Stochastic classification models

Peter McCullagh and Jie Yang *

Abstract. Two families of stochastic processes are constructed that are intended for use in classification problems where the aim is to classify units or specimens or species on the basis of measured features. The first model is an exchangeable cluster process generated by a standard Dirichlet allocation scheme. The set of classes is not pre-specified, so a new unit may be assigned to a previously unobserved class. The second model, which is more flexible, uses a marked point process as the mechanism generating the units or events, each with its associated class and feature. The conditional distribution given the superposition process is obtained in closed form for one particular marked point process. This distribution determines the conditional class probabilities, and thus the prediction rule for subsequent units.

Mathematics Subject Classification (2000). Primary 62H30; Secondary 68T10.

Keywords. Cluster process; Cox process; Dirichlet process; Gauss-Ewens process; Lack of interference; Marked point process; Permanent polynomial; Random subset; Supervised learning

1. Introduction

1.1. Classification. The problem of numerical taxonomy is to classify individual specimens or units u on the basis of measured variables or features $x(u) \in \mathcal{X}$. The units may be anything from tropical insects to bitmap images of handwritten digits or vocalizations of English words. The feature variables may be length or width or weight measurements in the case of insects, or the Fourier transformation at certain frequencies in the case of spoken words. The choice of feature variables is an important problem in its own right, but this matter is of little concern in the present paper.

A deterministic classification model is a rule or algorithm that associates with each feature value $x \in \mathcal{X}$ a class $y(x) \in \mathcal{C}$. Ordinarily the model must be primed or trained on a sample of units with measured features and known classes. In the dialect of artificial intelligence and computer science, the classifier learns the characteristics peculiar to each class and classifies subsequent units accordingly. When the training is over, each subsequent input is a feature value $x(u')$ for a new unit, and the output is the assigned class. The error rate is the fraction of wrong

*We are grateful to Jim Pitman for helpful comments. Support for this research was provided by NSF Grant DMS-0305009.

calls.

A stochastic classification model is a process determining a rule that associates with each feature value x a probability distribution $p(\cdot; x)$ on the set of classes. Once again, the classification model must be primed or trained on a sample of units with measured features and known classes. In statistical language, the classifier is a statistical model with unknown parameters to be estimated from the training data. Subsequent units are classified in the usual stochastic sense by computing the conditional distribution given the training data and the feature value for the new unit.

Three stochastic models are described in the sections that follow. The first of these is a regression model with independent components in which the feature values are treated as covariates. The second is an exchangeable cluster process closely related to Fisher's discriminant model, but different in several fundamental ways. The third model is also an exchangeable cluster process, called a permanent cluster process because the conditional distributions are expressed in terms of permanent polynomials.

The distinction between a closed classification model with a pre-determined set of labelled classes, and an open model with unlabelled classes is emphasized. A model of the latter type has a mathematical framework that permits a new unit to be assigned to a class that has not previously been observed and therefore does not have a name. The goal is to construct a classification model with no more than 4–5 parameters to be estimated regardless of the number of classes or the dimension of the feature space. In this way, the technically difficult problems associated with consistency and parameter estimation in high dimensional models are evaded. Ideally, the model should be capable of adapting to classification problems in which one or more classes occupies a non-convex region, or even several disconnected regions, in the feature space.

1.2. Remarks on the literature. The literature on stochastic classification is very extensive, the modern theory beginning with Fisher's discriminant model ([12]). Logistic regression models emerged in the 1960s, and with the advent of faster computing, smoothed versions using penalized likelihood became more popular. Stochastic models used in the statistical literature are sometimes complicated, but they are frequently of the most elementary form with independent components such that

$$\log\bigl(\mathrm{pr}(Y(u) = r \mid X)\bigr) = f_r(X(u)).$$

The goal is to estimate the functions f_r under certain smoothness conditions, which are enforced through penalty functions added to the log likelihood. For a good overview see [29], [15], [27] or [16].

At the more mathematical end of the statistical spectrum, the same model with independent components is frequently used, with f belonging to a suitable space of functions, usually a Besov space. The stated mathematical goal is to obtain the best estimate of f under the most adverse conditions in very large samples ([9]). Smoothing is usually achieved by shrinkage or thresholding of coefficients in

a wavelet expansion.

The past decade has seen an upsurge of work in the computer science community under the headings of artificial intelligence, data mining and supervised learning. Methods used include neural nets, support vector machines and tree classifiers. The emphasis is primarily on algorithms, regularization, efficiency of computation, how best to combine weak classifiers ([13]), and so on. Few algorithms and methods of this type have an overt connection with a generative stochastic process beyond the simple additive form with independent components.

In the Bayesian literature, more complicated processes are constructed using mixture models with Dirichlet priors for the class frequencies ([11], [2], [24], [14]). The cluster process in section 3 is in fact a simple special case of a more general classification model ([4], [8]). It is used here mainly for illustrative purposes because the distributions can be studied analytically, which is rare for processes generated by Dirichlet allocation schemes.

The semi-parametric models described in section 4 are of a different type. They are based on Cox processes ([5]) with a baseline intensity measure μ treated as an unknown parameter. One major attraction for practical work is that the conditional distribution of the class labels given the observed features does not depend on the baseline measure. The unknown nuisance parameter is eliminated by conditioning rather than by integration, and this conditional distribution is the basis for inference and classification.

2. Logistic discrimination

2.1. Non-interference and regression models. Let $\mathcal{U}$ be the set of units, the infinite set of objects such as plots or subjects or specimens, on which the process Y is defined. A covariate $x\colon \mathcal{U} \to \mathcal{X}$ is a function on the units, the values of which are thought to have an effect on the distribution. In a logistic regression model it is the class $Y(u) \in \mathcal{C}$ that is regarded as the response, and the measured feature $x(u)$ is the covariate.

In practical work, it is often helpful to distinguish between covariates such as sex, age and geographical position that are intrinsic to the unit, and treatment variables such as medication or variety that can in principle be controlled by the experimenter. For mathematical purposes it is more useful to distinguish between a covariate as a function on the units, and a relationship as a function on pairs of units. Examples of the latter include distance if the units are arrayed in space, temporal ordering for time points, genetic or familial relationships if the units are individual organisms, or a block factor as an equivalence relation on units. The statistical distinction, roughly speaking, is that a covariate affects one-dimensional marginal distributions, while a relationship affects bivariate distributions. For present purposes, however, distinctions of this sort are unnecessary.

