

Random  
partitions

Lars Holst  
Symposium

Random Partitions with an Application to  
Ecological Genetics  
by  
Timo Koski MAI/LiU

Lars Holst Symposium

June 1, 2007

# Unsupervised classification

Random  
partitions

Lars Holst  
Symposium

Unsupervised classification of a set of observed data is understood as learning a postulated hidden structure underlying a set of observations.



*Adansonia digitata*



*Drosophila melanogaster*

fruit fly

# Example from population genetics

Random  
partitions

Lars Holst  
Symposium

- The existence of several panmictic units, *i.e.* parts of a population<sup>1</sup> of interest in which random mating pattern can be considered as a feasible approximation. This can be properly investigated only on basis of some empirical observations<sup>2</sup> from the target population. The observations take often the form of alleles of molecular marker genes.

---

<sup>1</sup>later *Drosophila melanogaster* in Africa

<sup>2</sup>to be called *feature vectors*

# Inference of population structure

Random  
partitions

Lars Holst  
Symposium

The fundamental paper is

J.K. Pritchard, M. Stephens & P. Donnelly (2000): Inference of Population Structure Using Multilocus Genotype Data.

*Genetics* 155, 945–959.

Here it is postulated that there are  $k$  populations ( $k$  may be unknown), each of which is characterized by a set of allele frequencies at each locus. The method of Pritchard et.al. attempts to assign individuals to populations on the basis of their genotypes, while simultaneously estimating the population allele frequencies. This assumes that the loci are unlinked and that there is HW-equilibrium within populations.

# Inference of population structure

Random  
partitions

Lars Holst  
Symposium

The method of Pritchard et.al. was developed in K.J. Dawson & Kh. Belkhir (2001): A Bayesian Approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, 78, pp. 59–77.  
and here we hope to improve on Dawson and Belkhir in terms of interpretation and computational efficiency.

Classification is here seen as merely a partition of items (represented by an observed data set). A class is then a cell in the partition. We are not discussing hierarchic classification.

# On the shoulders of Giants

Random  
partitions

Lars Holst  
Symposium



Carl von Linné



Michel Adanson (1727–1806)

# Michel Adanson (from Enc. Britannica)

Random  
partitions

Lars Holst  
Symposium

**Adanson was the first to use statistics in plant and animal classification.** He proposed his universal method, a system of classification distinct from those of Buffon and Linnaeus. Adanson founded his classification of all organized beings on the consideration of each individual organ. As each organ gave birth to new relations, so he established a corresponding number of arbitrary arrangements. Those beings possessing the greatest number of similar organs were referred to one great division, and the relationship was considered more remote in proportion to the dissimilarity of organs. **Adanson established natural classification as a fundamental aim of biology.**

# Natural classification

Random  
partitions

Lars Holst  
Symposium

With description of a class, we should be able to predict (well) the properties of an item in this class without having investigated the item → a 'natural classification' → predictive probability distribution (e.g. in Dawson & Belkhir)

# Notation

Random  
partitions

Lars Holst  
Symposium

Let us consider a set  $N$  of  $n$  items, these can be individuals sampled from a population (e.g. for which we wish to study the existence of genetic structure).

The items in  $N$  are represented by discrete valued *feature vectors*  $\mathbf{x}^{(i)}$ ,  $i = 1, \dots, n$ .

We use

$$\mathbf{x}^{(N)} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$$

to denote the set of the  $n$  feature vectors.

Also,  $\mathbf{x}^{(s)}$  is used to denote the feature vectors for a subset  $s \subseteq N$  of the items.

# Posterior on Partitions

Random  
partitions

Lars Holst  
Symposium

Let  $S = (s_1, \dots, s_k)$  ( $1 \leq k \leq n$ ) be a partition of  $N$ .  $\mathcal{S}$  is the family of all partitions  $S$  of  $N$ .

