Cluster processes

Peter McCullagh
University of Chicago

November 4, 2015

1 Cluster processes

A $\mathcal{R}^d$-valued cluster process is a pair (Y, B) in which $Y = (Y_1, \ldots)$ is an $\mathcal{R}^d$-valued random sequence and B is a random partition of the index set $\mathbb{N}$. The process is said to be exchangeable if, for each finite sample $[n] \subset \mathbb{N}$, the restricted process $(Y[n], B[n])$ is invariant under permutation $\sigma\colon [n] \to [n]$ of sample elements.

The Gauss-Ewens process is the simplest non-trivial example for which the distribution is as follows. First fix the parameter values $\lambda > 0$, and the matrices Σ^0, Σ^1, both symmetric positive definite of order d. In the first step, B has the Ewens distribution with parameter λ. Given B, Y is a zero-mean $\mathcal{R}^d$-valued Gaussian sequence with conditional covariances

$$\operatorname{cov}(Y_{ir}, Y_{js} \,|\, B) = \delta_{ij}\Sigma^0_{rs} + B_{ij}\Sigma^1_{rs},$$

where δ_{ij} is the Kronecker symbol. For $d = 2$, a scatterplot colour-coded by blocks of the Y values in $\mathcal{R}^2$ shows that the points tend to be clustered, the degree of clustering being governed by the ratio of between to within-cluster variances.

For an equivalent construction we may proceed using a version of the Chinese restaurant process in which tables are numbered in order of occupancy, and $t(i)$ is the number of the table at which customer i is seated. In addition, $\epsilon_1, \ldots$ and $\eta_1, \ldots$ are independent Gaussian sequences with independent components $\epsilon_i \sim N_d(0, \Sigma^0)$, and $\eta_i \sim N_d(0, \Sigma^1)$. The sequence t determines B, and the value for individual i is a vector $Y_i = \eta_{t(i)} + \epsilon_i$ in $\mathcal{R}^d$, or $Y_i = \mu + \eta_{t(i)} + \epsilon_i$ if a constant non-zero mean vector is included.

Each row of Fig. 1 is an independent realization of the Gauss-Ewens process with $\lambda = 2$, $\Sigma^0 = I_2$ and $\Sigma^1 = 9I_2$, with values colour-coded by block. In each row, the pattern of colours in the right panel matches closely that on the left, but the two rows are independent runs, so there is no relation between the blocks in the first row and those in the second.

Figure 2 is a similar illustration of the two-level tree-structured Gauss-Ewens process with two partitions $B' \le B$ representing classes and sub-classes. The

