The Conditional-Potts Clustering Model: Supplementary Material

Alejandro Murua¹, and Nicolas Wicker²

¹Département de mathématiques et de statistique, Université de Montréal,
CP 6128, succ. centre-ville, Montréal, Québec H3C 3J7 Canada

²Université Lille 1 - UFR de Mathématiques,
Cité Scientifique, 59655 Villeneuve d'Ascq, France

The supplementary material shown here consists on the proof of the convergence theorem for PMC2 (Section 1), the proof of the probability bound theorem used to construct an informative data-driven prior for the temperature parameter of the conditional-Potts model (Section 2), and a more detailed description of the data sets used in our simulations (Section 3).

1 Proof of the convergence Theorem

Let $\theta = (\sigma, T) \in \Theta$. We will work with $\tilde{\pi}_m(\theta) \propto \pi_m(\theta) \times \frac{\sum_{g=1}^d e^{c_{\beta;m}(g)}}{\sum_{g=1}^d e^{c_m(g)}}$. Note that this does not change the target distribution since the last term does not depend on θ . Let $A \subset \Theta$ be a Borel set. Consider the proposal kernel

$$P_m(\theta, A) = \int_A \min \left\{ 1, \frac{\tilde{\pi}_m(\theta') q_m(\theta|\theta')}{\tilde{\pi}_m(\theta) q_m(\theta'|\theta)} \right\} q_m(\theta'|\theta) d\theta',$$

whose invariant probability distribution is exactly $\tilde{\pi}_m(\theta)$.

Note that since $\beta(\theta)$ is a continuous function of θ , it is bounded in Θ . Let β_{min}, β_{max} be lower and upper bounds respectively. Also, by assumption there exist $T_{min} > 0$ such that $T > T_{min}$ for all T . Moreover T is bounded from above. Hence, there exist constants $0 < K_{min} < K_{max}$ such that $K_{min} < k_{ij}(\sigma)/T < K_{max}$ for all $(\sigma, T) \in \Theta$. These easily imply

$$\begin{aligned} \exp\{n(\beta_{min} - \beta_{max})\} \kappa_{min} &\leq \frac{Z_{\beta;m}(\theta)}{\sum_{g=1}^d e^{c_{\beta;m}(g)}} \leq \exp\{n(\beta_{max} - \beta_{min})\} \kappa_{max}, \\ \exp\{n(K_{min} - K_{max})\} \kappa_{min} &\leq \frac{Z_m(\theta)}{\sum_{g=1}^d e^{c_m(g)}} \leq \exp\{n(K_{max} - K_{min})\} \kappa_{max}. \end{aligned}$$

Note that for a k -nearest-neighbor graph, the number of data points n could be replaced by the number of neighbors k in the exponents. From these inequalities, we obtain

$$\tilde{\pi}_m(\theta) > \frac{\exp\{n(K_{min} - K_{max})\} \kappa_{min}}{\exp\{n(\beta_{max} - \beta_{min})\} \kappa_{max}} C_{min}, \quad (1)$$

$$\tilde{\pi}_m(\theta) < \frac{\exp\{n(K_{max} - K_{min})\} \kappa_{max}}{\exp\{n(\beta_{min} - \beta_{max})\} \kappa_{min}} C_{max} \quad (2)$$

for all $\theta \in \Theta$, where C_{min} and C_{max} are lower and upper bounds for the prior (assumption (P4)). Therefore, (1), (2) and assumption (P2) implies the existence of a constant $\epsilon > 0$ independent of θ and m such that

$$P_m(\theta, A) \geq \epsilon \int_A d\theta'. \quad (3)$$

By Theorem 8 of Roberts and Rosenthal (2004),

$$\|P_m^j(\theta', \cdot) - \tilde{\pi}_m(\cdot)\|_{TV} \leq (1 - \epsilon)^j$$

for all j , where TV stands for the “Total Variation” norm. This shows that assumption (A3) of Theorem 2.1 of Atchade et al. (2010) holds. Also, (3) is exactly the condition (B3) of Theorem 3.1 of Atchade et al. (2010). Moreover, the existence of the bounds $K_{min}, K_{max}, \beta_{min}, \beta_{max}$ is equivalent to the assumption (B1) of Theorem 3.1 of Atchade et al. (2010). These conditions together with (P1) implies that Proposition 3.1 and Theorem 3.1 of Atchade et al. (2010) also hold. In particular, $e^{c_m(g)} / \sum_{g'=1}^d e^{c_m(g')}$ converges a.s. to $C_1 Z(\sigma^g, T^g, X)$, and $e^{c_{\beta;m}(g)} / \sum_{g'=1}^d e^{c_{\beta;m}(g')}$ converges a.s. to $C_2 Z(\sigma^g, T^g)$. These imply (see the proof of Theorem 3.1 in (Atchade et al. 2010)) that