A regression model is a process in which the joint distribution of the response $(Y(u_1), \ldots, Y(u_n))$ on n units is determined by the covariate values $x = (x(u_1), \ldots, x(u_n))$ on those units. We write $P_n(\cdot; x)$ for the joint distribution

on an ordered set of n distinct units, implying that two sets of units having the same ordered list of covariate values, also have the same distribution. In other words, if $(x(u_1), \ldots, x(u_n)) = (x(u'_1), \ldots, x(u'_n))$ then $(Y(u_1), \ldots, Y(u_n))$ and $(Y(u'_1), \ldots, Y(u'_n))$ are both distributed as $P_n(\cdot; x)$.

In general, the probability assigned to an event $A \subset \mathcal{C}^n$ depends on the covariate vector $(x_1, \ldots, x_n)$. However, the lack of interference condition

$$P_n(A; (x_1, \ldots, x_n)) = P_{n+1}(A \times \mathcal{C}; (x_1, \ldots, x_n, x_{n+1})) \tag{2.1}$$

implies that the probability assigned by P_{n+1} to the event $A \times \mathcal{C}$ does not depend on the final component x_{n+1} of x. The failure of this condition means that the probability assigned by P_2 to an event of the form $Y(u_1) = 0$ depends on the value of $x(u_2)$. Since the value assigned by P_1 to the same event depends only on $x(u_1)$, the two probability distributions are mutually inconsistent. At the very least, interference of this sort may lead to ambiguities in the calculation of probabilities.

Consider two disjoint sets of units with associated vectors $X^{(1)}, Y^{(1)}, X^{(2)}, Y^{(2)}$, all regarded as random variables. Lack of interference is equivalent to the condition that the response $Y^{(1)}$ be conditionally independent of $X^{(2)}$ given $X^{(1)}$. The condition is asymmetric in X and Y. As a consequence, the covariate value on unit u' has no effect on the joint distribution for other units. The same term is used in the applied statistical literature ([6], section 2.4; [26]) with a similar meaning, though usually interpreted as a physical or biological property of the system rather than a mathematical property of the model. Without this property, it is difficult to give the model a causal interpretation, so lack of interference is often taken for granted as a logical necessity in applications involving deliberate intervention or assignment of treatment to units.

For applications in which the x-values are generated by a process, the preceding argument is not compelling, and the non-interference condition is in fact unduly restrictive. The classification model in section 3 is derived from an exchangeable bivariate process $(Y(u), X(u))_{u \in \mathcal{U}}$ with finite-dimensional distributions Q_n. The conditional distributions $Q_n(\cdot \mid X = x)$ determine the joint classification probabilities for n units having the given covariate values as generated by the process. This is not a regression model because the non-interference condition (2.1) is not satisfied by the conditional distributions. As a result, the response distribution for a set of units selected on the basis of their covariate values is not easily determined and is not equal to $Q_n(\cdot \mid X = x)$.

We argue that condition (2.1) is unnecessarily strong for certain applications, and that a weaker condition is sufficient for applications in which intervention does not arise. Consider a family of distributions $P_n(\cdot; x)$, one such distribution for each covariate configuration. It may happen that there exists a bivariate process with distributions Q_n such that, for each covariate configuration x and each event $A \subset \mathcal{C}^n$, the conditional distributions satisfy $P_n(A; x) = Q_n(A \mid X = x)$. The distributions $\{P_n(\cdot; x)\}$ are then said to be weakly compatible with one another. If such a bivariate process exists, it is not unique because the marginal distribution of the X-process is arbitrary. Since the units in the bivariate process have no covariates to distinguish one from another, the bivariate process is ordinarily

exchangeable. Lack of interference implies weak compatibility, but the converse is false.

2.2. Logistic regression. In a logistic regression model, the components $Y(u_1), \dots$ are independent, so the joint distributions are determined by the one-dimensional marginal distributions. The dependence on x is determined by a suitable collection of discriminant functions, $f_j \colon \mathcal{X} \to \mathcal{R}$, which could be the coordinate projections if $\mathcal{X} = \mathcal{R}^q$, but might include quadratic or other non-linear functions. For a unit u whose feature value is $x = x(u)$ the class probabilities are

$$\log \operatorname{pr}(Y(u) = r) = \sum_j \beta_{rj} f_j(x),$$

where the coefficients β_{rj} are parameters to be estimated from the training data. In particular, if there are only two classes, the log odds for class 0 are

$$\log\bigl(\operatorname{pr}(Y(u) = 0)/\operatorname{pr}(Y(u) = 1)\bigr) = \sum_j (\beta_{0j} - \beta_{1j}) f_j(x). \tag{2.2}$$

For a model with k classes and q linearly independent discriminant functions, the number of parameters is $q(k-1)$, which can be large.

The lack of interference condition is automatically satisfied by the logistic regression model, and in fact by any similar model with independent components.

3. An exchangeable cluster process

3.1. Random permutations and random partitions. A partition B of the set $[n] = \{1, \dots, n\}$ is a set of disjoint non-empty subsets called blocks whose union is the whole set. The symbol $\#B$ denotes the number of blocks, and for each block $b \in B$, $\#b$ is the number of elements. The partition is also an equivalence relation on $[n]$, i.e. a function $B \colon [n] \times [n] \to \{0, 1\}$ that is reflexive, symmetric and transitive. Finally, B is also a symmetric binary matrix with components $B(i, j)$. No distinction is made in the notation between B as a set of subsets, B as a matrix, and B as an equivalence relation. If the partition is regarded as a matrix, $\#B$ is its rank.

Denote by $\mathcal{B}_n$ the set of partitions of $[n]$. Thus, $\mathcal{B}_2 = \{12, 1|2\}$ has two elements, and $\mathcal{B}_3$ has five elements

$$123, \quad 12|3, \quad 13|2, \quad 23|1, \quad 1|2|3,$$

where $13|2$ is an abbreviation for $\{\{1, 3\}, \{2\}\}$, containing two blocks. The 15 elements of $\mathcal{B}_4$ can be grouped by block sizes as follows

$$1234, \quad 123|4\ [4], \quad 12|34\ [3], \quad 12|3|4\ [6], \quad 1|2|3|4$$

where $12|34$ [3] is an abbreviation for the three distinct partitions $12|34$, $13|24$, $14|23$, each having two blocks of size two. The number of elements in $\mathcal{B}_n$ is the nth

Bell number, the coefficient of $t^n/n!$ in the generating function $\exp(e^t - 1)$. The first few values are 1, 2, 5, 15, 52, 203, 877,.., increasing rapidly with n.

