligned}\mathbf{E}\big[(u_{i_1} + t_{j_1})^3 (u_{i_2} + t_{j_2})^3\big] &= 3\big[\mathbf{E}(u_{i_1}u_{i_2}) + \mathbf{E}(u_{i_1}t_{j_2}) + \mathbf{E}(u_{i_2}t_{j_1}) + \mathbf{E}(t_{i_1}t_{j_2})\big] \\ &\times \big[\mathbf{E}(u_{i_1}u_{i_2}) + 2\big(\mathbf{E}(u_{i_1}t_{j_2}) + \mathbf{E}(u_{i_2}t_{j_1}) + \mathbf{E}(t_{j_1}t_{j_2})\big)^2 \\ &+ 3\big(\mathbf{E}(u_{i_1}^2) + 2\mathbf{E}(u_{i_1}t_{j_1}) + \mathbf{E}(t_{j_1}^2)\big)\big(\mathbf{E}(u_{i_2}^2) + 2\mathbf{E}(u_{i_2}t_{j_2}) + \mathbf{E}(t_{j_2}^2)\big)\big]. \end{aligned} \tag{4.14}$$

Note that evaluating μ requires a sum over all possible $i_1, i_2 = 1 \ldots m$ and $j_1, j_2 = 1 \ldots n-1$ so the complexity of the calculation is $\mathrm{O}(m^2 n^2)$, making the minimum computational burden higher than for just the Gaussian estimate. In practice, the running time of either of the maximum-entropy estimates is not significant for small matrices: our implementations of both run in under a second for $m, n \lesssim 32$. On the other hand, very large matrices of size $m, n \gtrsim 512$ can take well over an hour, and running time can also be an issue when one needs estimates for a large number of smaller matrices. For cases where running time is a concern, Section (b) of Appendix C gives our recommendations for various parameter values.

5. Sequential importance sampling

Sequential importance sampling (SIS) is a computational technique that in the present case can be used either to sample from the set of non-negative integer matrices $A(\mathbf{r}, \mathbf{c})$ [2, 3] or to find the size $\Omega(\mathbf{r}, \mathbf{c})$ of the set. In this section, we review the standard SIS approach and show how it can be improved by exploiting our new linear-time estimate. Advances in SIS for contingency tables have been made in the past through the incorporation of faster and more accurate approximations for $\Omega(\mathbf{r}, \mathbf{c})$ [3, 4]. For example, Eisinger and Chen [4] used the GC and GMK estimates to optimize SIS in the sparse and dense regimes respectively. Here we take a similar approach with our EC estimate, but since the EC estimate performs well across all regimes from sparse to dense, it allows us to perform sampling using the same approximation in all cases. Moreover, as shown in Fig. 2, the EC estimate offers roughly a one-hundred-fold improvement in accuracy over the GC estimate when used to estimate $\Omega(\mathbf{r}, \mathbf{c})$, which results in a corresponding gain in efficiency for importance sampling.

The main ingredient of importance sampling is a "trial distribution" $q(X)$ over matrices X which is nonzero if and only if $X \in A(\mathbf{r}, \mathbf{c})$. If we can sample matrices from this distribution then we have

$$\mathbf{E}_q\left[\frac{1}{q(X)}\right] = \sum_{X \in A(\mathbf{r},\mathbf{c})} q(X)\frac{1}{q(X)} = |A(\mathbf{r}, \mathbf{c})| = \Omega(\mathbf{r}, \mathbf{c}). \tag{5.1}$$

Thus if we can draw N matrices $X^{(1)} \ldots X^{(N)}$ from $q(X)$ we can estimate $\Omega(\mathbf{r}, \mathbf{c})$ as

$$\widehat{\Omega}(\mathbf{r}, \mathbf{c}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{q(X^{(i)})}\,, \tag{5.2}$$

and the accompanying statistical error can be estimated in the conventional manner.

Alternatively, we can use the same approach to estimate an expectation under the uniform distribution over all integer matrices with fixed margins: $\mu = \mathbf{E}_{\mathrm{uni}}[f(X)]$. We compute an

estimator for this average as

$$\hat{\mu} = \frac{\sum_{i=1}^{N} f(X_i)/q(X_i)}{\sum_{i=1}^{N} 1/q(X_i)}. \tag{5.3}$$

For example, if we choose $f(X) = 1$ if the χ^2 statistic of X is greater than some value χ_0^2 and zero otherwise, this expression estimates the p-value of the χ^2-statistic for the uniform distribution over contingency tables, which has been proposed by Diaconis and Efron as an alternative to more traditional tests of independence [11].

In principle, these estimates converge regardless of the form of $q(X)$, but they become more efficient the closer the distribution is to being uniform over $A(\mathbf{r}, \mathbf{c})$, since the values of the sums in Eqs. (5.2) and (5.3) are dominated by the states with the smallest $q(X)$, which are unlikely to be sampled when $q(X)$ is highly nonuniform. The key to making the method work well lies in finding a $q(X)$ that is sufficiently close to the uniform distribution while still being straightforward to work with. The latter condition can be difficult to satisfy. We can trivially choose $q(X)$ to be exactly uniform by setting its value to a constant, but in that case the constant is $q(X) = 1/\Omega(\mathbf{r}, \mathbf{c})$, so calculating the value of $q(X)$ would be exactly as hard as calculating $\Omega(\mathbf{r}, \mathbf{c})$ in the first place.

SIS gets around these difficulties by sampling the matrix X one column at a time. (This is the "sequential" part of sequential importance sampling.) The idea is to first sample values X_1 of the first column of X with probabilities as close as possible to the probability with which they appear under the uniform distribution, which can be written as

$$p(X_1) = \frac{\Omega(\mathbf{r}', \mathbf{c}')}{\Omega(\mathbf{r}, \mathbf{c})}, \tag{5.4}$$

where $\mathbf{r}'$ and $\mathbf{c}'$ denote the row and column sums of the matrix after the first column is removed.

After the first column is sampled we repeat the process and sample values of the second column, then the third, and so forth until one has a sample of the entire matrix. If at each step the exact probabilities $p(X_i)$ in Eq. (5.4) are used, this process will sample the matrices $X \in A(\mathbf{r}, \mathbf{c})$ exactly uniformly, and indeed this is the approach taken by some methods [8], although these approaches are computationally costly and moreover require us to know $\Omega(\mathbf{r}, \mathbf{c})$ exactly and hence are not suitable for calculating $\Omega(\mathbf{r}, \mathbf{c})$ itself. For most purposes a better approach is to approximate the exact distribution $p(X_1)$ of Eq. (5.4) with some other distribution $q(X_1)$ that is easier to compute, at the expense of modestly nonuniform sampling. Despite the non-uniformity, we still get a convergent estimate for $\Omega(\mathbf{r}, \mathbf{c})$ using Eq. (5.2) as $N \to \infty$.

In choosing a value for $q(X_1)$, the various linear-time estimates for $\Omega(\mathbf{r}, \mathbf{c})$ in Section 3 provide an elegant route forward, and specifically, given its good performance on test cases, we propose using our effective columns estimate $\Omega^{\mathrm{EC}}(\mathbf{r}, \mathbf{c})$ of Eq. (2.1) to define a distribution over the column $X_1 = (x_{i1})$ thus:

$$q(X_1) = \frac{\Omega^{\mathrm{EC}}(\mathbf{r}', \mathbf{c}')}{\Omega^{\mathrm{EC}}(\mathbf{r}, \mathbf{c})} \propto \prod_{i=1}^{m} \binom{r_i - x_{i1} + \alpha_{\mathbf{c}'} - 1}{\alpha_{\mathbf{c}'} - 1} \mathbf{1}_{\sum_i x_{i1} = c_1} \mathbf{1}_{0 \le x_{i1} \le r_i}. \tag{5.5}$$

This expression combines our combinatorial estimate with hard constraints that impose the correct sum of the generated column $\sum_i x_{i1} = c_1$ and prevent any entry from surpassing the value of the remaining row sums $0 \le x_{i1} \le r_i$, which would make it impossible to complete the rest of the columns.

