Direct Sampling from Conditional Distributions by Sequential Maximum Likelihood Estimations

Shuhei Mano **The Institute of Statistical Mathematics**

The direct sampling algorithm

Definition 1. Let $A = (a_{ij}) \in \mathbb{Z}^{d\times m}$ be a matrix of integers such that no row or column is zero vector and $(1,\dots,1) \in \mathrm{rowspan}(A)$. Let $x \in \mathbb{R}^m_{>0}$. The *log-affine model* associated with the *configuration matrix* A is the set of probability distributions

$$\mathcal{M}_A := \mathrm{cl}\{p \in \mathrm{int}(\Delta_{m-1}) : \log p \in \log x + \mathrm{rowspan}(A)\},$$

where Δ_{m-1} is the standard m-dimensional simplex. If $x = 1$, the model is called the *log-linear model.*

Consider a sample consisting of observations in which counts of the j-th state is u_j, $j \in [m]$. The conditional distribution given the complete minimal sufficient statistics b in the affine semigroup $\mathbb{N}A := \{Av : v \in \mathbb{N}^m\}$, where $\mathbb{N}$ is the set of non-negative integers, is

$$\mathbf{P}(U = u|AU = b) = \frac{1}{Z_A(b;x)}\frac{x^u}{u!}, \quad x^u := \prod_{j\in[m]} x_j^{u_j}, \quad u! := \prod_{j\in[m]} u_j!.$$

The support $\mathcal{F}_A(b) := \{v \in \mathbb{N}^m : Av = b\}$ is called the *fiber*. The normalization constant $Z_A(b;x) := \sum_{v\in\mathcal{F}_A(b)} x^v/v!$ is called the *A-hypergeometric polynomial* defined by Gel'fand, Kapranov, and Zelevinsky in the 1980's. We adopt the convention $Z_A(b;x) = 0$ if $b \notin \mathbb{N}A$.

The direct sampling algorithm [2017, EJS] is a Markov chain on

Definition 2. The *Markov lattice* $\mathcal{L}_A(b)$ is the bounded integer lattice embedded in $\mathbb{N}A$ equipped with the partial order

$$\beta \in \mathbb{N}A \text{ and } \beta - a_j \in \mathbb{N}A \quad \Rightarrow \quad \beta - a_j \prec \beta,$$

and the maximum and the minimum are b and 0, respectively. Here, a_j denotes the j-th column vector of A.

Example 1 (2×2 contingency table)**.**

$$\begin{array}{cc|c} u_{11} & u_{12} & u_{1\cdot} \\ u_{21} & u_{22} & u_{2\cdot} \\ \hline u_{\cdot 1} & u_{\cdot 2} & |u| \end{array}, \quad A = \begin{pmatrix} 1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1 \end{pmatrix}, \quad u = \begin{pmatrix} u_{11}\\u_{12}\\u_{21}\\u_{22} \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} u_{1\cdot}\\u_{2\cdot}\\u_{\cdot 1}\\u_{\cdot 2} \end{pmatrix},$$

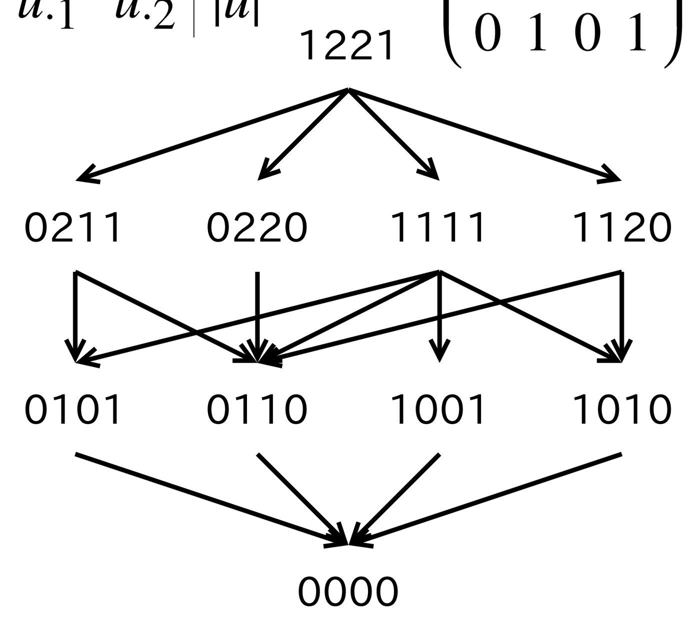

Figure 1: The Hasse diagram of the Markov lattice $\mathcal{L}_A(b)$ for the maximum $b = {}^t(1,2,2,1)$.

The Markov kernel is represented by the transition probability

$$P(\beta, \beta - a_j; x) := \frac{Z_A(\beta - a_j; x)}{Z_A(\beta; x)} \frac{x_j}{\deg(\beta)}, \quad j \in [m] \tag{1}$$

for the transition from β to $\beta - a_j$, where $\deg(\beta)$ is the degree of $Z_A(\beta;x)$. The computation of (1) dominates the computational cost. (1) is the ratio of the UMVUE (uniform minimum variance unbiased estimator) of the expected counts to the total number of counts if $\beta = b$.