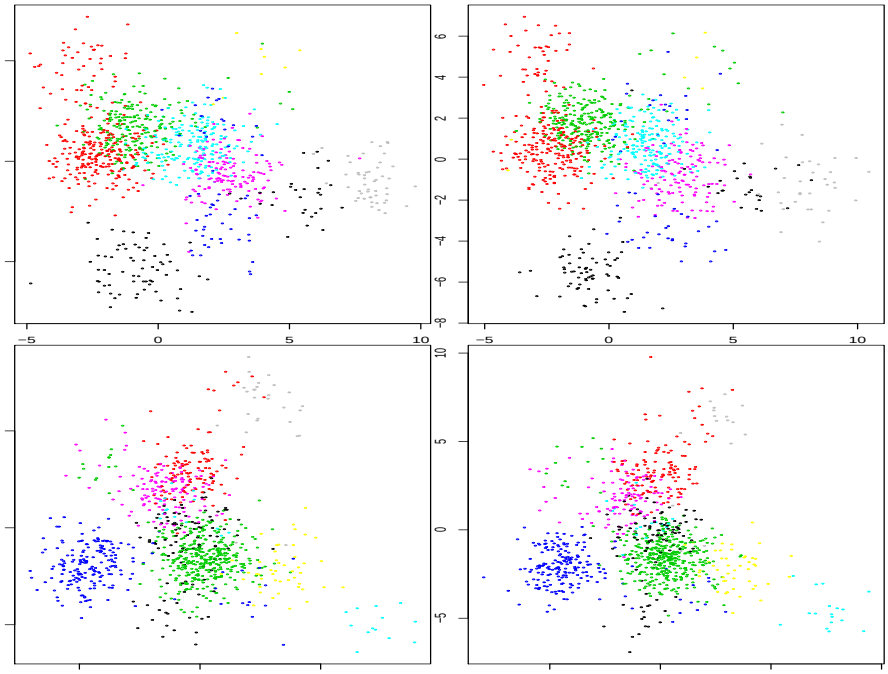

Figure 1. The zero-mean Gauss-Ewens process in $\mathcal{R}^2$. The left box shows the first 1000 values colour-coded by block; the right box shows the next 1000 values using the same colour code. Each row is an independent realization with same parameter, so the blocks in row 1 are unrelated those in row 2.

parameters in this case are $\lambda = 1$, $\Sigma^0 = I_2$, $\Sigma^1 = 100I_2$ and $\Sigma^2 = 5I_2$. The conditional distribution given B, B' is zero-mean Gaussian with covariance matrix

$$\mathrm{cov}(Y_{ir}, Y_{js} \mid B, B') = \delta_{ij}\Sigma^0_{rs} + B_{ij}\Sigma^1_{rs} + B'_{ij}\Sigma^2_{rs}.$$

The sub-partition has the effect of splitting each main cluster randomly into sub-clusters.

In effect, each major cluster is an independent copy of the ordinary Gauss-Ewens process, so the configurations in each major cluster are random and independent with the same distribution.

2 Classification using cluster processes

Despite the absence of class labels, cluster processes lend themselves naturally to prediction and classification, also called supervised learning. The description that follows is taken from McCullagh and Yang (2006) but, with minor modifications, the same description applies equally to more complicated non-linear versions associated using generalized linear mixed models (Blei, Ng and Jordan

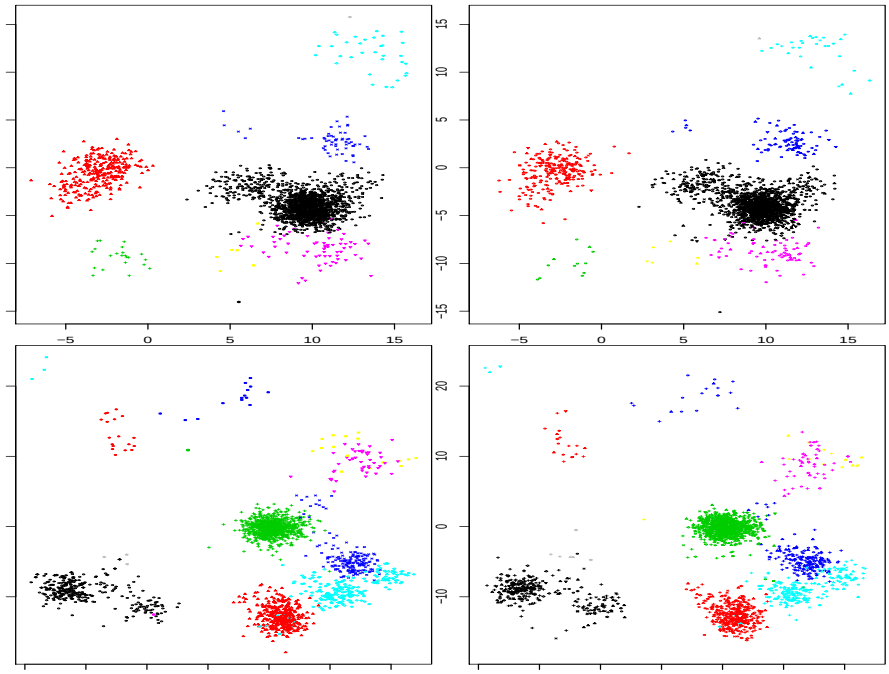

Figure 2. The two-level zero-mean Gauss-Ewens process in $\mathcal{R}^2$ colour-coded by major class. Each of the main blocks is partitioned at random into sub-blocks. The sub-partition is not illustrated, but can be inferred to some extent from the configuration.