The posterior probability of  $S$  given the feature vectors  $\mathbf{x}^{(N)}$  equals then

$$p(S|\mathbf{x}^{(N)}) = \frac{p(\mathbf{x}^{(N)}|S)p(S)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|S)p(S)}.$$

Maximum A Posterior Partition (MAPP)

$$\hat{S} = \arg \max_{S \in \mathcal{S}} p(S|\mathbf{x}^{(N)}).$$

# Probability on Partitions

Random  
partitions

Lars Holst  
Symposium

We need to determine three things

- $p(S)$
- $p(\mathbf{x}^{(N)}|S) \leftrightarrow$  a predictive distribution of  $\mathbf{x}^{(N)}$  w.r.t  $S$
- an effective algorithm to find a MAPP  $\hat{S}$ .

# Probability on Partitions

Random  
partitions

Lars Holst  
Symposium

The statement of the population structure problem with probability mass on partitions has been developed from the work by K.J. Dawson & Kh. Belkhir (2001) loc.cit.

A difficulty of K.J. Dawson & Kh. Belkhir is in their treatment of the priors. We use the urn model due to A.J. Stam (1983)<sup>3</sup>

The theory of random partitions (of integers) due to J.F.C. Kingman applied in analysis of neutral genetic evolution with an infinite alleles model is also applicable here.

---

<sup>3</sup>Generation of a random partition of a finite set by an urn model. *J. Combin. Theory. Ser. A*, **35**, 231-240.

# Label Switching

Random  
partitions

Lars Holst  
Symposium

By considering random partitions we avoid the difficulties encountered in unsupervised learning (of mixtures) connected to the phenomenon of *label switching* (an identifiability problem):

M. Stephens (2000): Dealing with label switching in mixture models. *Journal of the Royal Statistical Society*, **B**, 62, pt 4, 795–809.

# Prior Distributions on Partitions

Random  
partitions

Lars Holst  
Symposium

Let  $u$  be a stochastic number of urns into which the items of  $N$  are allocated randomly. Consequently, the contents of the resulting  $k$  nonempty urns determine a *stochastic* partition of  $N$ . The implied probability measure equals

$$p(S) = \sum_{u=k}^{\infty} \pi(S|u)p(u), S \in \mathcal{S}, \quad (1)$$

for  $S$  with  $k$  cells,  $1 \leq k \leq n$ , where

$$\pi(S|u) = \frac{u(u-1)(u-2)\cdots(u-k+1)}{u^n}$$

# Uniform Prior on Partitions

Random  
partitions

Lars Holst  
Symposium

Let  $B_n(= |\mathcal{S}|)$  be the  $n$ :th Bell number, i.e., the number of partitions of  $n$  items. This satisfies Dobinski's formula:  $B_n = e^{-1} \sum_{m=1}^{\infty} \frac{m^n}{m!}$ . Stam showed that if we pick  $u$  from the probability mass function

$$p(u = m) = \frac{1}{B_n} e^{-1} \frac{m^n}{m!}, \quad m = 1, 2, \dots,$$

and thereafter place the  $n$  items at random in the  $u = m$  urns ( $m - k$  is the number of empty urns), then

$$p(S) = |\mathcal{S}|^{-1}, S \in \mathcal{S}.$$

# Gibbs Prior on Partitions

Random  
partitions

Lars Holst  
Symposium

J.W. Pitman<sup>4</sup> characterized all  $p(u)$  yielding the uniform distribution. Using the urn procedure of Stam he noted that if

$$p(u = m) = \frac{m^n}{c_n} e^{-\lambda} \frac{\lambda^m}{m!}, \quad m = 1, 2, \dots,$$

where  $c_n = e^{-\lambda} \sum_{r=0}^{\infty} r^n \frac{\lambda^r}{r!}$ , then there is the Gibbs distribution

$$p(S) = \frac{\lambda^k}{c_n}, \quad S \in \mathcal{S}.$$

if the number of cells in  $S$  is  $k$ .