The UMVUE and the rational MLE

The number of critical points of the likelihood over the complex space is one if and only if the MLE admits a specific form called the Horn uniformization of the A-discriminant [Huh 2014, J. Alg. Stat]. Moreover,

Theorem 1. [Duarte et al. 2021, Bernoulli] *A sample has the rational MLE if and only if there is a vector $\lambda \in \mathbb{R}^m$ and a matrix $H = (h_{ij}) \in \mathbb{Z}^{l\times m}$ such that $\mathcal{M}_A$ is the image of the rational map $\Phi : \mathbb{R}^m_{>0} \dashrightarrow \mathbb{R}^m_{>0}$:*

$$\Phi_j(u_1,\dots,u_m) = \lambda_j \prod_{i=1}^{l} \Big(\sum_{k=1}^{m} h_{ik}u_k\Big)^{h_{ij}} \quad \textit{with} \sum_{j\in[m]} \Phi_j(u) = 1.$$

Remark 1. [+Takayama, arXiv: 2110.14922] The rational MLE is an analogue of the Gauss hypergeometric theorem for A-hypergeom. polynomials.

Example 2 (2×2 contingency table, cont.)**.** We know the rational MLE of the expected count $\hat{\mu}(b;1) = |u|\Phi(u)$ with

$$\Phi(u) = \left(\frac{u_{1\cdot}u_{\cdot 1}}{|u|^2}, \frac{u_{1\cdot}u_{\cdot 2}}{|u|^2}, \frac{u_{2\cdot}u_{\cdot 1}}{|u|^2}, \frac{u_{2\cdot}u_{\cdot 2}}{|u|^2}\right),$$

and we can read off

$$\lambda = (4,4,4,4), \quad \text{and} \quad H = \begin{pmatrix} 1&1&0&0\\0&0&1&1\\-2&-2&-2&-2\\1&0&1&0\\0&1&0&1 \end{pmatrix}.$$

If we replace the UMVUE with the MLE, the transition probability becomes $\frac{\hat{\mu}(b;1)}{|u|} = \frac{u_{i\cdot}}{|u|}\frac{u_{\cdot j}}{|u|}$, where a_{ij} denotes the column vector of A cording the (i,j)-cell.

Theorem 2. [1] *The UMVUE and the MLE of a log-linear graphical model coincide if and only if the model is decomposable. Moreover, the MLE is rational.*

The approximate algorithm

Corollary 1. *For a log-linear model, the approximate algorithm obtained by replacing the UMVUE with the MLE becomes exact if and only if the model is a decomposable graphical model.*

Nevertheless, if β lie in the boundary of $\mathbb{N}A$ and $\beta - a_j \notin \mathbb{N}A$, the MLE $\hat{\mu}_j(\beta;x)$ coincides with the UMVUE and vanishes exactly without bias for any log-linear model. In practical terms, we have

Proposition 1. *The approximate algorithm is a Markov chain on $\mathcal{L}_A(b)$.*

Please see [1] why we can avoid computation of the Gröbner bases.

Ideal	Algorithm	States	Investigate	Reference
$I_A \subset k[p]$	Metropolis	$\mathcal{F}_A(b)$	Markov basis	Diaconis–Sturmfels† 1998, AS
$H_A(b) \subset \mathbb{C}\langle p\rangle\langle\partial\rangle$	Direct	$\mathcal{L}_A(b)$	Connection matrix	2017, EJS
	Approx.-Direct	$\mathcal{L}_A(b)$	None	[1]

Table 1: Summary of sampling algorithms.

Numerical experiments

For a no-three-way interaction model of three-way table of rank 81, the exact direct sampling was prohibitive [+Takayama, arXiv: 2110.14922].

$\lvert u\rvert$	(burn-in,length)= $(10^3, 10^4)$	$(10^4, 10^4)$	$(10^5, 10^4)$	$(10^5, 10^5)$
36	0.037	0.027	0.034	0.029
90	0.030	0.039	0.036	0.017
180	0.035	0.031	0.031	0.013

Table 2: The total variation distances between the empirical distributions of chi-square values generated by the Metropolis algorithm and that by the approximate direct sampling algorithm with 10^6 tables. 10-trials average.

$\lvert u\rvert$	ESS	Metropolis (sec.)	Approx.-Direct (sec.)
36	844	4.012	83.7
90	717	4.241	130.3
180	544	4.574	165.0

Table 3: The times to draw 10^4 tables by the Metropolis algorithm, the effective sample sizes (ESS) of the 10^4 tables, the times to draw the same number of tables by the approximate direct sampling algorithm. 100-trials average.

Reference

[1] Direct sampling from conditional distributions by sequential maximum likelihood estimations, preprint. arXiv: 2502.00812

 E-mail: smano@ism.ac.jp