As described by Harrison and Miller [3], it is possible to sample the column X_1 from a distribution of the form (5.5) in time $\mathrm{O}(mc_1^2)$. The full SIS method samples each of the n columns in turn for a total time complexity of roughly $\mathrm{O}(N^2 m/n)$ per full sample. Performance can be improved by a numerical factor (but not in overall complexity) by arranging the elements of $\mathbf{c}$ in non-increasing order.

(a) Results

The method described above performs well, as shown in Fig. 2. The leftmost panel, labeled "EC-based SIS," shows results for our method, while the other panels show two other methods for comparison. "GC-based SIS," considered by Eisinger and Chen [4], employs a similar approach to ours but with a trial distribution based on the Good-Crook (GC) estimate [9], which appears to have the second-best broad performance behind our EC estimate (see Fig. 1). We find that the fractional error for the GC-based method is between 10 and 100 times larger than that for the EC-based method.

The third panel in Fig. 2, labeled "Greedy SIS," shows results from the method of Chen, Diaconis, Holmes, and Liu [2]. In this method the entry x_{11} is directly sampled from the distribution

$$\mathbf{Pr}(x_{11} = k) \propto \min(r_2, c_1 - k) + \max(0, c_1 + r_1 + r_2 - N - k) + 1, \tag{5.6}$$

and similarly for each remaining entry of X. This approach gives faster sampling than ours, at a rough complexity of $\mathrm{O}(Nm)$ per full sample, but at the expense of greater non-uniformity in $q(X)$. The trade-off turns out not to be beneficial. Convergence is slowed considerably for all but the smallest of matrix sizes and overall accuracy suffers, as shown in Fig. 2.

Based on these results, we have chosen the EC-based SIS technique for computing the ground-truth estimates of $\Omega(\mathbf{r}, \mathbf{c})$ employed in our work. We emphasize that this does not in any way bias the outcome of our benchmarking comparisons in Fig. 1 in favor of the EC estimate. All SIS methods, regardless of their choice of trial distribution, give convergent estimates; the choice of an EC-based trial distribution merely improves the rate of convergence of those estimates.

6. Conclusions

In this paper we have studied the problem of estimating the number $\Omega(\mathbf{r}, \mathbf{c})$ of non-negative integer matrices with given row and column sums, which arises for example in statistical and information theoretic calculations involving contingency tables. There is no known exact expression for $\Omega(\mathbf{r}, \mathbf{c})$, but a variety of methods for approximating it have been proposed in the past. We have presented two new methods that improve upon these previous approaches. First, we have proposed a closed-form approximation based on the concept of *effective columns*, which can be evaluated in time linear in the number $m + n$ of rows plus columns of the matrix and returns results of accuracy similar to or better than other linear-time estimates in the extensive benchmark tests presented here. Second, the same effective columns approximation is also used to derive a sequential importance sampling (SIS) algorithm for sampling such tables, which can be used to make numerical estimates of $\Omega(\mathbf{r}, \mathbf{c})$ with significantly faster convergence than previous SIS methods, resulting in estimates about 100 times more accurate in comparable running times.

Acknowledgments

We thank Alexander Barvinok, Peter Bickel, Brendan McKay and Igor Pak for helpful comments. This work was supported in part by the US National Science Foundation under grant DMS–2005899 and by computational resources provided by the Advanced Research Computing initiative at the University of Michigan. For code implementing the methods described here see `https://github.com/maxjerdee/contingency_count`.

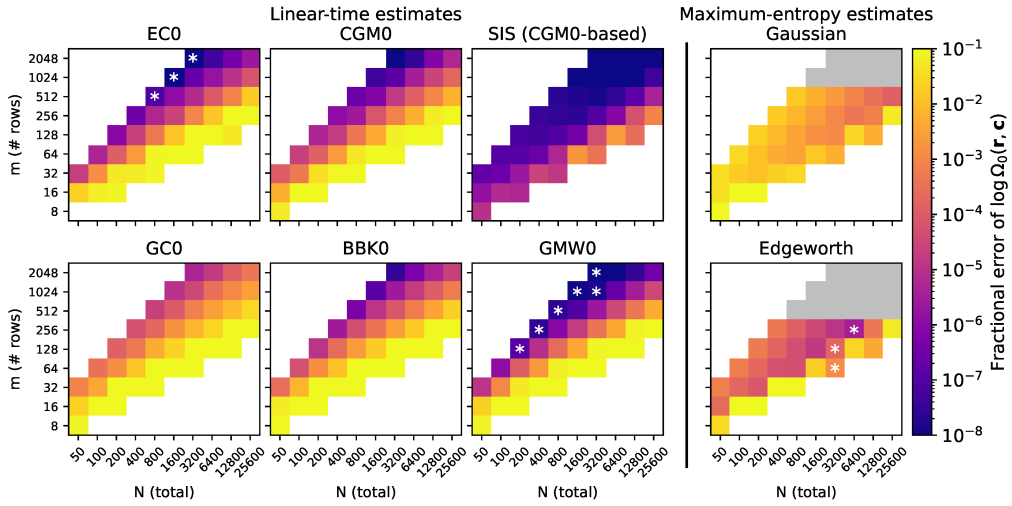

Figure 3. Fractional error of various estimates of $\log \Omega_0(\mathbf{r}, \mathbf{c})$ for square $m \times m$ matrices that sum to N. For compactness the estimated errors of the SIS method used for benchmarking are also plotted under the linear-time category. As in Fig. 1, white regions denote impossible parameter combinations, while gray regions indicate estimates that cannot be completed in an hour of run time on our hardware.

Appendices

A. Counting 0-1 matrices

A parallel problem to that of counting non-negative integer matrices is that of counting matrices whose elements take only the values zero and one, with row and columns sums once again fixed at given values $\mathbf{r}$ and $\mathbf{c}$. This problem is important in its own right and has been the subject of considerable work. The methods of this paper can be applied to the case of 0-1 matrices, but we do not emphasize this approach because in practice it does not improve upon existing methods. Nonetheless, for the sake of completeness, we describe the approach in this appendix and compare its performance with other available estimates.

(a) Summary of results

Let $\Omega_0(\mathbf{r}, \mathbf{c})$ be the number of 0-1 matrices with margins $\mathbf{r}, \mathbf{c}$. Figure 3 summarizes the performance of various estimates of $\log \Omega_0(\mathbf{r}, \mathbf{c})$ for square $m \times m$ matrices that sum to N. In the linear-time category we consider five estimates. The first, an analogue for the 0-1 case of our effective columns estimate, performs fairly well, but it does not reliably outperform the other linear-time estimates, and in general all the linear-time methods struggle with dense matrices.

To generate Fig. 3, for each combination of parameters (N, m), ten margin pairs $\mathbf{r}, \mathbf{c}$ were drawn uniformly from the set of all values that correspond to at least one 0-1 matrix (i.e., uniformly over values satisfying the *Gale-Ryser condition* [23, 24]). The ground truth is computed using sequential importance sampling as described in Section 5 with a trial distribution based on the CGM0 estimate as in [3]. The resulting estimated errors on the sequential importance sampling are also shown in Fig. 3.

(b) Effective columns estimate

Motivated by the same "effective columns" reasoning we used for our estimate of the number of contingency tables, we propose the following estimate for the number of 0-1 matrices with

margins $\mathbf{r}, \mathbf{c}$:

$$\Omega_0^{\mathrm{EC}}(\mathbf{r},\mathbf{c}) = \left| \binom{m\alpha_{\mathbf{c}}^{(0)}}{N}^{-1} \prod_{i=1}^{m} \binom{\alpha_{\mathbf{c}}^{(0)}}{r_i} \prod_{j=1}^{n} \binom{m}{c_j} \right|, \tag{A 1}$$

where

$$\alpha_{\mathbf{c}}^{(0)} = \frac{N^2 - N - (N^2 - c^2)/m}{c^2 - N}. \tag{A 2}$$

As we now explain, this estimate stands on less certain ground than our estimate of $\Omega(\mathbf{r},\mathbf{c})$, but the resulting formula nonetheless appears to be quite accurate.