Consider a probability distribution on the symmetric group $\mathcal{S}_n$ in which the probability assigned to the permutation σ depends on the number of cycles as follows:

$$p_n(\sigma) = \lambda^{\#\sigma}\Gamma(\lambda)/\Gamma(n+\lambda), \tag{3.1}$$

where $\lambda > 0$, and the ratio of gamma functions is the required normalizing constant. This is the exponential family generated from the uniform distribution with weight function $\lambda^{\#\sigma}$, canonical parameter $\log\lambda$ and canonical statistic $\#\sigma$ the number of cycles. It is evident that the distribution is invariant under the action of the group on itself by conjugation, so p_n is finitely exchangeable. Less obvious but easily verified is the fact that p_n is the marginal distribution of p_{n+1} under the natural deletion operation $\sigma' \mapsto \sigma$ from $\mathcal{S}_{n+1}$ into $\mathcal{S}_n$, which operates as follows. Write σ' in cycle form, for example $\sigma' = (1,3)(5)(2,6,4)$ for $n = 5$, and delete element $n+1 = 6$ giving $\sigma = (1,3)(5)(2,4)$. This construction, together with the associated Chinese restaurant process, is described by Pitman ([24], section 4). The projection $\mathcal{S}_{n+1} \to \mathcal{S}_n$ is not a group homomorphism, but successive deletions are commutative. For each $\lambda > 0$, these distributions determine an exchangeable permutation process closely related to the Ewens process on partitions.

The cycles of the permutation $\sigma \in \mathcal{S}_n$ determine a partition of the set $[n]$, and thus a map $\mathcal{S}_n \to \mathcal{B}_n$. The inverse image of $B \in \mathcal{B}_n$ contains $\prod_{b\in B}\Gamma(\#b)$ permutations all having the same probability. Thus, the marginal distribution on partitions induced by (3.1) is

$$p_n(B;\lambda) = \frac{\Gamma(\lambda)\lambda^{\#B}}{\Gamma(n+\lambda)}\prod_{b\in B}\Gamma(\#b) \tag{3.2}$$

for $B \in \mathcal{B}_n$ and $\lambda > 0$ ([10], [1]). This distribution is symmetric in the sense that for each permutation $\sigma\colon [n] \to [n]$, the permuted matrix $(B^\sigma)_{ij} = B_{\sigma(i),\sigma(j)}$ has the same distribution as B. The partition B^σ has the same block sizes as B, which are maximal invariant, and the probability $p_n(B;\lambda)$ depends only on the block sizes. In addition if $B' \sim p_{n+1}(\cdot;\lambda)$ is a random partition of $[n+1]$, the leading $n \times n$ submatrix B is a random partition of $[n]$ whose distribution is $p_n(\cdot;\lambda)$ ([19]). For each $\lambda > 0$, the sequence of distributions $\{p_n\}$ determines an exchangeable process called the Ewens partition process. For further details, see Pitman ([25]).

The Ewens process is by no means the only example of an exchangeable partition process, but it is one of the simplest and most natural, and it is sufficient to illustrate the ideas in the sections that follow. Some simple extensions are described by Pitman ([24]).

3.2. Gauss-Ewens cluster process. A cluster process with state space $\mathcal{X}$ is an infinite sequence of $\mathcal{X}$-valued random variables $X(u)$ for $u \in \mathcal{U}$, together with a random partition $B\colon \mathcal{U}\times\mathcal{U} \to \{0,1\}$, which determines the clusters. An observation on a finite set of units $\{u_1,\ldots,u_n\}$ consists of the values $X(u_1),\ldots,X(u_n)$ together with the components of the matrix $B_{ij} = B(u_i,u_j)$.

The finite-dimensional distributions on $\mathcal{B}_n \times \mathcal{X}^n$ with densities p_n satisfy the obvious Kolmogorov consistency condition:

$$p_n(B, x_1, \ldots, x_n) = \sum_{B' : \phi B' = B} \int_{\mathcal{X}} p_{n+1}(B', x_1, \ldots, x_{n+1})\, dx_{n+1}$$

where $\phi\colon \mathcal{B}_{n+1} \to \mathcal{B}_n$ is the deletion operator that removes the last row and column.

In the Gauss-Ewens process, $\mathcal{X} = \mathcal{R}^q$ is a vector space. The observation $(B, (X_1, \ldots, X_n))$ on a finite set of n units has a joint density in which B is a partition with distribution (3.2). The conditional distribution given B is Gaussian with constant mean vector, here taken to be zero, and covariance matrix $\Sigma_B = I_n \otimes \Sigma + B \otimes \Sigma_1$, where Σ, Σ_1 are $q \times q$ covariance matrices. In component form

$$\operatorname{cov}(X_{ir}, X_{js} \,|\, B) = \delta_{ij}\Sigma_{rs} + B_{ij}\Sigma_{1\,rs}.$$

This construction implies that X is a sum of two independent processes, one i.i.d. on the units, and one with i.i.d. components for each block.

If $\mathcal{X} = \mathcal{R}$, the coefficient matrices are scalars and the joint density is

$$p_n(B, x) = \frac{\Gamma(\lambda)\lambda^{\#B}}{\Gamma(n+\lambda)} \prod_{b \in B} \Gamma(\#b) \times (2\pi)^{-n/2} |\Sigma_B|^{-1/2} \exp(-x'\Sigma_B^{-1}x/2).$$

It is helpful here to re-parameterize by writing $\bar{x}_b$ for the mean in block b, $\theta = \sigma_1^2/\sigma^2$ for the ratio of variance components, $w_b = \#b/(1 + \theta\,\#b)$ and $\bar{x} = \sum w_b \bar{x}_b / \sum w_b$, in which case we have

$$|\Sigma_B|^{-1/2} = \sigma^{-n} \prod_{b \in B} (1 + \theta\,\#b)^{-1/2},$$
$$x'\Sigma_B^{-1}x = \sum_{b \in B} (S^2(b) + w_b \bar{x}_b^2)/\sigma^2,$$

where $S^2(b)$ is the sum of squares for block b.

A permutation of the units sends $X_1, \ldots, X_n$ to $X_{\sigma(1)}, \ldots, X_{\sigma(n)}$ and also transforms the components of B in such a way that the i, j component of B^σ is $B_{\sigma(i)\sigma(j)}$. Evidently, the distribution p_n is unaffected by such permutations, so the Gauss-Ewens process is infinitely exchangeable. As it stands, the Gauss-Ewens process is not a mixture of independent and identically distributed processes because the observation space $\mathcal{B}_n \times \mathcal{X}^n$ for a finite set of n units is not an n-fold product space. However, if the blocks are labelled at random, the new process is equivalent in every way to the original, and the new process does follow the de Finetti characterization ([25], p. 44).