$$\tilde{\pi}_m(\sigma, T) \rightarrow \frac{Z(\sigma, T, X)}{Z(\sigma, T)} \pi(T|\sigma) \pi(\sigma) = \pi(\sigma, T|X)$$

a.s. as $m \rightarrow +\infty$. Moreover, by (2) and the Lebesgue dominated convergence theorem, for any bounded measurable function $h : \Theta \rightarrow \mathbf{R}$,

$$\int h(\theta) \tilde{\pi}_m(\theta) d\theta \rightarrow \int h(\theta) \pi(\theta|X) d\theta$$

as m goes to $+\infty$. This shows that assumption (A1) of Theorem 2.1 of Atchade et al. (2010) also holds. To finish the proof, we just need to show that assumption (A2) of Theorem 2.1 of Atchade et al. (2010) holds. That is, we need to show that $\sup_{\theta \in \Theta} \|P_{m+1}(\theta, \cdot) - P_m(\theta, \cdot)\|_{TV} \rightarrow 0$ a.s. as $m \rightarrow +\infty$. Let $h : \Theta \rightarrow \mathbf{R}$ be a measurable function such that $|h(\theta)| \leq 1$. By definition, $P_m h(\theta) = \int \min\left\{1, \frac{\tilde{\pi}_m(\theta') q_m(\theta|\theta')}{\tilde{\pi}_m(\theta) q_m(\theta'|\theta)}\right\} h(\theta') q_m(\theta'|\theta) d\theta'$. We have

$$\begin{aligned} |P_{m+1} h(\theta) - P_m h(\theta)| &= \left| \int \left(\min\left\{1, \frac{\tilde{\pi}_{m+1}(\theta') q_{m+1}(\theta|\theta')}{\tilde{\pi}_{m+1}(\theta) q_{m+1}(\theta'|\theta)}\right\} q_{m+1}(\theta'|\theta) - \right. \right. \\ &\quad \left. \left. \min\left\{1, \frac{\tilde{\pi}_m(\theta') q_m(\theta|\theta')}{\tilde{\pi}_m(\theta) q_m(\theta'|\theta)}\right\} q_m(\theta'|\theta) \right) h(\theta') d\theta' \right| \\ &\leq \left| \int \left(\min\left\{1, \frac{\tilde{\pi}_{m+1}(\theta') q_{m+1}(\theta|\theta')}{\tilde{\pi}_{m+1}(\theta) q_{m+1}(\theta'|\theta)}\right\} - \min\left\{1, \frac{\tilde{\pi}_m(\theta') q_{m+1}(\theta|\theta')}{\tilde{\pi}_m(\theta) q_{m+1}(\theta'|\theta)}\right\} \right) h(\theta') q_{m+1}(\theta'|\theta) d\theta' \right. \\ &\quad \left. + \int \left(\min\left\{1, \frac{\tilde{\pi}_m(\theta') q_{m+1}(\theta|\theta')}{\tilde{\pi}_m(\theta) q_{m+1}(\theta'|\theta)}\right\} - \min\left\{1, \frac{\tilde{\pi}_m(\theta') q_m(\theta|\theta')}{\tilde{\pi}_m(\theta) q_m(\theta'|\theta)}\right\} \right) h(\theta') q_{m+1}(\theta'|\theta) d\theta' \right. \\ &\quad \left. + \int \min\left\{1, \frac{\tilde{\pi}_m(\theta') q_m(\theta|\theta')}{\tilde{\pi}_m(\theta) q_m(\theta'|\theta)}\right\} \left(q_{m+1}(\theta'|\theta) - q_m(\theta'|\theta) \right) h(\theta') d\theta' \right| \\ &\leq \left| \int \left[\left(\frac{\tilde{\pi}_{m+1}(\theta')}{\tilde{\pi}_{m+1}(\theta)} - \frac{\tilde{\pi}_m(\theta')}{\tilde{\pi}_m(\theta)} \right) q_{m+1}(\theta|\theta') + \left(\frac{q_{m+1}(\theta|\theta')}{q_{m+1}(\theta'|\theta)} - \frac{q_m(\theta|\theta')}{q_m(\theta'|\theta)} \right) \frac{\tilde{\pi}_m(\theta')}{\tilde{\pi}_m(\theta)} q_{m+1}(\theta'|\theta) \right] h(\theta') d\theta' \right. \\ &\quad \left. + \int \min\left\{1, \frac{\tilde{\pi}_m(\theta') q_m(\theta|\theta')}{\tilde{\pi}_m(\theta) q_m(\theta'|\theta)}\right\} \left(q_{m+1}(\theta'|\theta) - q_m(\theta'|\theta) \right) h(\theta') d\theta' \right|. \end{aligned}$$