Let $A_0(\mathbf{c})$ be the set of 0-1 matrices that have column sums $\mathbf{c}$. The number of such matrices can be found by independently choosing one column at a time. For each column j there are c_j elements equal to 1 and the rest are 0, so there are $\binom{m}{c_j}$ ways to distribute the 1s in the column. Since the columns are independent we then have

$$|A_0(\mathbf{c})| = \prod_{j=1}^{n} \binom{m}{c_j}. \tag{A 3}$$

Given this number, $\Omega(\mathbf{r},\mathbf{c})$ can be estimated as before from a knowledge of the conditional distribution $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$, and for this we again take inspiration from the unconditional distribution of $\mathbf{r}$,

$$\mathbf{Pr}(\mathbf{r}) = \binom{mn}{N}^{-1} \prod_{i=1}^{m} \binom{n}{r_i}, \tag{A 4}$$

replacing the number of columns n with an effective number $\alpha_{\mathbf{c}}^{(0)}$:

$$\tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)}) = \binom{m\alpha_{\mathbf{c}}^{(0)}}{N}^{-1} \prod_{i=1}^{m} \binom{\alpha_{\mathbf{c}}^{(0)}}{r_i}. \tag{A 5}$$

Unlike the case of non-negative integer matrices, where we were left with a well-defined distribution for any $\alpha_{\mathbf{c}} > 0$, $\tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)})$ is not quite a probability distribution over $\mathbf{r}$. Away from the poles, when $\alpha_{\mathbf{c}}^{(0)} \notin \{0, 1/m, \ldots, (N-1)/m\}$, Eq. (A 5) is properly normalized

$$\sum_{\mathbf{r}|\sum_i r_i = N} \tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)}) = 1, \tag{A 6}$$

but it is no longer non-negative for all $\mathbf{r}$: we can have $\tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)}) < 0$. In spite of this we press on and evaluate the "expectations" and "co-variances" of the r_i weighted by $\tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)})$:

$$\mathbf{E}(r_i) = \frac{N}{m}, \qquad \mathbf{cov}(r_i, r_k) = \frac{N(m\alpha_{\mathbf{c}}^{(0)} - N)}{m(m\alpha_{\mathbf{c}}^{(0)} - 1)} \left(\delta_{ik} - m^{-1}\right). \tag{A 7}$$

The true probability density $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ is again a mixture of independent columns with expectation and covariances

$$\mathbf{E}(r_i) = \frac{N}{m}, \qquad \mathbf{cov}(r_i, r_k) = \frac{Nm - c^2}{m(m-1)} \left(\delta_{ik} - m^{-1}\right). \tag{A 8}$$

The choice of the parameter $\alpha_{\mathbf{c}}^{(0)}$ such that the covariances of $\tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)})$ and $\mathbf{Pr}(\mathbf{r}|\mathbf{c})$ match is then

$$\alpha_{\mathbf{c}}^{(0)} = \frac{N^2 - N - (N^2 - c^2)/m}{c^2 - N}, \tag{A 9}$$

and our estimate of $\Omega_0(\mathbf{r},\mathbf{c})$ is given by $\tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)})|A_0(\mathbf{c})|$. In most cases, this expression can be used directly, but on the occasional instances where $\tilde{P}(\mathbf{r}|\alpha_{\mathbf{c}}^{(0)})$ is negative the resulting estimate can be negative as well. We remedy this issue in an ad-hoc way by taking the absolute value of the

result, which yields the estimate Eq. (A 1). In spite of this uncontrolled step the estimate performs reasonably well in the tests shown in Fig. 3.

(c) Other estimates

We also consider four other linear-time estimates of $\Omega_0(\mathbf{r},\mathbf{c})$ drawn from the literature, several of which are related to those for the case of general non-negative integer matrices. Good and Crook (GC0) [10] give an estimate which can be understood as our effective columns estimate but with the number of effective columns equal to the number of true columns:

$$\Omega_0^{\mathrm{GC}}(\mathbf{r},\mathbf{c}) = \binom{mn}{N}^{-1} \prod_{i=1}^{m} \binom{n}{r_i} \prod_{j=1}^{n} \log \binom{m}{c_j}. \tag{A 10}$$

Békéssy, Békéssy, and Komlós (BBK0) [13] provide an estimate suited to the sparse regime:

$$\Omega_0^{\mathrm{BBK}}(\mathbf{r},\mathbf{c}) = \frac{N!}{\prod_{i=1}^{m} r_i! \prod_{j=1}^{m} c_j!} \exp\left[-\frac{2}{N^2} \sum_{i=1}^{m} \binom{r_i}{2} \sum_{j=1}^{n} \binom{c_j}{2}\right] \left[1 + \mathrm{O}(N^{-1})\right], \tag{A 11}$$

which is improved by Greenhill, McKay, and Wang (GMW0) thus [25]:

$$\begin{aligned} \Omega_0^{\mathrm{GMW}}(\mathbf{r},\mathbf{c}) = \frac{N!}{\prod_{i=1}^{m} r_i! \prod_{j=1}^{n} c_j!} \exp\Bigg[&-\frac{R_2C_2}{2N^2} - \frac{R_2C_2}{2N^3} + \frac{R_3C_3}{3N^3} - \frac{R_2C_2\,(R_2+C_2)}{4N^4} \\ &- \frac{R_2^2C_3 + R_3C_2^2}{2N^4} + \frac{R_2^2C_2^2}{2N^5} + O\left(\frac{r_{\max}^3 c_{\max}^3}{N^2}\right)\Bigg]. \end{aligned} \tag{A 12}$$

Canfield, Greenhill, and McKay (CGM0) [26] provide an estimate for dense 0-1 matrices that can be understood as a correction to the GC estimate:

$$\Omega_0^{\mathrm{CGM}}(\mathbf{r},\mathbf{c}) = \binom{mn}{N}^{-1} \prod_{i=1}^{m} \binom{n}{r_i} \prod_{j=1}^{n} \binom{m}{c_j} \exp\left[-\frac{1}{2}\left(1 - \frac{R}{2Amn}\right)\left(1 - \frac{C}{2Amn}\right)\right], \tag{A 13}$$

where

$$R = \sum_{i=1}^{m} \left(r_i - \frac{N}{m}\right)^2, \quad C = \sum_{j=1}^{n} \left(c_j - \frac{N}{n}\right)^2, \quad \lambda = \frac{N}{mn}, \quad A = \tfrac{1}{2}\lambda(1-\lambda). \tag{A 14}$$

Canfield *et al.* show that this is in fact asymptotically correct under certain conditions—loosely when the matrix is relatively square and has density not too close to 0 or 1. Finally, Barvinok and Hartigan [27] give maximum-entropy estimates in Gaussian and Edgeworth-corrected varieties analogous to those of Section 4.

Figure 3 shows a quantitative comparison of the accuracy of each of these estimates, along with our effective columns (EC0) estimate and numerical SIS results for a range of sizes of square matrices and values of the sum N of all elements. Based on these results, it appears that the EC0, CGM0 and GMW0 estimates all provide good accuracy in general, particularly in the sparse regime, with EC0 and GMW0 exceeding the accuracy of SIS in some cases. The maximum-entropy estimates perform poorly in the sparse regime, but better for denser matrices. When the Edgeworth correction is included they provide the most accurate results in the dense regime, although at the expense of greater computational effort.

B. Validity of the effective columns estimate

In this appendix, we demonstrate various properties of our effective columns (EC) estimate as defined in Eq. (2.1). Specifically, we show exact behavior in the degenerate $\alpha_{\mathbf{c}} \to \infty$ limit, asymptotically correct behavior in the sparse $N \to \infty$ limit, and agreement with the DE estimate [11] in the dense limit.