3.3. Conditional distributions. Given the observed list of feature values $x = (x_1, \ldots, x_n)$, the conditional distribution on partitions induced by the one-dimensional Gauss-Ewens process is

$$p_n(B \,|\, x) \propto \prod_{b \in B} \lambda\Gamma(\#b)(1 + \theta\,\#b)^{-1/2} \exp\bigl(-S^2(b) - w_b\bar{x}_b^2)/(2\sigma^2)\bigr).$$

This is a distribution of the product-partition type $p_n(B) \propto \prod_{b\in B} C(b;x)$ ([17]) with cohesion function

$$C(b;x) = \lambda\Gamma(\#b)(1+\theta\,\#b)^{-1/2}\exp\bigl((-S^2(b) - w_b\bar{x}_b^2)/(2\sigma^2)\bigr)$$

depending on the feature values of the units in block b only. In particular, $C(b;x)$ does not depend on $\#B$ or on n. Evidently, two sets of units having the same ordered list of feature values are assigned the same conditional distribution. The marginal distribution on $\mathcal{B}_n$ induced from $p_{n+1}(\cdot\,|\,(x,x_{n+1}))$ by deleting the last component, depends on the value of x_{n+1}, so these conditional distributions do not determine a process. However, there is no contradiction here because these are conditional distributions, and the two conditioning events are different. Since they are derived from a bivariate process, the distributions are weakly compatible with one another in the sense of section 2.1.

For the multivariate Gauss-Ewens process, the conditional distributions are not of the product-partition type unless the coefficient matrices are proportional, i.e. $\Sigma_1 = \theta\Sigma$. When this condition is satisfied, the cohesion function is an obvious multivariate analogue of the univariate version.

Product partition distributions are certainly convenient for use in applied work, but the great majority of product partition models are incompatible with any process. Consider for example, the product partition model with cohesion function $C(b,x) = \lambda$, independent of the covariate values. For $\lambda = 1$, the distributions are uniform on each $\mathcal{B}_n$. But the distribution on $\mathcal{B}_n$ induced from the uniform distribution on $\mathcal{B}_{n+1}$ is not uniform. The Ewens distributions with cohesion function $\lambda\Gamma(\#b)$ are the only product partition models that are compatible with an exchangeable process.

3.4. Stochastic classification. Given the observation $(B, x(u_1), \ldots, x(u_n))$ on n units, plus the feature value $x(u')$ on a subsequent unit, we aim to calculate the conditional distribution $p_{n+1}(\cdot\,|\,\text{data})$ on $\mathcal{B}_{n+1}$ given the observed values generated by the process. The only missing piece of information is the block to which unit u' is assigned, so the conditional distribution is determined by the probabilities assigned to the events $u' \mapsto b$ for those blocks $b \in B$ or $b = \emptyset$.

A straightforward calculation for a product partition model shows that

$$\text{pr}(u' \mapsto b\,|\,\text{data}) \propto \begin{cases} C(b\cup\{u'\},(x,x'))/C(b,x) & b \in B \\ C(\{u'\},x') & b = \emptyset \end{cases}$$

where (x,x') is the complete list of $n+1$ observed feature values. For $b \in B$, the cohesion ratio for the univariate Gauss-Ewens process is

$$\#b\gamma^{1/2}\exp\bigl(-\gamma(x' - \theta\#b\bar{x}_b/(1+\theta\#b))^2/(2\sigma^2)\bigr)$$

where $\gamma = (1+\theta\#b)/(1+\theta(\#b+1))$. If $\theta\#b$ is large, blocks whose sample mean are close to x' have relatively high probability, which is to be expected.

The predictive distribution for the general multivariate Gauss-Ewens process involves a ratio of multivariate normal densities. Although preference is given

to larger blocks, the predictive distribution also puts more weight on those classes whose block means are close to x'. If x' is sufficiently far removed from all observed block means, the empty set (new class) is given relatively greater weight. When the empty set is excluded from consideration the parameter λ has no effect, and the predictive distribution is roughly the same as that obtained from the Fisher discriminant model with prior probabilities proportional to class sizes.

4. Point process models

4.1. Permanent polynomial. To each square matrix K of order n there corresponds a polynomial of degree n,

$$\mathrm{per}_t(K) = \sum_\sigma t^{\#\sigma} K_{1\sigma(1)} \cdots K_{n\sigma(n)}$$

where the sum runs over permutations of $\{1, \ldots, n\}$, and $\#\sigma$ is the number of cycles. The conventional permanent is the value at $t = 1$, and the determinant is $\det(K) = \mathrm{per}_{-1}(-K)$. The coefficient of t is the sum of cyclic products

$$\mathrm{cyp}(K) = \lim_{t\to 0} t^{-1} \mathrm{per}_t(K) = \sum_{\sigma:\#\sigma=1} K_{1\sigma(1)} \cdots K_{n\sigma(n)}.$$

For certain types of patterned matrices, the permanent polynomial can be evaluated in closed form or by recursion. Consider, for example, the matrix J of order n such that $J_{ii} = \zeta$ and $J_{ij} = 1$ otherwise. The permanent polynomial is the value $f_n(t)$ obtained by recursion

$$\begin{pmatrix} f_{n+1}(t) \\ h_{n+1}(t) \end{pmatrix} = \begin{pmatrix} \zeta t & n \\ t & n \end{pmatrix} \begin{pmatrix} f_n(t) \\ h_n(t) \end{pmatrix}$$

starting with $f_0(t) = h_0(t) = 1$. In particular, for $\zeta = 1$ and $t = \lambda$ we obtain the value $f_n(\lambda) = \Gamma(n+\lambda)/\Gamma(\lambda)$, which is the normalizing constant in the distribution (3.1).