---

<sup>4</sup>J.W. Pitman (1997): Some probabilistic aspects of set partitions. *Amer. Math. Monthly*, **104**, 201-209

# The Bell numbers

Random  
partitions

Lars Holst  
Symposium

There is an excellent approximation to the  $n$ :th Bell number

$$B_n \sim (\lambda_n + 1)^{-1/2} e^{n(\lambda_n - 1 - \frac{1}{\lambda_n}) - 1}$$

where  $\lambda_n$  solves

$$e^\lambda \lambda = n$$

This is shown (using a probabilistic method) in  
L. Holst (1981): On Numbers Related to Partitions of Unlike  
Objects and Occupancy Problems. *European Journal of  
Combinatorics*, **2**, 231–237.

# Exchangeability

Random  
partitions

Lars Holst  
Symposium

We assume that  $\mathbf{x}^{(s)}$  are *unrestrictedly infinitely exchangeable*  
<sup>5</sup> This implies that, if we combine any permutation of the item values  $x_j^{(1)}, \dots, x_j^{(|s|)}$ <sup>6</sup>, for a fixed  $j$ , with arbitrary corresponding permutations over the remaining features, the same predictive probability mass function for  $\mathbf{x}^{(s)}$  is obtained.

---

<sup>5</sup>c.f., Definitions 4.2, 4.3, and 4.13, and Propositions 4.2 and 4.18 in Bernardo and Smith (1994): *Bayes Theory*,

<sup>6</sup> $|s|$  is the cardinality of the cell  $s$

# Exchangeability

Random  
partitions

Lars Holst  
Symposium

Furthermore, the sequences  $x_j^{(1)}, \dots, x_j^{(|s|)}$ <sup>7</sup> are summarized in terms of the sufficient statistics  $n_{jl}, l = 1, \dots, r_j$ , where  $n_{jl}$  represents the number of copies of value  $l$  for feature  $j$  observed among the items in  $s$ .

---

<sup>7</sup> $|s|$  is the cardinality of the cell  $s$

# Exchangeability

Random  
partitions

Lars Holst  
Symposium

By extending the unrestricted exchangeability assumption to hold over the cells  $s_1, \dots, s_k$ , we obtain the joint probabilistic characterization for  $\mathbf{x}^{(N)} = \{x^{(s_1)}, \dots, x^{(s_k)}\}$  as

$$p(\mathbf{x}^{(N)}|S) = \int_{\Theta} \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} \theta_{cjl}^{n_{cjl}} dQ(\theta|S),$$

where  $n_{cjl}$  represents now the number of copies of value  $l$  for feature  $j$  observed among the items in  $s_c$ , and  $\theta_{cjl}, \theta$ , and  $Q(\theta|S)$ , are defined suitably.

# Dirichlet Density

Random  
partitions

Lars Holst  
Symposium

The assumption of sufficientness<sup>8</sup> leads to an explicit form of the prior beliefs  $Q(\theta|S)$ , the product Dirichlet distribution

$$Q(\theta|S) \propto \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} \theta_{cjl}^{\lambda_{cjl}-1},$$

where the hyperparameter  $\lambda_{cjl} > 0$ , for all index values.

---

<sup>8</sup>S.L. Zabell (1982): W.E. Johnson's 'Sufficientness' Principle, *Annals of Statistics*, 10, pp. 1091–1099.

# Exchangeability

Random  
partitions

Lars Holst  
Symposium

The sufficientness assumption implies that the observed values of features within an arbitrary class provide no information about the values observed in any other class . The probability mass function  $p(\mathbf{x}^{(N)}|S)$  defines a predictive probability model for the observed items in  $N$ .