(a) Degenerate $\alpha_{\mathbf{c}}$

The value of the parameter $\alpha_{\mathbf{c}}$ in our estimate is given by Eq. (2.2) to be

$$\alpha_{\mathbf{c}} = \frac{N^2 - N + (N^2 - c^2)/m}{c^2 - N}. \tag{A 1}$$

At first sight this expression appears potentially problematic, since the denominator could become zero or negative. In fact, it cannot be negative because

$$c^2 - N = \sum_{j=1}^{n} c_j^2 - \sum_{j=1}^{n} c_j = \sum_{j=1}^{n} c_j(c_j - 1) \geq 0. \tag{A 2}$$

The value could however be zero if c_j is either zero or one for all j, and this would cause $\alpha_{\mathbf{c}}$ to diverge. In practice we can ignore columns with $c_j = 0$ since these have no effect on the number of matrices $\Omega(\mathbf{r}, \mathbf{c})$, so let us assume that all such columns have been removed. What then happens if all remaining columns have $c_j = 1$? In this case it turns out that the limit $\alpha_{\mathbf{c}} \to \infty$ of the estimate of $\Omega(\mathbf{r}, \mathbf{c})$ does give the correct result, as we now show.

If all columns have $c_j = 1$ then all elements in a column are zero except for a single 1. The constraint on the row sums then demands that r_i out of the n columns have their 1 in row i for all $i = 1 \ldots m$. The number of possible arrangements satisfying this requirement is

$$\Omega(\mathbf{r}, \mathbf{c} = (1, \ldots, 1)) = \frac{n!}{\prod_{i=1}^{m} r_i!}. \tag{A 3}$$

We now show that the $\alpha_{\mathbf{c}} \to \infty$ limit of our estimate Eq. (2.1) for this situation gives this exact result.

Recall that our EC estimate is given by

$$\Omega^{\mathrm{EC}}(\mathbf{r}, \mathbf{c}) = \binom{N + m\alpha_{\mathbf{c}} - 1}{m\alpha_{\mathbf{c}} - 1}^{-1} \prod_{i=1}^{m} \binom{r_i + \alpha_{\mathbf{c}} - 1}{\alpha_{\mathbf{c}} - 1} \prod_{j=1}^{n} \binom{c_j + m - 1}{m - 1}. \tag{A 4}$$

Noting that when $c_j = 1$ for all j we have $N = n$, we write

$$\binom{N + m\alpha_{\mathbf{c}} - 1}{m\alpha_{\mathbf{c}} - 1} = \frac{\Gamma(n + m\alpha_{\mathbf{c}})}{\Gamma(m\alpha_{\mathbf{c}})\Gamma(n + 1)}, \qquad \binom{r_i + \alpha_{\mathbf{c}} - 1}{\alpha_{\mathbf{c}} - 1} = \frac{\Gamma(r_i + \alpha_{\mathbf{c}})}{\Gamma(\alpha_{\mathbf{c}})\Gamma(r_i + 1)}, \tag{A 5}$$

and apply Stirling's approximation in the form

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z \left[1 + \mathrm{O}(z^{-1})\right], \tag{A 6}$$

which in the limit of large $\alpha_{\mathbf{c}}$ gives

$$\binom{n + m\alpha_{\mathbf{c}} - 1}{m\alpha_{\mathbf{c}} - 1} = \frac{(m\alpha_{\mathbf{c}})^n}{\Gamma(n + 1)} \left[1 + \mathrm{O}(\alpha_{\mathbf{c}}^{-1})\right], \tag{A 7}$$

$$\binom{r_i + \alpha_{\mathbf{c}} - 1}{\alpha_{\mathbf{c}} - 1} = \frac{\alpha_{\mathbf{c}}^{r_i}}{\Gamma(r_i + 1)} \left[1 + \mathrm{O}(\alpha_{\mathbf{c}}^{-1})\right]. \tag{A 8}$$

Then our estimate, Eq. (A 4), is

$$\begin{aligned} &\binom{n + m\alpha_{\mathbf{c}} - 1}{m\alpha_{\mathbf{c}} - 1}^{-1} \prod_{i=1}^{m} \binom{r_i + \alpha_{\mathbf{c}} - 1}{\alpha_{\mathbf{c}} - 1} \prod_{j=1}^{n} \binom{1 + m - 1}{m - 1} \\ &\quad = \left[\frac{\Gamma(n + 1)}{(m\alpha_{\mathbf{c}})^n} \prod_{i=1}^{m} \frac{\alpha_{\mathbf{c}}^{r_i}}{\Gamma(r_i + 1)} \prod_{j=1}^{n} m\right] \left[1 + \mathrm{O}(\alpha_{\mathbf{c}}^{-1})\right] = \frac{n!}{\prod_{i=1}^{m} r_i!} \left[1 + \mathrm{O}(\alpha_{\mathbf{c}}^{-1})\right]. \end{aligned} \tag{A 9}$$

So the $\alpha_{\mathbf{c}} \to \infty$ limit indeed recovers the correct result.

This is a nice property of our estimate. In a practical implementation we can recognize the case $\mathbf{c} = (1, \ldots, 1)$ and either return the exact result, Eq. (A 3), or simply evaluate the usual estimate at

a large value of $\alpha_{\mathbf{c}}$. The latter prescription is also equivalent to writing

$$\alpha_{\mathbf{c}} = \frac{N^2 - N + (N^2 - c^2)/m}{c^2 - N + \epsilon} \tag{A 10}$$

for ϵ small and positive.

(b) Effective columns estimate in the sparse limit

In this section, we show that in the sparse limit where $N \to \infty$ but the row and column sums are bounded, the EC estimate is asymptotically exact:

$$\Omega^{\mathrm{EC}}(\mathbf{r}, \mathbf{c}) = \Omega(\mathbf{r}, \mathbf{c})\left[1 + \mathrm{O}\left(N^{-1}\right)\right]. \tag{A 11}$$

To demonstrate this, suppose that the row sums are bounded above by $r_{\max}$ and the column sums by $c_{\max}$. We then define the following quantities, equal to the fraction of row and column sums that take on each possible value:

$$\hat{r}_k = \frac{1}{m}\left|\{i | r_i = k\}\right|, \quad \text{for } k = 1 \ldots r_{\max} \tag{A 12}$$

$$\hat{c}_l = \frac{1}{n}\left|\{j | r_j = l\}\right|, \quad \text{for } l = 1 \ldots c_{\max}. \tag{A 13}$$

We now observe the following expressions for various sums which appear in the estimates we consider:

$$N = \sum_{i=1}^{m} r_i = \sum_{k=1}^{r_{\max}} m\hat{r}_k k = m\hat{r}^{(1)}, \qquad \sum_{i=1}^{m} r_i^2 = \sum_{k=1}^{r_{\max}} m\hat{r}_k k^2 = m\hat{r}^{(2)}, \tag{A 14}$$

$$N = \sum_{j=1}^{n} c_j = \sum_{l=1}^{c_{\max}} n\hat{c}_l l = n\hat{c}^{(1)}, \qquad \sum_{j=1}^{n} c_j^2 = \sum_{l=1}^{c_{\max}} n\hat{c}_l l^2 = n\hat{c}^{(2)}, \tag{A 15}$$

where $\hat{r}^{(1)}$ and $\hat{r}^{(2)}$ are the first and second moments of $\mathbf{r}$, and similarly for $\mathbf{c}$:

$$\hat{r}^{(1)} = \sum_{k=1}^{r_{\max}} \hat{r}_k k, \qquad \hat{r}^{(2)} = \sum_{k=1}^{r_{\max}} \hat{r}_k k^2, \tag{A 16}$$

$$\hat{c}^{(1)} = \sum_{l=1}^{c_{\max}} \hat{c}_l l, \qquad \hat{c}^{(2)} = \sum_{l=1}^{c_{\max}} \hat{c}_l l^2. \tag{A 17}$$

Since these moments are all bounded for constant $r_{\max}$ and $c_{\max}$, m and n grow as $\mathrm{O}(N)$ in this sparse limit.