4.2. Gaussian moments. The permanent polynomial arises naturally in statistical work associated with factorial moment measures of Cox processes as follows. Let Z be a zero-mean real Gaussian process on $\mathcal{X}$ with covariance function $\mathrm{cov}(Z(x), Z(x')) = K(x, x')/2$. The joint cumulant and the joint moment of the squared variables $|Z(x_1)|^2, \ldots, |Z(x_n)|^2$ are

$$\begin{aligned} \mathrm{cum}_n\big(|Z(x_1)|^2, \ldots, |Z(x_n)|^2\big) &= \mathrm{cyp}[K](x_1, \ldots, x_n)/2, \\ E\big(|Z(x_1)|^2 \cdots |Z(x_n)|^2\big) &= \mathrm{per}_{1/2}[K](x_1, \ldots, x_n), \end{aligned}$$

where $[K](x_1, \ldots, x_n)$ is the symmetric matrix of order n whose entries are $K(x_i, x_j)$. More generally, if $\Lambda(x) = |Z_1(x)|^2 + \cdots + |Z_k(x)|^2$ is the sum of squares of k inde-

pendent and identically distributed Gaussian processes, we have

$$\begin{aligned} \operatorname{cum}_n\big(\Lambda(x_1),\dots,\Lambda(x_n)\big) &= \alpha\ \operatorname{cyp}[K](x_1,\dots,x_n), \\ E\big(\Lambda(x_1)\cdots\Lambda(x_n)\big) &= \operatorname{per}_\alpha[K](x_1,\dots,x_n) \end{aligned} \tag{4.1}$$

with $\alpha = k/2$ ([22]). Thus, if Λ is the intensity function for a doubly stochastic Poisson process, the nth order product density at $x = (x_1,\dots,x_n)$ is $\operatorname{per}_\alpha[K](x)$. In other words, the expected number of ordered n-tuples of distinct events occurring in an infinitesimal ball of volume dx centered at $x \in \mathcal{X}^n$ is $\operatorname{per}_\alpha[K](x)\,dx$.

The analogous result for zero-mean complex-valued processes with covariance function $\operatorname{cov}(Z(x), \bar{Z}(x')) = K(x,x')$ and Λ as defined above is the same except that $\alpha = k$ rather than $k/2$. A proof for $\alpha = 1$ can be found in Macchi ([21]), and for general k in McCullagh and Møller ([22]). Although K is Hermitian, the polynomial is real because inverse permutations have conjugate coefficients.

4.3. Convolution semi-group properties. Permanent polynomials also have a semi-group convolution property that is relevant for probability calculations connected with the superposition of independent processes. In describing this property, it is helpful to regard the points $\mathbf{x} = \{x_1,\dots,x_n\}$ as distinct and unordered, so $\mathbf{x}$ is a finite subset of $\mathcal{X}$. Since $\operatorname{per}_\alpha[K](x_1,\dots,x_n)$ is a symmetric function of x, we may write $\operatorname{per}_\alpha[K](\mathbf{x})$ without ambiguity for non-empty sets. For the empty subset, $\operatorname{per}_\alpha[K](\emptyset) = 1$. It is shown in McCullagh and Møller ([22]) that

$$\sum_{\mathbf{w}\subset\mathbf{x}} \operatorname{per}_\alpha[K](\mathbf{w})\operatorname{per}_{\alpha'}[K](\bar{\mathbf{w}}) = \operatorname{per}_{\alpha+\alpha'}[K](\mathbf{x}) \tag{4.2}$$

where the sum is over all 2^n subsets, and $\bar{\mathbf{w}}$ is the complement of $\mathbf{w}$ in $\mathbf{x}$.

Suppose that $\operatorname{per}_\alpha[K](\mathbf{x})$ is the density at $\mathbf{x}$, with respect to some product measure $\mu(dx_1)\cdots\mu(dx_n)$, of a finite point process in $\mathcal{X}$. The convolution property implies that the superposition of two independent processes having the same covariance function K has a distribution in the same family with parameter $\alpha+\alpha'$. Furthermore, the ratio

$$q(\mathbf{w};\mathbf{x}) = \frac{\operatorname{per}_\alpha[K](\mathbf{w})\operatorname{per}_{\alpha'}[K](\bar{\mathbf{w}})}{\operatorname{per}_{\alpha+\alpha'}[K](\mathbf{x})} \tag{4.3}$$

determines a probability distribution on the subsets of $\mathbf{x}$. If in fact some components of $\mathbf{x}$ are duplicated, these duplicates must be regarded as distinct units that happen to have the same x-value, and q is then regarded as a distribution on subsets of the n units. In the extreme case where all components are identical, all components of the matrix $[K](\mathbf{x})$ are equal, and the distribution reduces to

$$q(\mathbf{w};\mathbf{x}) = \frac{\Gamma(\#\mathbf{w}+\alpha)\,\Gamma(\#\bar{\mathbf{w}}+\alpha')\,\Gamma(\alpha+\alpha')}{\Gamma(n+\alpha+\alpha')\,\Gamma(\alpha)\,\Gamma(\alpha')}.$$

In other words, $\#\mathbf{w}$ has the beta-binomial distribution.

The statistical construction ensures that the polynomial $\operatorname{per}_\alpha(K)$ is positive at all positive half-integer values of α provided only that K is real symmetric and positive semi-definite. In view of the convolution property, it is natural to ask whether

the permanent polynomial of a real symmetric positive semi-definite matrix is positive for all $\alpha \geq 1/2$. The numerical evidence on this point is compelling, but so far there is no proof. On the one hand, there exist positive semi-definite symmetric matrices such that $\mathrm{per}_\alpha(K) < 0$ for values in the interval $0 < \alpha < 1/2$. On the other hand, extensive numerical work has failed to produce a positive semi-definite matrix such that the permanent polynomial has a root whose real part exceeds one half. Although no proof is offered, it seems safe to proceed as if $\mathrm{per}_\alpha(K) \geq 0$ for all $\alpha \geq 1/2$ and positive semi-definite symmetric K. In applications where the covariance function is non-negative, the permanent polynomial is clearly positive for all $\alpha > 0$.

4.4. A marked point process. Consider a Poisson process $\mathbf{X}$ in $\mathcal{X}$ with intensity measure μ. In the first instance, $\mathbf{X}$ is a counting measure in $\mathcal{X}$ such that the number of events $\mathbf{X}(A)$ has the Poisson distribution with mean $\mu(A)$. In addition, for non-overlapping sets A, A', the event counts $\mathbf{X}(A)$ and $\mathbf{X}(A')$ are independent. The process is said to be regular if it has no multiple events at the same point and is finite on compact sets. In that case $\mathbf{X}$ is a random subset of $\mathcal{X}$ such that $\mathbf{X} \cap A$ is finite for compact sets A. For linguistic convenience, we use the terminology associated with random sets rather than the terminology associated with random measures or multisets. All processes are assumed to be regular.