Therefore, under the assumptions of the theorem, there exists a finite constant C_3 such that

$$\begin{aligned} |P_{m+1} h(\theta) - P_m h(\theta)| &\leq C_3 \left(\int \left| \frac{\tilde{\pi}_{m+1}(\theta')}{\tilde{\pi}_{m+1}(\theta)} - \frac{\tilde{\pi}_m(\theta')}{\tilde{\pi}_m(\theta)} \right| d\theta' \right. \\ &\quad \left. + \int \left| \frac{q_{m+1}(\theta|\theta')}{q_{m+1}(\theta'|\theta)} - \frac{q_m(\theta|\theta')}{q_m(\theta'|\theta)} \right| d\theta' + \int |q_{m+1}(\theta'|\theta) - q_m(\theta'|\theta)| d\theta' \right). \end{aligned}$$

After some more algebra, using the bounds previously found, we have

$$\begin{aligned} \sup_{\theta \in \Theta} |P_{m+1}h(\theta) - P_m h(\theta)| &\leq C_4 \int |\tilde{\pi}_{m+1}(\theta') - \tilde{\pi}_m(\theta')| d\theta' + C_5 \sup_{\theta \in \Theta} |\tilde{\pi}_{m+1}(\theta) - \tilde{\pi}_m(\theta)| \\ &\quad + C_6 \sup_{\theta \in \Theta} \left(\int |q_{m+1}(\theta'|\theta) - q_m(\theta'|\theta)| d\theta' + \int |q_{m+1}(\theta|\theta') - q_m(\theta|\theta')| d\theta' \right). \end{aligned}$$

The first term goes to zero by the Lebesgue dominated convergence theorem. The last term goes to zero by hypothesis. It only remains to proof that the second term goes to zero as well. Let $\tilde{Z}_m(\theta) = Z_m(\theta) / \sum_{g=1}^d e^{c_m(g)}$, and $\tilde{Z}_{\beta;m}(\theta) = Z_{\beta;m}(\theta) / \sum_{g=1}^d e^{c_{\beta;m}(g)}$. We have

$$\begin{aligned} |\tilde{\pi}_{m+1}(\theta) - \tilde{\pi}_m(\theta)| &= \left| \frac{\tilde{Z}_{m+1}(\theta)}{\tilde{Z}_{\beta;m+1}(\theta)} - \frac{\tilde{Z}_m(\theta)}{\tilde{Z}_{\beta;m}(\theta)} \right| \pi(\theta) \\ &\leq C_7 |\tilde{Z}_{m+1}(\theta) \tilde{Z}_{\beta;m}(\theta) - \tilde{Z}_m(\theta) \tilde{Z}_{\beta;m+1}(\theta)| \\ &\leq C_8 |\tilde{Z}_{m+1}(\theta) - \tilde{Z}_m(\theta)| + C_9 |\tilde{Z}_{\beta;m+1}(\theta) - \tilde{Z}_{\beta;m}(\theta)| \end{aligned}$$

Using exactly the same arguments in the proof of Theorem 3.1 of Atchade et al. (2010), one concludes that

$$\sup_{\theta \in \Theta} |\tilde{Z}_{m+1}(\theta) - \tilde{Z}_m(\theta)| \leq C_{10} \left(\gamma_{m+1} + \sum_{g=1}^d \left[\sum_{a=1}^{m+1} \mathbf{1}_g(I^a) \right]^{-1} \right),$$

and

$$\sup_{\theta \in \Theta} |\tilde{Z}_{\beta;m+1}(\theta) - \tilde{Z}_{\beta;m}(\theta)| \leq C_{11} \left(\gamma_{\beta;m+1} + \sum_{g=1}^d \left[\sum_{a=1}^{m+1} \mathbf{1}_g(I_{\beta}^a) \right]^{-1} \right).$$

By the Wang-Landau algorithm's property (Proposition 3.1 of Atchade et al. (2010)),

$$\frac{1}{m} \sum_{a=1}^m \mathbf{1}_g(I^a) \rightarrow \frac{1}{d} \text{ a.s., and } \frac{1}{m} \sum_{a=1}^m \mathbf{1}_g(I_{\beta}^a) \rightarrow \frac{1}{d} \text{ a.s.}$$

Therefore, since both sequences γ_m and $\gamma_{\beta,m}$ goes to zero a.s., we conclude that $\sup_{\theta \in \Theta} |\tilde{\pi}_{m+1}(\theta) - \tilde{\pi}_m(\theta)|$ goes to zero a.s.

