SUPPLEMENT B TO “BAYESIAN COMPLEMENTARY CLUSTERING, MCMC AND ANGLO-SAXON PLACENAMES”: COMPUTATIONAL COMPLEXITY OF THE MODEL

By Giacomo Zanella[*]

University of Warwick

We provide some results and references coming from the complexity theory literature regarding the intractability of the model described in Section 3.6 of Zanella (2014).

We consider the full conditional distribution $\pi(\rho|\mathbf{x}, \sigma, \mathbf{p}^{(c)}, \lambda)$ arising from the model described in Section 3.6 of Zanella (2014). As in Zanella (2014), we will denote $\pi(\rho \mid \mathbf{x}, \sigma, \mathbf{p}^{(c)}, \lambda)$ by $\hat{\pi}(\rho)$.

In Section 3.6.1 of Zanella (2014) we described $\hat{\pi}(\rho)$ in terms of a hypergraph construction. There we show that every partition ρ that is admissible for our model can be interpreted as a partial matching. In such a way the sample space of $\hat{\pi}(\rho)$ can be seen as the space of partial matchings contained in G, where $G = (V, E)$ is the complete k-partite hypergraph induced by k sets $V_1, \dots, V_k$ (see Section 3.6.1 of Zanella (2014) for more details and definitions of those terms). Moreover $\hat{\pi}(\rho)$ is proportional to the weight of the matching ρ, defined as $\prod_{e\in\rho} w(e)$, where $w(e)$ is the weight of the hyperedge $e = \{x_1, \dots, x_s\} \in E$ defined in (3.7) of Zanella (2014)

1. Computational complexity of $\hat{\pi}(\rho)$.

We consider the complexity of the following tasks: finding the normalizing constant of $\hat{\pi}(\rho)$, finding the mode $\rho_{max} = \arg\max_{\rho\in\mathcal{P}_n} \hat{\pi}(\rho)$ and sampling from $\hat{\pi}(\rho)$.

We will distinguish between the two-color case ($k = 2$) and the multi-color case ($k \geq 3$) because they present substantially different complexity issues. Motivated by the corresponding literature in the theory of algorithms we often refer to those as two-dimensional and k-dimensional case.

1.1. *Finding the normalizing constant.*

The normalizing constant of $\hat{\pi}(\rho)$ is the sum of the weights of all the matchings ρ contained in G, that is the total weight of G. The problem of computing the total weight of a k-partite hypergraph is an $\#P$-hard counting problem (Valiant, 1979), even in the easiest version where $k = 2$ and the edge weights can only be 0 or

[*]Supported by EPSRC through a PhD position under the CRiSM grant EP/D002060/1

1 (i.e. counting the number of partial matchings in a bipartite graph). The $\#P$-hard complexity class for counting problems is analogous to the NP-hard complexity class for decision problems. A counting problem y is said to be $\#P$-hard if and only if every problem in $\#P$ (i.e. every polynomially checkable counting problem) is Cook-reducible to y (see Valiant (1979) for definitions of these terms).

1.2. *Finding the posterior mode.*

Finding $\rho_{max} = \text{argmax}_\rho \, \hat{\pi}(\rho)$ can be reduced to a k-dimensional optimal assignment problem as follows.

Problem 1. *(k-dimensional optimal assignment problem)*
Instance: k sets I_1,...,I_k of size n and a cost function $C : I_1 \times \cdots \times I_k \to \mathbb{R}$.
Problem: find an assignment A, i.e. a subset $A \subseteq I_1 \times \cdots \times I_k$ containing each point of I_1,...,I_k exactly once, that minimizes $\sum_{(i_1,\ldots,i_k)\in A} C(i_1,\ldots,i_k)$.

First note that $\rho_{max} = \text{argmax}_\rho \sum_{e\in\rho} \log(w(e))$, where ρ belongs to the set of matchings of $G = (V, E)$. Suppose now that V is made of $n_1,\ldots,n_k$ vertices of colors $1,\ldots,k$ respectively. For each color i add $n - n_i$ auxiliary points that do not contribute to the weight of any edge. Each partial matching ρ of G can then be seen as a complete matching $\tilde{\rho}$ (i.e. a matching connecting all the vertices in V) in the augmented version of G such that ρ and $\tilde{\rho}$ have the same weight. Expressing $\tilde{\rho}$ as an assignment A we obtain Problem 1.

For $k = 2$ this problem is efficiently solvable, in $O(n^3)$ steps, using the Hungarian Algorithm (Kuhn, 1955), which is based on concepts from Optimal Transportation Theory (Villani, 2009). In contrast for $k \geq 3$ this is an NP-hard optimization problem. Even more, unless P=NP, there is no deterministic polynomial-time approximation algorithm for a general cost function (i.e. the problem is not in APX). The same holds even if the cost function C is decomposable as $C(x_1,\ldots,x_k) = \sum_{i\neq j} d(x_i, x_j)$. Some polynomial time approximation algorithms exist if d satisfies the triangle inequality, but this is not our case (see, for example, Crama and Spieksma (1992) and Bandelt, Crama and Spieksma, 1994). Balas and Saltzman (1991) propose an heuristic algorithm for a general cost function C, but no constant of approximation is provided and only the case $k = 3$ is considered.

Finally De la Vega et al. (2003) propose a polynomial time approximation scheme to partition n points of $\mathbb{R}^d$ in m clusters that minimize the sum of the intra-clusters squared Euclidean distances. This problem is similar to ours but unfortunately the running time of their algorithm is polynomial in n but exponential in m and in our context it seems reasonable to suppose m to be of the same order of n in magnitude.