Applying these new expressions, the BBK estimate [13] (equivalent to a truncated GMK estimate [14]) is:

$$\log \Omega^{\mathrm{BBK}}(\mathbf{r}, \mathbf{c}) = \log N! - \sum_{i=1}^{m} r_i! - \sum_{j=1}^{m} c_j! + \frac{2}{N^2}\sum_{i=1}^{m}\binom{r_i}{2}\sum_{j=1}^{n}\binom{c_j}{2}. \tag{A 18}$$

The last term can be written as

$$\begin{aligned}\frac{2}{N^2}\sum_{i=1}^{m}\binom{r_i}{2}\sum_{j=1}^{n}\binom{c_j}{2} &= \frac{1}{2N^2}\sum_{i=1}^{m}(r_i^2 - r_i)\sum_{j=1}^{n}(c_j^2 - c_j) \\ &= \frac{1}{2mn\hat{r}^{(1)}\hat{c}^{(1)}}\left(m\hat{r}^{(2)} - m\hat{r}^{(1)}\right)\left(n\hat{c}^{(2)} - n\hat{c}^{(1)}\right) = \frac{1}{2}\left(\frac{\hat{r}^{(2)}}{\hat{r}^{(1)}} - 1\right)\left(\frac{\hat{c}^{(2)}}{\hat{c}^{(1)}} - 1\right).\end{aligned} \tag{A 19}$$

From the sparse limit guarantee of this estimate, we have that the true $\log \Omega(\mathbf{r}, \mathbf{c})$ behaves in this limit as

$$\log \Omega(\mathbf{r}, \mathbf{c}) = \log N! - \sum_{i=1}^{m} r_i! - \sum_{j=1}^{m} c_j! + \frac{1}{2}\left(\frac{\hat{r}^{(2)}}{\hat{r}^{(1)}} - 1\right)\left(\frac{\hat{c}^{(2)}}{\hat{c}^{(1)}} - 1\right) + \mathrm{O}\left(N^{-1}\right). \tag{A 20}$$

We now demonstrate that the EC estimate has the same behavior in this limit, and hence that it is asymptotically equal to the true value of $\log \Omega(\mathbf{r}, \mathbf{c})$. We start from the definition of the EC estimate as

$$\Omega^{\mathrm{EC}}(\mathbf{r}, \mathbf{c}) = \binom{N + m\alpha_{\mathbf{c}} - 1}{m\alpha_{\mathbf{c}} - 1}^{-1} \prod_{i=1}^{m} \binom{r_i + \alpha_{\mathbf{c}} - 1}{\alpha_{\mathbf{c}} - 1} \prod_{j=1}^{n} \binom{c_j + m - 1}{m - 1}, \tag{A 21}$$

where

$$\begin{aligned} \alpha_{\mathbf{c}} &= \frac{N^2 - N + (N^2 - c^2)/m}{c^2 - N} = \frac{(n\hat{c}^{(1)})^2 - n\hat{c}^{(1)} + [(n\hat{c}^{(1)})^2 - n\hat{c}^{(2)}]/m}{n\hat{c}^{(2)} - n\hat{c}^{(1)}} \\ &= nS + \mathrm{O}(1), \end{aligned} \tag{A 22}$$

where

$$S = \frac{(\hat{c}^{(1)})^2}{\hat{c}^{(2)} - \hat{c}^{(1)}}. \tag{A 23}$$

Thus $\alpha_{\mathbf{c}}$ grows asymptotically as N. The quantity S can be understood as the factor by which the number of effective columns differs from the number of true columns n. The value of S is positive and finite in general, since $\hat{c}^{(1)} > 0$ and $\hat{c}^{(2)} \geq \hat{c}^{(1)}$, the only exception being when $\hat{c}_1 = 1$, i.e., when $c_j = 1$ for all j so that $\hat{c}^{(2)} = \hat{c}^{(1)}$. This, however, is precisely the degenerate case considered in Section (a) of this appendix, where we showed that the exact correct result is obtained in the $\alpha_{\mathbf{c}} \to \infty$ limit.

Given that $\alpha_{\mathbf{c}}$ is of order N, all of the binomial factors in the EC estimate benefit from the following expansion, derived by application of Stirling's approximation with $y \gg x$:

$$\begin{aligned} \log \binom{x+y}{y} &= \log(x+y)! - \log y! - \log x! \\ &= (x+y)\log(x+y) - (x+y) + \tfrac{1}{2}\log(2\pi(x+y)) + \frac{1}{12(x+y)} \\ &\quad - y\log y + y - \tfrac{1}{2}\log(2\pi y) - \frac{1}{12y} - \log x! + \mathrm{O}((x+y)^{-2}) + \mathrm{O}(y^{-2}) \\ &= (x+y)\left[\log y + \log(1 + x/y)\right] - y\log y + \tfrac{1}{2}\log(1 + x/y) \\ &\quad - x - \log x! + \mathrm{O}(xy^{-2}) \\ &= x\log y - \log x! + \frac{x(x+1)}{2y} + \mathrm{O}(x^2/y^2). \end{aligned} \tag{A 24}$$

Applying this approximation to the EC estimate, we have

$$\begin{aligned} \log \Omega^{\mathrm{EC}}(\mathbf{r}, \mathbf{c}) &= \log N! - \sum_{i=1}^{m} r_i! - \sum_{j=1}^{m} c_j! + \sum_{i=1}^{m} r_i \log(\alpha_{\mathbf{c}} - 1) + \tfrac{1}{2}\sum_{i=1}^{m} \frac{r_i(r_i+1)}{\alpha_{\mathbf{c}} - 1} \\ &\quad + \sum_{j=1}^{n} c_j \log(m-1) + \tfrac{1}{2}\sum_{j=1}^{n} \frac{c_j(c_j+1)}{m-1} \\ &\quad - N\log(m\alpha_{\mathbf{c}} - 1) - \frac{N(N+1)}{2(m\alpha_{\mathbf{c}} - 1)} + \mathrm{O}(N^{-1}) \\ &= \log N! - \sum_{i=1}^{m} r_i! - \sum_{j=1}^{m} c_j! + N\log\frac{(\alpha_{\mathbf{c}} - 1)(m-1)}{m\alpha_{\mathbf{c}} - 1} \\ &\quad + \frac{n(\hat{c}^{(2)} + \hat{c}^{(1)})}{2(m-1)} + \frac{m(\hat{r}^{(2)} + \hat{r}^{(1)})}{2(\alpha_{\mathbf{c}} - 1)} - \frac{N(N+1)}{2(m\alpha_{\mathbf{c}} - 1)} + \mathrm{O}(N^{-1}). \end{aligned} \tag{A 25}$$

Note that this implies that the sub-leading behavior of $\alpha_{\mathbf{c}}$ beyond order N is irrelevant for finding $\log \Omega^{\mathrm{EC}}$ to order N^{-1}, and hence that it is adequate to retain only the $\mathrm{O}(N)$ terms in Eq. (A 22).

Now, applying our expression for the leading behavior, we have

$$\begin{aligned}\log \Omega^{\mathrm{EC}}(\mathbf{r},\mathbf{c}) &= \log N! - \sum_{i=1}^{m} r_i! - \sum_{j=1}^{m} c_j! - N\frac{m+nS}{mnS} \\ &\quad + \frac{n(\hat{c}^{(2)}+\hat{c}^{(1)})}{2m} + \frac{m(\hat{r}^{(2)}+\hat{r}^{(1)})}{2nS} - \frac{N^2}{2mnS} + \mathrm{O}(N^{-1}). \\ &= \log N! - \sum_{i=1}^{m} r_i! - \sum_{j=1}^{m} c_j! - \frac{\hat{c}^{(1)}+\hat{r}^{(1)}S}{S} \\ &\quad + \frac{\hat{r}^{(1)}(\hat{c}^{(2)}+\hat{c}^{(1)})}{2\hat{c}^{(1)}} + \frac{\hat{c}^{(1)}(\hat{r}^{(2)}+\hat{r}^{(1)})}{2\hat{r}^{(1)}S} - \frac{\hat{r}^{(1)}\hat{c}^{(1)}}{2S} + \mathrm{O}(N^{-1}). \\ &= \log N! - \sum_{i=1}^{m} r_i! - \sum_{j=1}^{m} c_j! + \frac{1}{2}\left(\frac{\hat{r}^{(2)}}{\hat{r}^{(1)}} - 1\right)\left(\frac{\hat{c}^{(2)}}{\hat{c}^{(1)}} - 1\right) + \mathrm{O}(N^{-1}). \end{aligned} \tag{A 26}$$