2 Proof of the probability bound theorem

p_G is such that

$$p_G \leq \frac{1}{|G(n, r)|} \sum_{k=r+1}^{n/2} \binom{n}{k} \kappa(G(k, r)) |G(n-k, r)|,$$

where $\kappa(G(k, r))$ stands for the number of spanning trees in r -regular graphs. Note that in order to have a component of order k , k vertices need to be selected and connected by a spanning tree. Using the bound (Biggs 1994, page 42)

$$\kappa(G(k, r)) \leq \frac{1}{k} \left(\frac{kr}{k-1} \right)^{k-1},$$

it follows that

$$p_G \leq \frac{1}{|G(n, r)|} \sum_{k=r+1}^{n/2} \binom{n}{k} \frac{1}{k} \left(\frac{kr}{k-1} \right)^{k-1} |G(n-k, r)|.$$

Using the asymptotic approximation of Corollary 2.17 in (Bollobas 2001, page 55)

$$|G(n, r)| \sim \sqrt{2} e^{-(r^2-1)/4} \left(\frac{r^{r/2}}{e^{r/2} r!} \right)^n n^{rn/2} \text{ as } n \rightarrow +\infty,$$

we get

$$p_G \leq \sum_{k=r+1}^{n/2} \binom{n}{k} \frac{1}{k} \left(\frac{kr}{k-1} \right)^{k-1} \left(\frac{r^{r/2}}{e^{r/2} r!} \right)^{n-k} (n-k)^{r(n-k)/2} \left(\frac{r^{r/2}}{e^{r/2} r!} \right)^{-n} n^{-rn/2}.$$

Next, using the Stirling inequality, we obtain

$$\begin{aligned} p_G &\leq \sum_{k=r+1}^{n/2} \binom{n}{k} \frac{1}{k} \left(\frac{kr}{k-1} \right)^{k-1} \left(e^{r/2} \left(\frac{r}{e} \right)^r \frac{\sqrt{2\pi r} e^{\frac{1}{12r}}}{r^{r/2}} \right)^k (1-k/n)^{rn/2} \frac{1}{(n-k)^{rk/2}}. \\ p_G &\leq \sum_{k=r+1}^{n/2} \binom{n}{k} \frac{1}{k} \left(\frac{kr}{k-1} \right)^{k-1} r^{kr/2} e^{-\frac{kr}{2} + \frac{k}{12r}} \frac{\sqrt{2\pi r}^k}{(n-k)^{rk/2}}, \text{ as } 1-k/n < 1 \\ &\leq \sum_{k=r+1}^{n/2} \binom{n}{k} \frac{1}{k} \left(\frac{kr}{k-1} \right)^{k-1} r^{kr/2} \sqrt{2\pi r}^k e^{-\frac{kr}{2} + \frac{k}{12r}} \frac{e^{-\frac{kr}{2} + \frac{k}{12r}}}{(n/2)^{rk/2}}, \text{ as } k \leq n/2 \\ &\leq \sum_{k=r+1}^{n/2} \frac{n^{k-rk/2}}{k!} \frac{1}{k} \left(\frac{(r+1)r}{r} \right)^{k-1} r^{kr/2} \sqrt{2\pi r}^k e^{-\frac{kr}{2} + \frac{k}{12r}} 2^{rk/2}, \text{ as } \frac{k}{k-1} \text{ is a decreasing function of } k \\ &\leq \sum_{k=r+1}^{n/2} \frac{n^{k-rk/2}}{\left(\frac{k}{e} \right)^k \sqrt{2\pi k} e^{\frac{1}{12k+1}}} (r+1)^{k-2} r^{k\frac{r+1}{2}} \sqrt{2\pi}^k e^{-\frac{kr}{2} + \frac{k}{12r}} 2^{rk/2}, \text{ using the Stirling's inequality} \\ &\leq \frac{1}{\sqrt{2\pi}} \sum_{k=r+1}^{n/2} n^{k-rk/2} (r+1)^{-5/2} r^{k\frac{r+1}{2}} \sqrt{2\pi}^k e^{k(1-\frac{r}{2}) + \frac{k}{12r}} 2^{rk/2} \\ &\leq \frac{1}{\sqrt{2\pi} (r+1)^{5/2}} \sum_{k=r+1}^{n/2} s^k, \text{ with } s = n^{1-\frac{r}{2}} r^{\frac{r+1}{2}} \sqrt{2\pi} e^{1-\frac{r}{2} + \frac{1}{12r}} 2^{\frac{r}{2}} \\ &\leq s^{r+1} / \left(\sqrt{2\pi} (r+1)^{5/2} (1-s) \right), \text{ as } s < 1 \text{ for } n \text{ large enough.} \end{aligned}$$

