Sampling a binary correlation matrix

June 17, 2023

Abstract

Key-words: multivariate Binary variables; simulation; Markov chain Monte Carlo; constrained optimisation; hit-and-run.

1 Problem description

Exposition here will draw mostly from Leisch et al. (1998) and Schäfer (2010). See also section 3.6.2 (pp 69) in Schäfer (2012).

Let $\boldsymbol{X} \in \{0, 1\}^d$ be a d-dimensional binary random variable. We will say that $\boldsymbol{R}$ is the correlation matrix associated with $\boldsymbol{X}$ if

$$\boldsymbol{R}_{ij} = E[(X_i - m_i)(X_j - m_j)]/\sqrt{m_i(1-m_i)m_j(1-m_j)},$$

where $m_k := E[X_k]$, $k = 1, \dots, d$. Writing $p_{ij} = \Pr(X_i = 1, X_j = 1)$ for the joint probability, we can re-write the correlation as

$$\boldsymbol{R}_{ij} = \frac{p_{ij} - m_i m_j}{\sqrt{m_i(1-m_i)m_j(1-m_j)}}. \tag{1}$$

For a fixed vector of marginal probabilities $\boldsymbol{m} = \{m_1, \dots, m_d\}$, the structure of the problem induces a number of constraints on the possible values of joint probabilities p_{ij}, which we will detail shortly. Consider the space of all cross-moment matrices with entries on $(0, 1)$ that are compatible with $\boldsymbol{m}$:

$$\mathcal{A}(\boldsymbol{m}) = \{\boldsymbol{A} \in \mathcal{M}_{d\times d} : \boldsymbol{A}_{ij} \in (0,1), A_{kk} = m_k\},$$

for $k = 1, \dots, d$. It turns out that not all members of $\mathcal{A}(\boldsymbol{m})$ will yield valid joint probability matrices – see Section 3.3 in Leisch et al. (1998). There are two sets of constraints that a matrix in $\mathcal{A}(\boldsymbol{m})$ will need to fulfill:

- **Pairwise constraints** (C1): $\max(m_i + m_j - 1, 0) \leq \boldsymbol{A}_{ij} \leq \min(m_i, m_j)$ for every pair $(i, j) \in \{1, \dots, d\}^2$;
- **Triplet-wise constraints** (C2): $m_i + m_j + m_k - 1 \leq \boldsymbol{A}_{ij} + \boldsymbol{A}_{ik} + \boldsymbol{A}_{jk}$ for all triplets $(i, j, k) \in \{1, \dots, d\}^3$.

Note that satisfying C1 is not sufficient; C1 and C2 need to be satisfied simultaneously. Our goal is to sample a random variable defined on the space

$$\mathcal{Q}(\boldsymbol{m}) = \{\boldsymbol{Q} \in \mathcal{A}(\boldsymbol{m}) : \text{C1 } \underline{\text{and}} \text{ C2 are satisfied}\}.$$

It is clear that once we can produce a sample $\boldsymbol{Q} \in \mathcal{Q}(\boldsymbol{m})$, the expression (1) can be used entry-wise to produce a sample correlation matrix $\boldsymbol{R}$. For an alternative representation of the constraints, see Proposition 3.2.1 (pp 46) in Schäfer (2012).

Schäfer (2010) suggests one could sample the correlation matrix directly by finding $\boldsymbol{R}$ such that

$$\boldsymbol{A} = \boldsymbol{R} \cdot \boldsymbol{s}\boldsymbol{s}^T + \boldsymbol{m}\boldsymbol{m}^T,$$

where $s_i = m_i(1 - m_i)$.

Problem: Design a simulation-efficient algorithm to draw from a uniform measure on $\mathcal{Q}(\boldsymbol{m})$, for any $\boldsymbol{m} \in (0,1)^d$.

From C1, we can get some crude triplet-wise bounds:

$$\boldsymbol{A}_{ij} + \boldsymbol{A}_{ik} + \boldsymbol{A}_{jk} \leq \min(p_i, p_j) + \min(p_i, p_k) + \min(p_j, p_k),$$

for all triplets $(i, j, k) \in \{1, \ldots, d\}^3$.

Compare with the algorithm proposed in Section 3.6.2. of Schäfer (2012), with the proviso that thei algorithm samples $\boldsymbol{m}$ uniformly, whereas here we are interested in sampling *conditional* on a given set of marginals $\boldsymbol{m}$.

References

Leisch, F., Weingessel, A., and Hornik, K. (1998). On the generation of correlated artificial binary data.

Schäfer, C. (2010). Parametric families for Monte Carlo on binary spaces. *arXiv preprint arXiv:1008.0055*.

Schäfer, C. (2012). *Monte Carlo methods for sampling high-dimensional binary vectors*. PhD thesis, Université Paris Dauphine-Paris IX.