Comparing this with the limiting form of $\log \Omega^{\mathrm{EC}}(\mathbf{r},\mathbf{c})$ in Eq. (A 20), we see that we have agreement in the limit:

$$\log \Omega^{\mathrm{EC}}(\mathbf{r},\mathbf{c}) = \log \Omega(\mathbf{r},\mathbf{c}) + \mathrm{O}(N^{-1}). \tag{A 27}$$

We also observe that this sparse-limit behavior is a result of the specific $\mathbf{c}$-dependence of $\alpha_{\mathbf{c}}$. If we repeat the same analysis for the GC estimate of Eq. (3.21), which is equivalent to our EC estimate but with $\alpha_{\mathbf{c}} = n$, we find a constant error:

$$\log \Omega^{\mathrm{GC}}(\mathbf{r},\mathbf{c}) = \log \Omega(\mathbf{r},\mathbf{c}) + \frac{1}{2}\left(\frac{\hat{r}^{(2)}}{\hat{r}^{(1)}} - 1 - \hat{r}^{(1)}\right)\left(\frac{\hat{c}^{(2)}}{\hat{c}^{(1)}} - 1 - \hat{c}^{(1)}\right) + \mathrm{O}(N^{-1}), \tag{A 28}$$

and hence the GC estimate of $\log \Omega(\mathbf{r},\mathbf{c})$ is not asymptotically equal to the true value.

(c) Effective columns estimate in the dense limit

In this section, we show that in the dense limit where the relative sizes of the row and column sums are fixed as $N \to \infty$, our effective columns estimate asymptotically agrees with the estimate of Diaconis and Efron (DE) [11]:

$$\Omega^{\mathrm{EC}}(\mathbf{r},\mathbf{c}) = \Omega^{\mathrm{DE}}(\mathbf{r},\mathbf{c})\left[1 + \mathrm{O}(N^{-1})\right]. \tag{A 29}$$

Specifically, we define rescaled versions of the row and column sums thus:

$$\tilde{r}_i = \frac{r_i}{N}, \qquad \tilde{c}_j = \frac{c_j}{N}. \tag{A 30}$$

If these are constant up to terms of order N^{-1} as $N \to \infty$ then Eq. (A 29) applies. This is a dense limit since the dimensions m and n are fixed, so the density N/mn goes to infinity.

To demonstrate this result we first consider the DE estimate in the form:

$$\begin{aligned}\log \Omega^{\mathrm{DE}}(\mathbf{r},\mathbf{c}) &= (m-1)(n-1)\log\left(N + \tfrac{1}{2}mn\right) + (K_{\mathbf{c}} - 1)\sum_{i=1}^{m} \log \bar{r}_i \\ &\quad + (m-1)\sum_{j=1}^{n} \log \bar{c}_j + \log \Gamma(mK_{\mathbf{c}}) - n\log\Gamma(m) - m\log\Gamma(K_{\mathbf{c}}), \end{aligned} \tag{A 31}$$

where

$$\begin{aligned} w &= \frac{N}{N + \frac{1}{2}mn}, \qquad \bar{r}_i = \frac{1-w}{m} + \frac{wr_i}{N}, \qquad \bar{c}_j = \frac{1-w}{n} + \frac{wc_j}{N}, \\ K_{\mathbf{c}} &= \frac{m+1}{m\bar{c}^2} - \frac{1}{m}, \qquad \bar{c}^2 = \sum_j \bar{c}_j^2. \end{aligned} \tag{A 32}$$

Only the leading constant behavior of these quantities is needed to find $\log \Omega^{\mathrm{DE}}(\mathbf{r},\mathbf{c})$ to $\mathrm{O}(N^{-1})$ in the dense limit we consider. We have:

$$\begin{aligned} w &= 1+\mathrm{O}(N^{-1}), \qquad \bar{r}_i = \tilde{r}_i + \mathrm{O}(N^{-1}), \qquad \bar{c}_j = \tilde{c}_j + \mathrm{O}(N^{-1}), \\ K_{\mathbf{c}} &= k_{\mathbf{c}} + \mathrm{O}(N^{-1}), \qquad k_{\mathbf{c}} = \frac{m+1}{m\sum_j \tilde{c}_j^2} - \frac{1}{m}, \end{aligned} \tag{A 33}$$

where we have defined $k_{\mathbf{c}}$. Substituting into Eq. (A 31), this gives

$$\begin{aligned} \log \Omega^{\mathrm{DE}}(\mathbf{r},\mathbf{c}) = (m-1)(n-1)\log N + (k_{\mathbf{c}}-1)\sum_{i=1}^{m}\log \tilde{r}_i + (m-1)\sum_{j=1}^{n}\log \tilde{c}_j \\ + \log\Gamma(mk_{\mathbf{c}}) - n\log\Gamma(m) - m\log\Gamma(k_{\mathbf{c}}) + \mathrm{O}(N^{-1}). \end{aligned} \tag{A 34}$$

Now we also compute the EC estimate in the same limit. First, we consider the parameter $\alpha_{\mathbf{c}}$, which we can write as

$$\alpha_{\mathbf{c}} = \frac{N-1+N\left(1-\sum_j \tilde{c}_j^2\right)/m}{N\sum_j \tilde{c}_j^2 - 1} = \frac{m+1}{m\sum_j \tilde{c}_j^2} - \frac{1}{m} + \mathrm{O}(N^{-1}) = k_{\mathbf{c}} + \mathrm{O}(N^{-1}). \tag{A 35}$$

Crucially, we observe that $\alpha_{\mathbf{c}}$ is constant in N up to terms of order N^{-1}. Now expanding the logarithms of binomials in the EC estimate using Eq. (A 24), we have:

$$\begin{aligned} \log \Omega^{\mathrm{EC}}(\mathbf{r},\mathbf{c}) &= -(m\alpha_{\mathbf{c}}-1)\log N + \sum_{i=1}^{m}\left[(\alpha_{\mathbf{c}}-1)\log(N\tilde{r}_i) - \log(\alpha_{\mathbf{c}}-1)!\right] \\ &\quad + \log(m\alpha_{\mathbf{c}}-1)! + \sum_{j=1}^{n}\left[(m-1)\log(N\tilde{c}_j) - \log(m-1)\right] + \mathrm{O}(N^{-1}) \\ &= \left[m(\alpha_{\mathbf{c}}-1) + n(m-1) - (m\alpha_{\mathbf{c}}-1)\right]\log N + (\alpha_{\mathbf{c}}-1)\sum_{i=1}^{m}\log\tilde{r}_i \\ &\quad + (m-1)\sum_{j=1}^{n}\log\tilde{c}_j + \log\Gamma(m\alpha_{\mathbf{c}}) - n\log\Gamma(m) - m\log\Gamma(\alpha_{\mathbf{c}}) + \mathrm{O}(N^{-1}) \\ &= (m-1)(n-1)\log N + (k_{\mathbf{c}}-1)\sum_{i=1}^{m}\log\tilde{r}_i + (m-1)\sum_{j=1}^{n}\log\tilde{c}_j \\ &\quad + \log\Gamma(mk_{\mathbf{c}}) - n\log\Gamma(m) - m\log\Gamma(k_{\mathbf{c}}) + \mathrm{O}(N^{-1}). \end{aligned} \tag{A 36}$$

Comparing with Eq. (A 34), we see that this agrees with the dense limit of the DE estimate and hence the EC and DE estimates are asymptotically equal in this limit. This agreement does not come as a surprise, given that the Dirichlet-multinomial distribution upon which the EC estimate is based is, in the dense limit, the same as the Dirichlet distribution that the DE estimate uses.