In conclusion the literature does not appear to provide a generic bounded-complexity method to obtain (or approximate) ρ_{max}. We might therefore resort to heuristics suited to our particular case.

1.3. *Approximate sampling.* In Statistical Physics a monomer-dimer system is a collection of n sites covered by molecules occupying one site, monomers, or two sites, dimers. It can be described with the following model.

Model 1. *(monomer-dimer system)*
Instance: A finite undirected graph $G = (V, E)$ with non negative edge weights $w : E \to [0, \infty)$ such that $w(e) > 0$ for at least one $e \in E$.
State space: the set of partial matchings $\rho \subseteq E$.
Probability distribution: $\hat{\pi}(\rho) \propto \prod_{e \in \rho} w(e)$.

Although monomer-dimer system are usually considered in lattice frameworks, the two-dimensional version of our model can be interpreted as a monomer-dimer system (see Section 3.6.1 of Zanella, 2014). Jerrum and Sinclair (1996) propose a Metropolis-Hastings (MH) random walk algorithm to obtain approximate samples from monomer-dimer systems distributions in polynomial time. Using a canonical paths argument they prove that for any starting state ρ the mixing time of their Markov Chain satisfy

$$\tau_\rho(\epsilon) \leq 4(\#E)(\#V)w'^2 \left(\log(\#E)\#E + \log\left(\epsilon^{-1}\right)\right), \tag{1.1}$$

where $w' = \max\{1, \max_{e \in E} w(e)\}$. Huber and Law (2012) consider the same Markov Chain starting from the mode ρ_{max} (which can be found in $O(\#V)^3$ by the Hungarian algorithm) and improve slightly the bound (1.1) to

$$\tau_{\rho_{max}}(\epsilon) \leq 4(\#E)(\#V)w'^2 \left(\log(2)\#E + \log\left(\epsilon^{-1}\right)\right). \tag{1.2}$$

Remark 1. *The bounds (1.1) and (1.2) seem to be very conservative in practice. For example in the framework of Section 4.1.2 of Zanella (2014) the bound in (1.2) is of order 10^9 (depending weakly on ϵ). Convergence diagnostic methods, though, suggest that order 10^5 steps are enough to approximate $\hat{\pi}$.*

Can we approximately sample from $\hat{\pi}(\rho)$ in polynomial time for $k \geq 3$ too? This is equivalent to approximately count matchings in hypergraphs in polynomial time (see Chapter 3 of Jerrum (2003) for the relationship between approximate sampling and approximate counting). Unfortunately, as far as we are aware, there are not many results in this field. Karpinski, Rucinski and Szymanska (2012) try to extend the methods of Jerrum and Sinclair

(1996) to a hypergraph setting but they managed to do it only for a specific class of sparse hypergraphs, that do not include our case. They also prove a negative result: unless NP=RP (RP is the analogous of P but for randomized decision algorithms), there cannot be any FPRAS (Fully Polynomial Random Approximation Scheme, see for example Jerrum, 2003) to obtain approximate samples from the k-dimensional version of the monomer-dimer system for $k \geq 6$ (see Proposition 3 of Karpinski, Rucinski and Szymanska, 2012). Strictly speaking, this still does not imply that such a scheme cannot exist for our problem, because our problem is constrained by additional conditions (e.g. our hypergraph is k-partite).

References.

Balas, E. and Saltzman, M. (1991). An algorithm for the three-index assignment problem. *Operations Research* **39** 150–161.

Bandelt, H., Crama, Y. and Spieksma, F. (1994). Approximation algorithms for multi-dimensional assignment problems with decomposable costs. *Discrete Applied Mathematics* **49** 25–50.

Crama, Y. and Spieksma, F. (1992). Approximation algorithms for three-dimensional assignment problems with triangle inequalities. *European Journal of Operational Research* **60** 273–279.

De la Vega, W., Karpinski, M., Kenyon, C. and Rabani, Y. (2003). Approximation Schemes for Clustering Problems in Finite Metrics and High Dimensional Spaces. *Proceedings on the thirty-fifth annual ACM symposium on Theory of Computing* 50–58.

Huber, M. and Law, J. (2012). Simulation reduction of the Ising model to general matchings. *Electronic Journal of Probability* **17** 1–15.

Jerrum, M. (2003). *Counting, Sampling and Integrating: Algorithms and Complexity. Lectures in Mathematics ETH Zürich.* Birkhäuser Verlag, Basel.

Jerrum, M. and Sinclair, A. (1996). The Markov chain Monte Carlo method: an approach to approximate counting and integration. *Approximation algorithms for NP-hard problems* 482–520.

Karpinski, M., Rucinski, A. and Szymanska, E. (2012). Approximate Counting of Matchings in Sparse Uniform Hypergraphs. *arXiv preprint arXiv:1204.5335* 1–13.

Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly* **2** 83–97.

Valiant, L. (1979). The complexity of enumeration and reliability problems. *SIAM Journal on Computing* **8**.

Villani, C. (2009). *Optimal transport: old and new* **338**. Springer.

Zanella, G. (2014). Bayesian Complementary Clustering, MCMC and Anglo-Saxon placenames. *arXiv preprint arXiv:1409.6994*.

Department of Statistics
University of Warwick
Coventry, CV4 7AL
United Kingdom
E-mail: g.zanella@warwick.ac.uk