# The predictive distribution

Random  
partitions

Lars Holst  
Symposium

The explicit predictive probability mass function for the items equals now

$$p(\mathbf{x}^{(N)}|S) = \prod_{c=1}^k \prod_{j=1}^d \frac{\Gamma(\sum_{l=1}^{r_j} \lambda_{cjl})}{\Gamma(\sum_{l=1}^{r_j} \lambda_{cjl} + n_{cjl})} \prod_{l=1}^{r_j} \frac{\Gamma(\lambda_{cjl} + n_{cjl})}{\Gamma(\lambda_{cjl})},$$

where  $\Gamma(\cdot)$  is the gamma function.

# Hyperparameters

Random  
partitions

Lars Holst  
Symposium

It has been argued<sup>9</sup> that the appropriate choice of hyperparameters is

$$\lambda_{cjl} = r_j^{-1}, l = 1, \dots, r_j; j = 1, \dots, d; c = 1, \dots, k,$$

as other choices can lead to substantial amount of information in the genetic data being contained in the prior.

A theoretical argument for this is due to W. Perks (1947)<sup>10</sup>.

---

<sup>9</sup>E.C. Anderson & E.A. Thompson: A Model-Based Method for Identifying Species Hybrids Using Multilocus Genetic Data. *Genetics*, 160, 1217–1229.

<sup>10</sup>Some observations on inverse probability including a new indifference rule. *J. Institute Actuaries*, **73**, pp. 285–334.

# Parallel MCMC on Partitions for MAPP

Random  
partitions

Lars Holst  
Symposium

MCMC algorithms for simulating several independent parallel Markov chains are well known. Although the method seems to perform reliably even for fairly large data sets, the computational effort required may become prohibitive for complex data sets.

# Parallel MCMC on Partitions

Random  
partitions

Lars Holst  
Symposium

We are able to improve the parallel MCMC strategy significantly, by introducing dependence between the chains, and by modifying the transition mechanisms of the individual chains.

# Proposal Mechanism

Random  
partitions

Lars Holst  
Symposium

The proposal mechanism to derive  $S^*$  from  $S$  was constructed from the following four different possibilities:

- With probability  $1/2$ , merge two randomly chosen classes  $s_c, s_{c^*}$ .
- With probability  $1/2$  split a randomly chosen class  $s_c$  into two new classes, whose cardinalities are uniformly distributed between 1 and  $|s_c| - 1$ , and whose elements are randomly chosen from  $s_c$ .
- Move an arbitrary item from a randomly chosen class  $s_c, |s_c| > 1$ , into another randomly chosen class  $s_{c^*}$ .
- Choose one item randomly from each of two randomly chosen classes  $s_c$  and  $s_{c^*}$ , and exchange them between the classes.

# Non-reversible MCMC on Partitions

Random  
partitions

Lars Holst  
Symposium

An algorithm, based on parallel Markov chains  $S_j = \{S_{tj}\}$  with the transition kernels, with the probability of a transition from a current state  $S$  to a proposed (by the mechanism above) new state  $S^*$ , as

$$\min \left( 1, \frac{p(\mathbf{x}^{(N)}|S^*)}{p(\mathbf{x}^{(N)}|S)} \right).$$

We use

$$p(S) = |\mathcal{S}|^{-1}, S \in \mathcal{S}.$$

The chains draw at independent random times a value from  $P(S_j | \mathbf{x}^{(N)})$ . This forces the chains to move to regions of maximal posterior probability.

# The Parallel Search

Random  
partitions

Lars Holst  
Symposium

Define the sequence of probabilities  $\{\alpha_t, t = 2, 3, \dots\}$  according to

$$\alpha_t = \frac{1}{q \log t},$$

where  $q \geq 1$  can be chosen suitably, for instance  $q \in [5, 10]$ .  
 $Z_0 = 0$ , and  $P(Z_t = 1) = \alpha_t, P(Z_t = 0) = 1 - \alpha_t$ ,  
independently for  $t = 1, 2, \dots$ .