As with our earlier result for the sparse limit, the equivalence of the effective columns and DE estimates is fundamentally an effect of the $\mathbf{c}$-dependence of $\alpha_{\mathbf{c}}$. If $\alpha_{\mathbf{c}} = n$ as in the GC estimate, then there is no such equivalence and there is again a constant error between the estimates.

C. Numerical calculations

In this appendix we give some technical details of the numerical tests reported in Section 2.

(a) Generation of test cases

The process by which the test values of $\mathbf{r},\mathbf{c}$ are sampled for benchmarking can impact results like those in Fig. 1. In this section, we describe the scheme we use, explore the impact of using a different scheme, and examine the effect of changing the shape m,n of the matrix while keeping

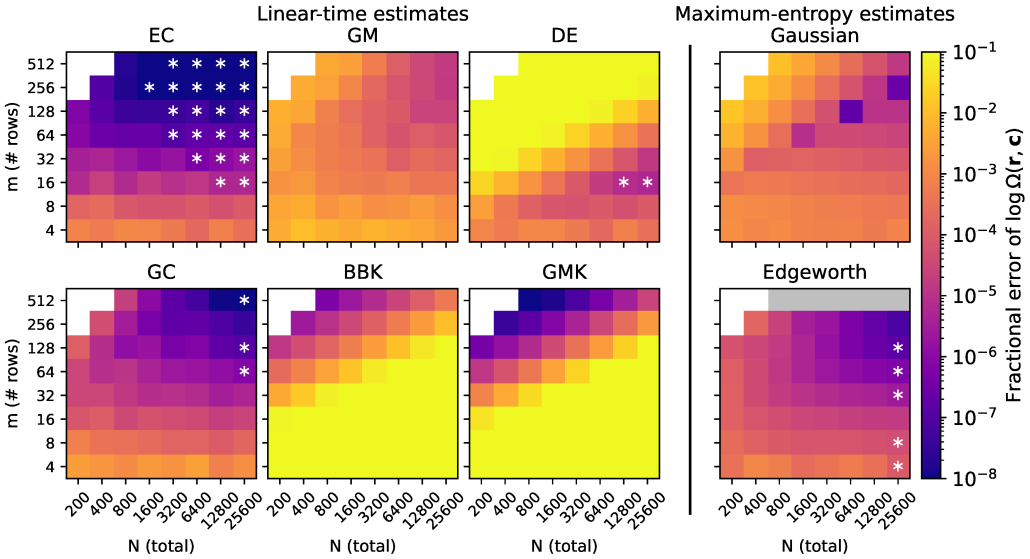

Figure 4. Fractional error of various estimates of $\log \Omega(\mathbf{r}, \mathbf{c})$. As in Fig. 1 the results are for square $m \times m$ matrices of total sum N. Unlike Fig. 1, however, the values of $\mathbf{r}, \mathbf{c}$ are the observed margins of uniformly sampled matrices, rather than being uniformly sampled themselves.

the sum N of all elements fixed. In all cases we find that our estimate for $\Omega(\mathbf{r}, \mathbf{c})$ appears to outperform other linear-time methods.

In generating the test cases we sample uniformly over possible margins $\mathbf{r}$ and $\mathbf{c}$ that have the required sizes m, n and sum to a given N. We also require that all r_i, c_j be nonzero, since cases with zeros can be trivially simplified by removing the zeros. Thus, for example, $\mathbf{r}$ is drawn uniformly from the set of all m-element vectors with strictly positive integer entries that sum to N.

There are other possible approaches, however. One could sample the margins by first generating a matrix, sampled uniformly from the set of non-negative integer $m \times n$ matrices that sum to N, and then take the row and column sums of this matrix to form $\mathbf{r}$ and $\mathbf{c}$. In effect, this process samples the margins $\mathbf{r}$ and $\mathbf{c}$ weighted by the number of possible matrices $\Omega(\mathbf{r}, \mathbf{c})$ with those margins. In practice this yields more uniform margins, particularly for larger and denser matrices, because there are larger numbers of matrices with relatively uniform margins than with non-uniform ones.

Making the margins more uniform typically improves the accuracy of estimates for $\Omega(\mathbf{r}, \mathbf{c})$, as shown in Fig. 4. Comparing to Fig. 1 we see that all of the estimates generally perform better for the more uniform margins. Our EC estimate, however, still stands out as performing particularly well and moreover is now competitive with the SIS benchmark and with the maximum-entropy methods for large N and m. In fact, when the margins are completely uniform our EC estimate (with $m \leq n$) appears numerically to always give results within the bounds conjectured by Canfield and McKay [19]. Collectively these observations suggest that the more uniform margins comprise the "easy cases" for approximating $\Omega(\mathbf{r}, \mathbf{c})$ and the more heterogeneous margins of Fig. 1 provide a more stringent test.

In Fig. 1 we also consider only square $m \times m$ matrices, since performance seems to be driven primarily by the value of N and the product of the dimensions mn. Figure 5 offers some evidence for this claim. In this figure we show the results of tests in which m and n are varied while keeping N fixed at a value of 1600, and we see that most of the performance is indeed explained by the combination mn—constant mn in this figure corresponds to diagonal lines from top-left to bottom-right. These patterns are also observed for other choices of N, although there

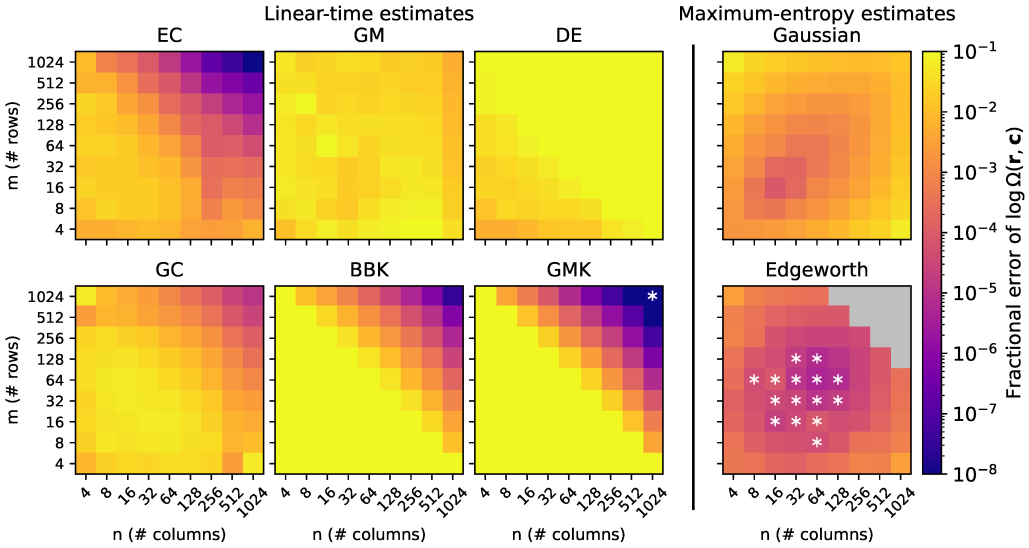

Figure 5. Fractional error of various estimates of $\log \Omega(\mathbf{r}, \mathbf{c})$. In these tests the totals of all matrices are fixed at $N = 1600$ while the numbers of rows and columns are varied. Each data point is averaged over ten sets of margins $\mathbf{r}, \mathbf{c}$ drawn uniformly at random. We observe that the performance of the linear-time estimates depends chiefly on the product mn, while the maximum-entropy estimates are best when $m \simeq n$ and poorer for highly oblong matrices.

are some deviations. The maximum-entropy estimates seem to have difficulty when m and n are very different, and some of the linear-time estimates (EC, GM, and DE) also show some mild asymmetry. By definition, the true $\Omega(\mathbf{r}, \mathbf{c})$ is symmetric under the interchange of $\mathbf{r}$ and $\mathbf{c}$, but asymmetries arise in the approximations. Based on the numerical evidence, we find that the EC estimate generally performs better for $m \leq n$, so if this is not the case we recommend interchanging the rows and columns before applying the estimate.