A Poisson process driven by a random intensity measure $\Lambda(x)\mu(dx)$ is called a doubly stochastic Poisson process, or a Cox process. Details of such processes can be found in the books by Kingman ([20]) and Daley and Vere-Jones ([7]).

Let μ be a non-random measure in $\mathcal{X}$ serving as a baseline for the construction of subsequent point processes. For probabilistic purposes, μ is a fixed measure defined on a suitable algebra of subsets of $\mathcal{X}$ that includes all singletons. For statistical purposes, μ is a parameter to be estimated, if necessary, from the data. Given a random non-negative intensity function $\Lambda(x)$, the associated Cox process is such that the expected number of events occurring in an infinitesimal ball dx centered at x is $E(\Lambda(x))\,\mu(dx)$. Likewise, the expected number of ordered pairs of distinct events in the infinitesimal product set $dx\,dx'$ at (x, x') is $E(\Lambda(x)\Lambda(x'))\,\mu(dx)\,\mu(dx')$, and so on. In general, for $\mathbf{x} = (x_1, \ldots, x_n)$,

$$m^{(n)}(\mathbf{x}) = E\big(\Lambda(x_1) \cdots \Lambda(x_n)\big)$$

is called the nth order product density at $\mathbf{x} \in \mathcal{X}^n$. These expectations are the densities of the factorial moment measures of the process with respect to the product measure μ^n. The order is implicit from the argument $\mathbf{x} \in \mathcal{X}^n$, so we usually write $m(\mathbf{x})$ rather than $m^{(n)}(\mathbf{x})$.

Ordinarily, in typical ecological applications or studies of the spatial interactions of particles, an observation on a point process consists of a census $\mathbf{X} \cap S$ of all events occurring in the bounded set S. The observation tells us not only that an event occurred at certain points in S, but also that no events occurred elsewhere in S. For the sorts of applications with which we are concerned, however, the training sample is not exhaustive, so the observation is regarded as a simple random sample of the events in $\mathcal{X}$. Such an observation tells us only that

an event occurred at certain points in $\mathcal{X}$, and says nothing about the occurrence or non-occurrence of events elsewhere.

Suppose now that $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)}$ are k independent Cox process on $\mathcal{X}$ driven by independent random intensity functions $\Lambda_1(x), \ldots, \Lambda_k(x)$, all relative to the same measure μ. The marked process can be represented by the pair $(\mathbf{X}, y)$ in which $\mathbf{X} = \cup \mathbf{X}^{(r)}$ is the superposition process, and $y\colon \mathbf{X} \to \mathcal{C}$ is the list of labels. Then the rth component process $\mathbf{X}^{(r)} = y^{-1}(r)$ is the inverse image of label r.

Let $\mathbf{x} \subset \mathcal{X}$ be a given finite point configuration consisting of n points. Given that $\mathbf{x} \subset \mathbf{X}$, i.e. that the superposition process contains $\mathbf{x}$, each event $x \in \mathbf{x}$ has a label $y(x)$ in the marked process so there are k^n possible values for the labels of the events in $\mathbf{x}$. Denote by $\mathbf{x}^{(r)}$ the subset $\mathbf{x} \cap y^{-1}(r)$, possibly empty, consisting of those events in $\mathbf{x}$ having label r. The conditional distribution of the class labels given $\mathbf{x} \subset \mathbf{X}$ is proportional to the product of the product densities of the component processes

$$p_n(y \,|\, \mathbf{x}) = \frac{\prod_{r \in \mathcal{C}} m_r(\mathbf{x}^{(r)})}{m_\cdot(\mathbf{x})} \tag{4.4}$$

In this expression, $m_r(\mathbf{x}^{(r)})$ is the product density of order $\#\mathbf{x}^{(r)}$ at $\mathbf{x}^{(r)}$ for the process labelled r, and $m_\cdot(\mathbf{x})$ is the nth order product density for the superposition process at $\mathbf{x}$. For the empty set, $m_r(\emptyset) = 1$. A key point to note is that the conditional distribution of the class labels depends only on the product densities, and not on the baseline measure μ.

The conditional distribution of the unlabelled partition B is obtained by ignoring labels, in effect by multiplying by the combinatorial coefficient $k!/(k - \#B)!$. Since the combinatorial coefficient depends on the number of blocks, the conditional distribution of the unlabelled partition is not a product partition model, but it is a distribution of Gibbs type ([25], p. 26)

These conditional distributions do not determine a regression model because they fail to satisfy the lack of interference condition (2.1). However, they are derived from a bona fide bivariate process, so they are mutually compatible in the weak sense.

In this context of prediction, it may be helpful to think of each event as a unit or specimen, in such a way that $x(u)$ is the position or feature value of the event, and $y(u)$ is the label. To classify a new unit or event u' such that $x(u') = x'$, it is sufficient to calculate the conditional distribution as determined by p_{n+1} given the extended configuration $\mathbf{x}' = \mathbf{x} \cup \{x'\}$ plus the labels of those points in $\mathbf{x}$. The conditional probabilities are proportional to the ratio of product densities

$$p_{n+1}(y(u') = r \,|\, \text{data}) \propto m_r(\mathbf{x}^{(r)} \cup \{x'\})/m_r(\mathbf{x}^{(r)}) \tag{4.5}$$

for $r \in \mathcal{C}$.

4.5. Specific examples. We consider two examples, one in which the intensity is the square of a Gaussian process with product density (4.1), and one in which the intensity is log normal.

Permanent process
Suppose that each component process is a permanent process and that the product density for process r is $m_r(\mathbf{x}) = \mathrm{per}_{\alpha_r}[K](\mathbf{x})$. Then the product density for the superposition process is $\mathrm{per}_{\alpha_\cdot}[K](\mathbf{x})$ and the conditional distribution of the labels given $\mathbf{x}$ is

$$p_n(y \mid \mathbf{x}) = \frac{\mathrm{per}_{\alpha_1}[K](\mathbf{x}^{(1)}) \cdots \mathrm{per}_{\alpha_k}[K](\mathbf{x}^{(k)})}{\mathrm{per}_{\alpha_\cdot}[K](\mathbf{x})} \tag{4.6}$$

This distribution determines a random labelled partition of the given events into k classes, some of which may be empty. It is the 'multinomial' generalization of (4.3), and is closed under aggregation of classes.