3 Data sets description

The artificial data sets which are also depicted in Figure 1 were:

- A 5-clump-3-arc data set which consists of 340 points mixed in five Gaussian clumps and three arcs with uniformly scattered points. These data were first used in (Murua et al. 2008) to study the performance of Potts model clustering. Besides their variation in shape and distribution, the clusters are not very well separated.

- The Bull’s eye data set. This was obtained as in (Blatt et al. 1997). It consists of three ring clusters generated by Gaussians $N(R, 0.025)$. The radius of the rings, R , are set to 1, 2 and 3, for the first, second and third cluster, respectively. The cluster sizes are 800, 1600, and 2400, respectively. These data are a real challenge for most clustering methods.
- A 50-component Gaussian mixture data set. The Gaussian means were generated uniformly on the square $[0, 500] \times [0, 500]$. The variances were set to 1.0 for the first 25 clusters, and to 9.0 for the others. The difference in cluster volume may produce difficulties when choosing the appropriate temperature-bandwidth parameters.

The data are plotted in Figure 1. The three real data sets were:

- The well-known Iris data (Anderson 1935). The data consist of 50 four-dimensional instances associated with each of three Iris types (Iris Setosa, Versicolor and Virginica). As is common in the machine learning literature, Iris Versicolor and Iris Virginica were considered as forming only one class. These data are very simple to cluster. We have used them here because of their popularity with the statistical and machine learning community, and also to check that the procedures work.
- The Olive oil data set contains 572 samples of olive oil (Forina 1987) described by their content in eight fatty acids. Nine different types of oil origins are represented, namely South-Apulia, North-Apulia, Calabria, Sicily, Inland-Sardinia, Coast-Sardinia, Umbria, East-Liguria and West-Liguria. The best results to our knowledge report an adjusted Rand index of about 0.70 (Stuetzle and Nugent 2010).
- The Yeast Cell Cycle data (Cho et al. 1998) record the fluctuations of the expression levels of about 6000 genes over two cell cycles comprising 17 time points. We use the 5-phase subset of the data (Cho et al. 1998). It consists of 420 genes of which 386 have been assigned to one of five phases of the cells cycle. These data are very difficult to cluster correctly. The best results so far published to our knowledge report an adjusted Rand Index of about 0.45 (Yeung et al. 2001).

References

- [1] Anderson, Edgar, “The irises of the Gaspé Peninsula,” *Bulletin of the American Iris Society*, 59: 2–5 (1935).
- [2] Atchade, Y. F., Lartillot, N., Robert, C. P., “Bayesian computation for statistical models with intractable normalizing constants,” *Brazilian Journal of Probability and Statistics*. To appear (2010).
- [3] Biggs, N., “Algebraic Graph Theory,” Cambridge Press (1994)
- [4] Blatt, M., Wiseman, S. and Domany, E., “Data clustering using a model granular magnet,” *Neural Computation*, 9, 1805–1842 (1997).
- [5] Bollobas, B., “Random Graphs,” Academic Press (2001).
- [6] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W., “A genome-wide transcriptional analysis of the mitotic cell cycle,” *Molecular Cell*, 2, 65–73 (1998).
- [7] Forina M., Armanino, C., “Eigenvector projection and simplified non-linear mapping of fatty acid content of Italian olive oils,” *Annali di Chimica*, 72:127–141 (1982).

Figure 1: The three artificially generated data sets: (left) the 5-clump-3-arc data set, (center) the Bull's eye, and (right) the 50-component Gaussians.

- [8] Murua, A., Stanberry, L., and Stuetzle, W., “On Potts Model Clustering, Kernel K-Means and Density Estimation,” *Journal of Computational and Graphical Statistics*, 17(3), 629-658 (2008).
- [9] Roberts, G. O., and Rosenthal, J. S., “General state space Markov chains and MCMC algorithms,” *Probability Surveys*, 1, 20–71 (2004).
- [10] Stuetzle, W. and Nugent, R., “A generalized single linkage method for estimating the cluster tree of a density,” *Journal of Computational and Graphical Statistics*, 19, 397–418 (2010).
- [11] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L., “Model-based clustering and data transformations for gene expression data,” *Bioinformatics*, 17, 977-987 (2001).