(b) Benchmarking

Benchmarking of our estimates requires us to compute accurate ground-truth values of $\Omega(\mathbf{r}, \mathbf{c})$ for comparison. In this section we describe various methods for doing this, and in particular address the following question: if you have one hour of computation time (on standard hardware *circa* 2022) to get the highest quality estimate of $\Omega(\mathbf{r}, \mathbf{c})$, what method should you use? Under these conditions, linear-time estimates never give the best answer (although in applications where speed is important, such as when estimating $\Omega(\mathbf{r}, \mathbf{c})$ for a large number of small matrices, linear-time methods may be the best).

Figure 6 summarizes our results for the best method to use as a function of m and N. In certain regimes exact solutions are available. Barvinok's algorithm for counting integer points in convex polytopes [7, 28] can be applied to give an exact algorithm with running time polynomial in N, which we implement using the `count` function from the `latte` software package [29]. This allows very large values of N to be probed, but the complexity grows quickly in m so this method is limited to $m \lesssim 6$ on current hardware.

In sparse situations with bounded margins $\mathbf{r}$ and $\mathbf{c}$ we can compute $\Omega(\mathbf{r}, \mathbf{c})$ exactly using recursion-based methods. Harrison and Miller [8] have given an implementation of this approach which exploits repeated entries in the margins to improve running time. While not shown in Fig. 6, this method can also be used for most cases where $N \lesssim 100$.

For all other cases, we use approximate ground-truth estimates of $\Omega(\mathbf{r}, \mathbf{c})$ computed using sequential importance sampling (SIS), and among the various SIS methods the EC-based method

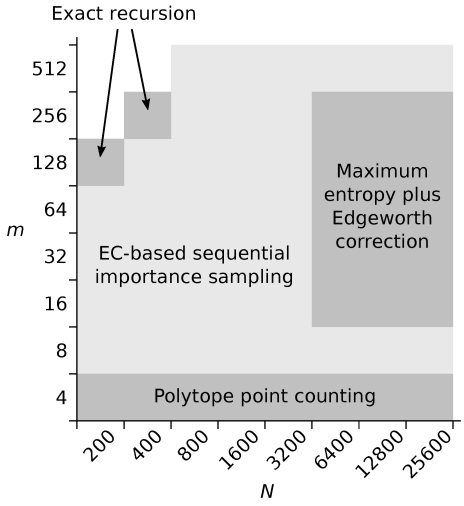

Figure 6. Schematic showing which methods give the best accuracy for estimating $\Omega(\mathbf{r}, \mathbf{c})$ in one hour of computer run time or less. The exact methods shown are Barvinok's polytope algorithm [7, 29] and the recursion-based approach of Harrison and Miller [8]. Where these are not applicable the Edgeworth-corrected maximum-entropy estimate is the winner for many large-N cases, while the EC-based sequential importance sampling approach of this paper is the method of choice for all the others. The empty region at the top left represents invalid parameter choices for which there are no matrices with the given margins. This diagram also represents the choice of method used to validate the SIS results in Fig. 2.

of this paper (Section 5) performs the best as shown in Fig. 2. In principle, the Edgeworth-corrected maximum-entropy method of Barvinok and Hartigan [15] (Section 4(b)) outperforms SIS in certain regimes as can be seen in Fig. 1, but this is not useful for our benchmarking since this is one of the approximations we are trying to evaluate. In a more general setting, however, where one simply wanted to make the best estimate of $\Omega(\mathbf{r}, \mathbf{c})$ in the time allotted, the maximum-entropy method could be useful.

References

1 Newman MEJ, Cantwell GT, Young JG. 2020 Improved mutual information measure for clustering, classification, and community detection. *Physical Review E* **101**, 042304.
2 Chen Y, Diaconis P, Holmes SP, Liu JS. 2005 Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association* **100**, 109–120.
3 Harrison MT, Miller JW. 2013 Importance sampling for weighted binary random matrices with specified margins. Arxiv preprint 1301.3928.
4 Eisinger RD, Chen Y. 2017 Sampling for conditional inference on contingency tables. *Journal of Computational and Graphical Statistics* **26**, 79–87.
5 Diaconis P, Gangolli A. 1995 Rectangular arrays with fixed margins. In Aldous D, Diaconis P, Spencer J, Steele JM, editors, *Discrete Probability and Algorithms* pp. 15–41, Springer, Berlin.
6 Dyer M, Kannan R, Mount J. 1997 Sampling contingency tables. *Random Structures & Algorithms* **10**, 487–506.
7 Barvinok AI. 1994 A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Mathematics of Operations Research* **19**, 769–779.
8 Miller JW, Harrison MT. 2013 Exact sampling and counting for fixed-margin matrices. *Annals of Statistics* **41**, 1569–1592.
9 Good IJ. 1976 On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics* **4**, 1159–1189.
10 Good IJ, Crook JF. 1977 The enumeration of arrays and a generalization related to contingency tables. *Discrete Mathematics* **19**, 23–45.

11 Diaconis P, Efron B. 1985 Testing for independence in a two-way table: New interpretations of the chi-square statistic. *Annals of Statistics* **13**, 845–874.
12 Gail M, Mantel N. 1977 Counting the number of $r \times c$ contingency tables with fixed margins. *Journal of the American Statistical Association* **72**, 859–862.
13 Békéssy A. 1972 Asymptotic enumeration of regular matrices. *Studia Scientiarum Mathematicarum Hungarica* **7**, 343–353.
14 Greenhill C, McKay BD. 2008 Asymptotic enumeration of sparse nonnegative integer matrices with specified row and column sums. *Advances in Applied Mathematics* **41**, 459–481.
15 Barvinok A. 2012 Matrices with prescribed row and column sums. *Linear Algebra and its Applications* **436**, 820–844.
16 Holmes RB, Jones LK. 1996 On uniform generation of two-way tables with fixed margins and the conditional volume test of Diaconis and Efron. *Annals of Statistics* **24**, 64–68.
17 Barvinok A. 2010 What does a random contingency table look like? *Combinatorics, Probability and Computing* **19**, 517–539.
18 Dittmer S, Lyu H, Pak I. 2020 Phase transition in random contingency tables with non-uniform margins. *Transactions of the American Mathematical Society* **373**, 8313–8338.
19 Canfield ER, McKay BD. 2007 Asymptotic enumeration of integer matrices with constant row and column sums. Arxiv preprint math/0703600.
20 Canfield ER, McKay BD. 2010 Asymptotic enumeration of integer matrices with large equal row and column sums. *Combinatorica* **30**, 655–680.
21 O'Neil PE. 1969 Asymptotics and random matrices with row-sum and column sum-restrictions. *Bulletin of the American Mathematical Society* **75**, 1276–1282.
22 Barvinok A, Hartigan JA. 2010 Maximum entropy Gaussian approximations for the number of integer points and volumes of polytopes. *Advances in Applied Mathematics* **45**, 252–289.
23 Gale D. 1957 A theorem on flows in networks. *Pacific Journal of Mathematics* **7**, 1073–1082.
24 Ryser HJ. 1957 Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics* **9**, 371–377.
25 Greenhill C, McKay BD, Wang X. 2006 Asymptotic enumeration of sparse 0–1 matrices with irregular row and column sums. *Journal of Combinatorial Theory, Series A* **113**, 291–324.
26 Canfield ER, Greenhill C, McKay BD. 2008 Asymptotic enumeration of dense 0–1 matrices with specified line sums. *Journal of Combinatorial Theory, Series A* **115**, 32–66.
27 Barvinok A, Hartigan JA. 2013 The number of graphs and a random graph with a given degree sequence. *Random Structures & Algorithms* **42**, 301–348.
28 Barvinok A, Pommersheim JE. 1999 An algorithmic theory of lattice points in polyhedra. *New Perspectives in Algebraic Combinatorics* **38**, 91–147.
29 Baldoni V, Berline N, De Loera JA, Dutra B, Köppe M, Moreinis S, Pinto G, Vergne M, Wu J. 2014 A user's guide for LattE integrale v1.7.2. *Optimization* **22**.