For a new unit u' such that $x(u') = x'$, the conditional probability of class r is proportional to the permanent ratio

$$p_{n+1}(y(u') = r \mid \mathrm{data}) \propto \mathrm{per}_{\alpha_r}[K](\mathbf{x}^{(r)}, x') / \mathrm{per}_{\alpha_r}[K](\mathbf{x}^{(r)}).$$

This expression is restricted to the set of k classes in $\mathcal{C}$, but it may include classes for which $\mathbf{x}^{(r)}$ is empty, i.e. named classes that do not occur in the training sample. In the extreme case where $\mathbf{x}$ is empty, the probability of class r is $\alpha_r/\alpha_\cdot$ regardless of x'.

The derivation of the conditional distribution from the marked point process requires each α to be a half-integer, and K to be positive semi-definite. Alternatively, K could be Hermitian and α_r a whole integer. However, if K is non-negative on $\mathcal{X}$, the distribution (4.6) exists for arbitrary $\alpha_r > 0$, even if K is not positive semi-definite. We shall therefore consider the limit in which $\alpha_r = \alpha$ and $k \to \infty$ such that $\alpha_\cdot = k\alpha = \lambda > 0$ is held fixed. The limit distribution for the unlabelled partition is

$$p_n(B \mid \mathbf{x}; \lambda) = \frac{\lambda^{\#B} \prod_{b \in B} \mathrm{cyp}[K](\mathbf{x}^{(b)})}{\mathrm{per}_\lambda[K](\mathbf{x})}, \tag{4.7}$$

which is a product partition model, and reduces to the Ewens distribution if K is constant on $\mathcal{X}$. For a new unit u' such that $x(u') = x'$, the conditional probability of assignment to block b is

$$p_{n+1}(u' \mapsto b \mid \mathrm{data}) \propto \begin{cases} \mathrm{cyp}[K](\mathbf{x}^{(b)}, x') / \mathrm{cyp}[K](\mathbf{x}^{(b)}) & b \in B \\ \lambda & b = \emptyset. \end{cases}$$

Our experience with these classification rules is restricted to the simplest versions of the model in which $\mathcal{X}$ is Euclidean space and $K(x, x') = \exp(-|x - x'|^2/\rho^2)$ or similar versions such as $\exp(-|x - x'|/\rho)$. On the whole, the smoother version is better, and the value of α in (4.6) has only minor effects. It is necessary to select a suitable value of the range parameter ρ, but the qualitative conclusions are the same for all ρ. The region in the $\mathcal{X}$-space for which the predictive probability of class r is high need not be convex or simply connected. In that sense, both of these classification rules are qualitatively different from the one derived from the Gauss-Ewens process.

Log Gaussian Cox processes
Suppose that each component process is log Gaussian, i.e. $\log \Lambda_r$ is a Gaussian process with mean and variance

$$E \log \Lambda_r(x) = \theta_r(x), \qquad \operatorname{cov}(\log \Lambda_r(x), \log \Lambda_r(x')) = K_r(x, x').$$

Then the nth order product density at $x = (x_1, \ldots, x_n)$ is

$$m_r(x) = \exp\Big(\sum_j \theta_r(x_j) + \tfrac{1}{2} \sum_{ij} K_r(x_i, x_j)\Big)$$

Given that $\mathbf{x}$ occurs in the superposition process, the conditional distribution of the labels satisfies

$$\log p_n(y \,|\, \mathbf{x}) = \sum_{x \in \mathbf{x}} \theta_{y(x)}(x) + \sum_{\substack{x, x' \in \mathbf{x} \\ y(x) = y(x')}} \tfrac{1}{2} K_{y(x)}(x, x') + \text{const.}$$

Finally, a new unit with $x(u') = x'$ generated from the process is assigned to class r with probability

$$\log p_{n+1}(y(u') = r \,|\, \text{data}) = \theta_r(x') + \tfrac{1}{2} K_r(x', x') + \sum_{x \in \mathbf{x}^{(r)}} K_r(x', x) + \text{const.}$$

Thus, if $\theta_r(x) = \sum_j \beta_{rj} f_j(x)$ as in section 2.2, and there are only two classes with $K_0 = K_1 = K$, the conditional log odds that the new unit is assigned to class 0 are

$$\sum_j (\beta_{0j} - \beta_{1j}) f_j(x') + \sum_{x \in \mathbf{x}^{(0)}} K(x', x) - \sum_{x \in \mathbf{x}^{(1)}} K(x', x). \tag{4.8}$$

coinciding with (2.4) when $K = 0$.

4.6. Numerical illustration. A simple artificial example suffices to illustrate the qualitative difference between classification models based on Cox processes, and classification models of the type described in section 3. We use the two-class permanent model (4.6) with $\alpha_1 = \alpha_2 = 1$. The feature space is a 3×3 square in the plane, the covariance function is $K(x, x') = \exp(-\|x - x'\|^2/\rho^2)$ with $\rho = 0.5$, and the true class is determined by a 3×3 chequerboard pattern with white in the center square. The training data consists of 90 units, with 10 feature values uniformly distributed in each small square as shown in the first panel of Fig. 1. The second panel is a density plot, and the third panel a contour plot, of the conditional probability that a new unit at that point is assigned to class 'white'. These probabilities were computed by an approximation using a cycle expansion for the permanent ratio.

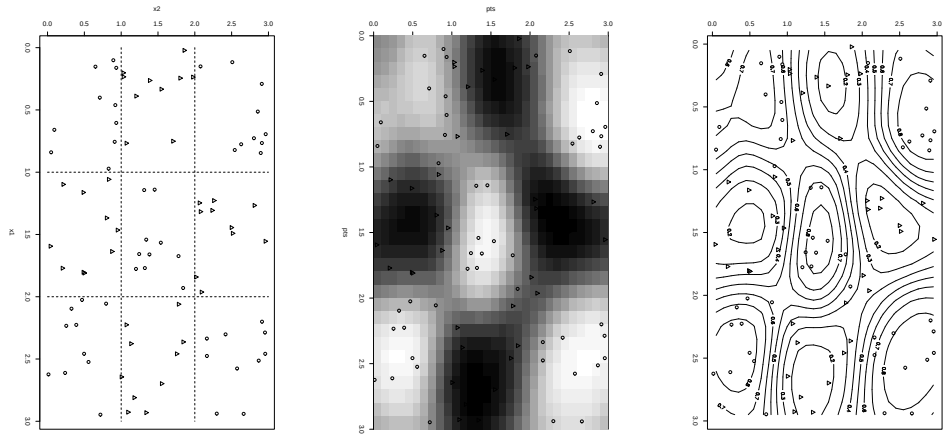

Figure 1. Predictive probability of class I using a permanent model.