2003). Given the observation $(Y[n], B[n])$ for the 'training sample' $[n]$, together with the feature vector Y_{n+1} for specimen u_{n+1}, the conditional distribution of $B[n+1]$ is determined by those events $u_{n+1} \mapsto b$ for $b \in B[n]$ and $b = \emptyset$ that are compatible with the observed partition. The assignment of a positive probability to the event that the new specimen belongs to a previously unobserved class seems highly desirable, even logically necessary, in many applications.

In the simplest version of the Gauss-Ewens model the between-blocks covariance matrix is proportional to the within-blocks matrix, i.e. $\Sigma^1 = \theta\Sigma^0$ for some scalar $\theta \geq 0$. The parameters in this reduced model are μ, θ, Σ_0, and an explicit analytic expression is available for the classification probability of a new unit such that $Y_{n+1} = y'$ as follows:

$$\mathrm{pr}(u_{n+1} \mapsto b \,|\, Y[n+1], B[n]) \propto \begin{cases} \#b \; q_b(y' - \mu - w_b(\bar{y}_b - \mu)) & b \in B[n]; \\ \lambda \; q_\emptyset(y' - \mu) & b = \emptyset. \end{cases}$$

Here, $\bar{y}_b$ is the mean for block b, the weight $w_b = \theta\,\#b/(1 + \theta\,\#b)$ is monotone increasing in block size, and $q_b(\cdot)$ is the density of the d-dimensional normal distribution with zero mean and covariance matrix $(1 + w_b)\Sigma_0$. In practice, the parameters must first be estimated from the block means and within-blocks sample covariance matrix. If $\theta = 0$, the y-values are irrelevant and the classification probabilities reduce to the Chinese restaurant process; otherwise, if $\theta > 0$, the conditional log odds

$$\log\left(\frac{\mathrm{pr}(u_{n+1} \mapsto b \,|\, \cdots)}{\mathrm{pr}(u_{n+1} \mapsto b' \,|\, \cdots)}\right)$$

is the difference of two quadratic forms, which is linear in y' if $\#b = \#b'$, and approximately linear otherwise. For classification purposes, this version of the Gauss-Ewens process may be viewed as a refinement of Fisher's linear discriminant model; it is more flexible in that the set of classes is not pre-specified, so that a new unit may be assigned with high probability to a previously unobserved class.

If the classes are tree-structured with two levels, we may generate a sub-partition $B' \leq B$ whose conditional distribution given B is Ewens restricted to the interval $[\mathbf{0}_n, B]$, with parameter λ'. This sub-partition has the effect of splitting each main cluster or genus, randomly into sub-clusters representing species. For the sample $[n]$, let $t'(i)$ be the number of the sub-cluster or the name of the species to which individual i belongs. Given B, B', the Gauss-Ewens two-level tree process illustrated in Fig. 2 is a sum of three independent Gaussian processes $Y_i = \eta_{t(i)} + \eta'_{t'(i)} + \epsilon_i$ for which the conditional distributions may be computed as before. In this situation, however, events that are compatible with the observation $B[n], B'[n]$ are of three types as follows:

$$u_{n+1} \mapsto b' \in B'[n], \quad u_{n+1} \mapsto \emptyset \subset b \in B[n], \quad u_{n+1} \mapsto \emptyset.$$

In all, there are $\#B' + \#B + 1$ disjoint events for which the conditional distribution given the observation $B[n], B'[n], Y[n+1]$ must be computed. An event

of the second type is one in which the new specimen belongs to the major class or genus $b \in B$, but not to any of the species previously observed for this class. An event of the third type is a new genus.

3 Acknowledgements

This is a revised version of parts of the article *Random permutations and partition models* from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science, Heidelberg: Springer Science +Business Media, LLC. Support for this research was provided in part by NSF Grant DMS-0906592.

References

[1] Blei, D., Ng, A. and Jordan, M. (2003) Latent Dirichlet allocation. *J. Machine Learning Research* **3**, 993–1022.

[2] McCullagh, P. and Yang, J. (2006) Stochastic classification models. Proc. International Congress of Mathematicians, 2006, vol. III, 669–686.

[3] McCullagh, P. and Yang, J. (2008). How many clusters? *Bayesian Analysis* **3**, 1–19.