# The Parallel Search

Random  
partitions

Lars Holst  
Symposium

For each  $t = 0, 1, \dots$ , define the distribution

$$P_t(S_{tj}) = \frac{p(\mathbf{x}^{(N)}|S_{tj})}{\sum_{j=1}^m p(\mathbf{x}^{(N)}|S_t)},$$

over the space of the current states  $\{S_{t1}, S_{t2}, \dots, S_{tm}\}$ . For each  $t = 0, 1, \dots$  such that  $Z_t = 1$ , the transition to the next state is determined according to this distribution, such that the next state for each chain is sampled to the non-reversible proposal-acceptance formulae, independently for  $j = 1, \dots, m$ .

# The Parallel Search

Random  
partitions

Lars Holst  
Symposium

For each  $t$ , such that  $Z_t = 0$ , transition to the next state  $S_{(t+1)j}$  is determined according to the non-reversible proposal-acceptance formulae above independently for  $j = 1, \dots, m$ .

# Consistency

Random  
partitions

Lars Holst  
Symposium

$$P_t(S_{tj}) = \frac{p(\mathbf{x}^{(N)}|S_{tj})}{\sum_{j=1}^m p(\mathbf{x}^{(N)}|S_t)},$$

converges as  $t \rightarrow \infty$ , to the posterior on  $\mathcal{S}$ . This follows since the parallel chains will eventually visit every partition between two interaction times  $\tau$  and  $\tau + 1$  with  $Z_\tau = 1$  and  $Z_{\tau+1} = 1$ .

# The Parallel Search

Random  
partitions

Lars Holst  
Symposium

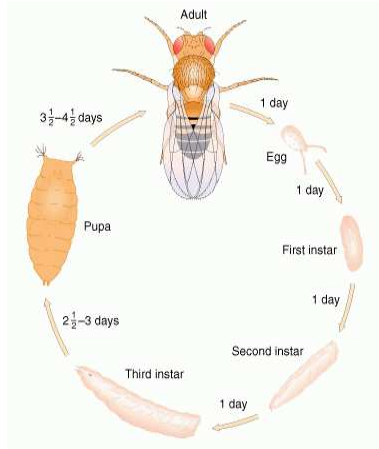
The simplified transition kernel has two advantages: the cardinalities of the classes  $|s_c|$  only affect the probability of proposing a particular solution, rather than the acceptance of it. Since the proposal probabilities need not to be explicitly calculated at every  $t$ , we are able to reduce the number of computational operations needed in the algorithm.

# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

As illustration we search for the population structure of *Drosophila melanogaster* (a fruit fly) in Africa.



# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

Tropical sub-Saharan regions are considered to be the geographical origin of *Drosophila melanogaster*. Starting from there, the species colonized the rest of the world after the last glaciation about 10000 years ago. Consistent with this demographic scenario, African populations have been shown to have higher levels of microsatellite and sequence variation than cosmopolitan populations. Nevertheless, limited information is available on the genetic structure of African populations.

# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

Dieringer *et al.* (2005) (exact reference later) investigated the population structure of *D. melanogaster* using 17 highly polymorphic microsatellite<sup>11</sup> loci. The 178 African individuals in the sample were originally collected from 13 distinct geographical regions.

---

<sup>11</sup>A microsatellite, also called simple sequence repeat, short tandem repeat or variable number tandem repeat, is a short (2-5 base) motif that is repeated multiple times and is flanked by unique DNA.