For the parameter values chosen, the range of predictive probabilities depends to a moderate extent on the configuration of x-values in the training sample, but the extremes are seldom below 0.1 or above 0.9 for a configuration of 90 points with 10 in each small square. The range of predictive probabilities decreases as ρ increases, but the 50% contour is little affected, so the classification is fairly stable. Given that the correct classification is determined by the chequerboard rule, the error rate for the permanent model using this particular training configuration can be computed exactly: it is around 13% for a point chosen uniformly at random from the large square. This error rate is a little misleading because most of those errors occur near an internal boundary where the predictive probability is close to 0.5. Gross errors are rare.

5. Parameter estimation

Let $(y, \mathbf{x})$ be the training data, and let x' be the feature value for a subsequent unit. In principle, the likelihood function should be computed for the full data including the value for the subsequent unit. In practice, it is more convenient to base the likelihood on the training data alone, i.e. $p_n(y, \mathbf{x}; \theta)$ at the parameter point θ. Ordinarily, the information sacrificed by ignoring the additional factor is negligible for large n, and the gain in simplicity may be substantial.

Likelihood computations are straightforward for logistic regression models, and the same is true for the Gauss-Ewens process, but the state of affairs is more complicated for point process models. Consider a marked permanent process model with $\alpha_r = \alpha$, in which $\mathcal{X}$ is a Euclidean space and $K(x, x') = \exp(-\|x - x'\|^2/\rho^2)$. The parameters of the process are the scalars α, ρ plus the baseline measure μ. However, the conditional likelihood given the observation $\mathbf{x}$ from the training sample depends only on α, ρ, and the predictive distribution also depends only on (α, ρ). In this setting, the distribution of $\mathbf{x}$ is governed largely by the baseline

measure μ, so the information for (α, ρ) in the superposition process must be negligible. Accordingly, we use the conditional likelihood instead of the full likelihood, for parameter estimation.

Even though the most troublesome component of the parameter has been eliminated, computation of the likelihood for the remaining parameters does present difficulties. In the case of the log Gaussian model, the normalizing constant is not available in closed form. In the case of the permanent models (4.6) or (4.7), for which the normalizing constants are available, the only remaining obstacle is the calculation of cyclic products and permanent polynomials. The permanent of a large matrix is notoriously difficult to compute exactly ([28]), and the permanent polynomial appears to be even more challenging. For $\alpha = 1$, polynomial-time algorithms are available for fixed-rank matrices ([3]). In addition, the existence of polynomial-time Monte Carlo algorithms for non-negative matrices, has been demonstrated but not implemented ([18]).

Our experience for positive definite matrices is less pessimistic than the preceding remarks suggest. Reasonably accurate polynomial-time continued-fraction approximations for the ratio of permanent polynomials can be developed without resorting to Monte Carlo approximation. We use a cycle expansion whose accuracy improves as α increases. Here, reasonably accurate means within 2-3% for typical covariance matrices of order $n = 100$, and for $\alpha \geq 1/2$. These expansions, which were used in the construction of Fig 1, will be described elsewhere.

References

[1] Aldous, D., Probability distributions on cladograms. In *Random Discrete Structures* (eds. D. Aldous and R. Pemantle). Springer, New York 1995, 1-18.

[2] Antoniak, C. E., Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2** (1974), 1152-1174.

[3] Barvinok, A. I., Two algorithmic results for the traveling salesman problem. *Mathematics of Operations Research* **21** (1996), 65-84.

[4] Blei, D., Ng, A., Jordan, M., Latent Dirichlet allocation. *J. Machine learning Research* **3** (2003), 993-1022.

[5] Cox, D. R., Some statistical methods connected with series of events. *J. Roy. Statist. Soc. B* **17** (1955), 129-164.

[6] Cox, D. R., *Planning of Experiments*. Wiley, New York, 1958.

[7] Daley D., Vere-Jones, D., *An Introduction to the Theory of Point Processes*, 2nd edition. Springer, New York, 2003.

[8] Daumé, H., Marcu, D., A Bayesian model for supervised clustering with the Dirichlet process prior. *Journal of Machine Learning Research* **6** (2005), 1551-1577.

[9] Donoho, D., Johnstone, I., Kerkyacharian, G., Picard, D., Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. B* **57** (1995), 301-369.

[10] Ewens, W. J., The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3** (1972), 87-112.

[11] Ferguson, T., A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1** (1973), 209-230.

[12] Fisher, R. A., The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** (1936), 179-188.

[13] Freund, Y., Schapire, R., Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kauffman, San Francisco 1996, 148-156.

[14] Gopalan, R., Berry, D., Bayesian multiple comparisons using Dirichlet process priors. *J. Amer. Statist. Assoc.* **93** (1998), 1130-1139.

[15] Green, P., Silverman, B., *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, 1994.

[16] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[17] Hartigan, J. A., Partition models. *Communications in Statistics* **19** (1990), 2745-2756.

[18] Jerrum, M., Sinclair, A., Vigoda, E., A polynomial-time approximation algorithm for approximating the permanent of a matrix with non-negative entries. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*. ACM Press, New York 2001, 712-721.

[19] Kingman, J. F. C., *Mathematics of Genetic Diversity*. SIAM, Philadelphia, 1980.

[20] Kingman, J. F. C., *Poisson Processes*. Oxford, 1993.

[21] Macchi, O., The coincidence approach to stochastic point processes. *Advances in Applied Probability* **7** (1975), 83-122.

[22] McCullagh, P., Møller, J., The permanent process. 2005.
Available via http://www.stat.uchicago.edu/ pmcc/permanent.pdf .

[23] Minc, H., *Permanents*. Addison-Wesley, Reading, MA, 1978.

[24] Pitman, J., Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell* (eds. T. S. Ferguson et al.). IMS Lecture Notes **30,** Hayward, CA 1996, 245-267.

[25] Pitman, J., *Combinatorial Stochastic Processes*. Springer, 2005.

[26] Rubin, D., Comment on Holland (1986): *J. Amer. Statist. Assoc.* **81** (1986), 961-962.

[27] Ripley, B., *Pattern recognition and Neural Networks*. Cambridge University Press, 1996.

[28] Valiant, L. G., The complexity of computing the permanent. *Theoretical Computer Science* **8** (1979), 189-201.

[29] Wahba, G., *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

Dept of Statistics, University of Chicago
E-mail: pmcc@galton.uchicago.edu

Dept of Statistics, University of Chicago
E-mail: jyang@galton.uchicago.edu