# *Drosophila melanogaster*

Random  
partitions

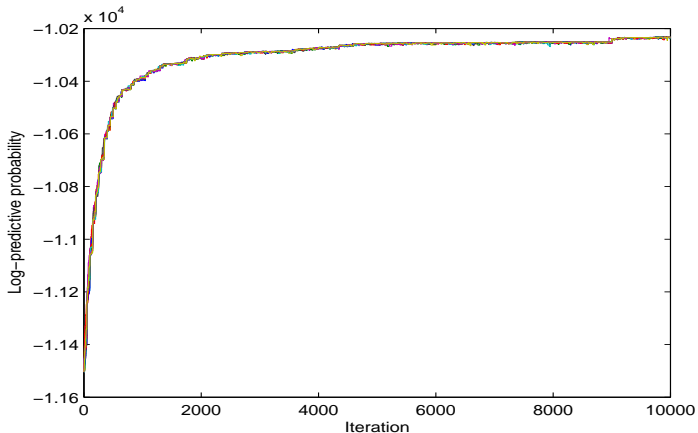
Lars Holst  
Symposium

For inference about the population structure we use the uniform prior over  $\mathcal{S}$ . In the simulation of the posterior distribution, 50 parallel interacting search processes were run for 10000 iterations. In the next figure the behavior of the logarithm of the predictive probability  $\log p(\mathbf{x}^{(N)}|S)$  over the iterations is shown.

# The Parallel Search

Random  
partitions

Lars Holst  
Symposium



# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

Two major groups emerged in the analysis: one consisting of the three sample populations from North Africa (Agadir, Djerba, Marrakech) (= a random mating unit), and the other including majority of individuals from Moribabougou and Kisoro, and additionally approximately one third of the individuals from both Malinga and Victoria Falls. Half of the Kampala sample and one individual from Harare formed another group, whereas the remaining individuals were included in groups of size 1-5.

# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

Investigation of the posterior  $p(S|\mathbf{x}^{(N)})$  probabilities for different configurations, revealed that for many individuals in the small groups, an alternative allocation to another non-major group was also plausible.

# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

Sample population	Major groups			Minor groups	Sum
	1	2	3		
Djerba, Tunisia	24	-	-	-	24
Marrakech, Morocco	8	-	-	-	8
Agadir, Morocco	4	-	-	-	4
Moribabogou, Mali	-	8	-	2	10
Abidjan, Ivory Coast	-	-	-	4	4
Kenya, multiple locations	-	-	-	18	18

# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

Sample population	Major groups			Minor groups	Sum
	1	2	3		
Malindi, Kenya	-	7	-	14	21
Sengwa, Zimbabwe	-	-	-	13	13
Harare, Zimbabwe	-	-	1	12	13
Lake Kariba, Zimbabwe	-	-	-	11	11
Victoria Falls, Zimbabwe	-	9	-	19	28
Kisoro, Uganda	-	8	-	7	15
Kampala, Uganda	-	1	4	5	10
Sum	36	33	5	104	178

# *Drosophila melanogaster*

Random  
partitions

Lars Holst  
Symposium

The results agree with the classification in Dieringer, D., Nolte, V. and Schlötterer, C. (2005): Population structure in African *Drosophila melanogaster* revealed by microsatellite analysis. *Mol. Ecol.* **14**, 563-573.

The first-named author gave kindly access to the data.

# BAPS

Random  
partitions

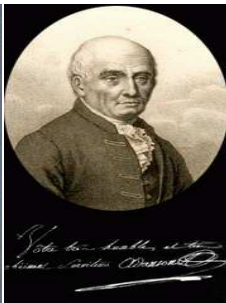
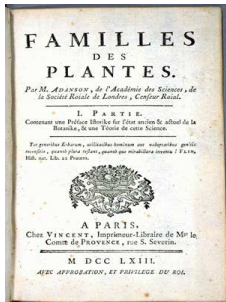
Lars Holst  
Symposium

The computations use the software BAPS:  
J. Corander, P. Waldmann & M. Sillanpää (2003): **B**ayesian  
**A**nalysis of Genetic Differentiation Between **P**opulations.  
*Genetics*, 163, 367–374.  
with a public freeware code written in MATLAB.

# Adansoniana : Familles des plantes & Acknowledgements

Random  
partitions

Lars Holst  
Symposium



This is joint work with Jukka Corander and Mats Gyllenberg, both from the Rolf Nevanlinna Institute/University of Helsinki.