J. R. Statist. Soc. A (1999)
162, *Part* 3, *pp.* 383–405

Classical multilevel and Bayesian approaches to population size estimation using multiple lists

Stephen E. Fienberg, Matthew S. Johnson and Brian W. Junker

Carnegie Mellon University, Pittsburgh, USA

[Received February 1999. Revised June 1999]

Summary. One of the major objections to the standard multiple-recapture approach to population estimation is the assumption of homogeneity of individual 'capture' probabilities. Modelling individual capture heterogeneity is complicated by the fact that it shows up as a restricted form of interaction among lists in the contingency table cross-classifying list memberships for all individuals. Traditional log-linear modelling approaches to capture–recapture problems are well suited to modelling interactions among lists but ignore the special dependence structure that individual heterogeneity induces. A random-effects approach, based on the Rasch model from educational testing and introduced in this context by Darroch and co-workers and Agresti, provides one way to introduce the dependence resulting from heterogeneity into the log-linear model; however, previous efforts to combine the Rasch-like heterogeneity terms additively with the usual log-linear interaction terms suggest that a more flexible approach is required. In this paper we consider both classical multilevel approaches and fully Bayesian hierarchical approaches to modelling individual heterogeneity and list interactions. Our framework encompasses both the traditional log-linear approach and various elements from the full Rasch model. We compare these approaches on two examples, the first arising from an epidemiological study of a population of diabetics in Italy, and the second a study intended to assess the 'size' of the World Wide Web. We also explore extensions allowing for interactions between the Rasch and log-linear portions of the models in both the classical and the Bayesian contexts.

Keywords: Log-linear models; Markov chain Monte Carlo methods; Multiple-recapture census; Quasi-symmetry; Rasch model

1. Introduction

Our goal in this paper is to re-examine the problem of estimating the size of a closed population by using multiple lists or sources, often referred to as the multiple-recapture population estimation problem (for example, see Bishop *et al.* (1975)) because of its origins for estimating wildlife and fish populations (for example, see Petersen (1896) and Schnabel (1938)). In effect, we treat our lists as having been generated by sampling multiple times from the population and we identify individuals or objects according to the lists in which they were included.

We wish to estimate N, the unknown size of the population of individuals or objects of interest (e.g. people, fish or software errors), and we do so using the information gleaned from which objects were included in each of the J lists drawn from the population. We let $i = 1, \ldots, N$ index the objects and $j = 1, \ldots, J$ index the lists. Our basic model has $N \times J$ random variables X_{ij} such that

Address for correspondence: Brian W. Junker, Department of Statistics, 232 Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA.
E-mail: brian@stat.cmu.edu

 0964–1998/99/162383

$$X_{ij} = \begin{cases} 1, & \text{if object } i \text{ appears on list } j, \\ 0, & \text{otherwise.} \end{cases}$$

We let $p_{ij} = P(X_{ij} = 1)$ and n be the number of objects that appear on at least one list. Our goal is to estimate the number of unobserved objects, $N - n$, or, equivalently, to estimate N. To do this, we need a model which specifies

(a) the probabilities of appearing in the various lists, i.e. capture probabilities,
(b) how the lists relate to one another, i.e. list dependences, and
(c) the ways in which these capture probabilities and list dependences vary across individuals.

The literature on capture–recapture methods is extensive and goes back many years to at least Petersen (1896). The earliest models for multiple-recapture methods (i.e. more than two lists) assumed that the various captures or lists were independent (e.g. Geiger and Werner (1924) and Schnabel (1938)) and that there were constant capture probabilities across individuals, although not necessarily across captures or lists. Although many researchers expressed concern about the assumption of independence among lists, a general way to cope with this problem awaited other developments in statistics. Thus it was not until the 1970s that Fienberg (1972) introduced the role of log-linear models to provide for dependences among the lists, and Sanathanan (1972) introduced the Rasch model to provide for the dependence induced by heterogeneity across individuals, but for independent lists. Fienberg (1992) and the International Working Group for Disease Monitoring and Forecasting (1995a, b) provide bibliographies of special relevance to the use of these methods in human populations.

The most tractable model that allows for differences in capture probabilities and heterogeneity among individuals is due to Rasch (1960), who derived it for scoring examination items in educational testing:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \theta_i + \beta_j, \qquad i = 1, \ldots, N, \quad j = 1, \ldots, J, \tag{1}$$

a logistic regression model with additive effects for each object i's catchability θ_i and each list j's catch effort β_j. When we set $\theta_i = 0$ in equation (1)—or to any constant independent of i—the log-odds of inclusion of object i on list j depends only on the list, and thus the model reduces to the traditional multiple-recapture model with independent lists. When the θ_i are non-constant and we treat them as random effects, this model is intrinsically multilevel, with lists at one level and individuals at another. Additional multilevel structure may be readily incorporated into this model, through either θ or β, depending on relevant object and list covariates. For example, see Johnson *et al.* (1998) for a Bayesian version of this extension and Wu *et al.* (1997) for a classical or missing data formulation. The Rasch model and its natural generalizations play central roles in our work.

In this paper, we draw on the lessons from what were, until recently, three seemingly separate literatures on

(a) log-linear models for multiple-recapture census problems,
(b) Rasch models for individual heterogeneity and
(c) Bayesian hierarchical model approaches.

For us these three approaches are intimately linked and this paper explores their relationships. In the remainder of this section we review aspects of the literature on heterogeneity, and Bayesian approaches. In Section 2, we introduce three examples to which we later apply

our methodology: data simulated directly from the Rasch model, data from an epidemiological study of a population of diabetics in Italy and data from six 'Web search engines' intended to assess the 'size' of the World Wide Web (WWW). Then, in Section 3, we outline the elements of the Rasch model and its relationship with the usual log-linear models for population size estimation, and in Section 4 we present a fully Bayesian hierarchical approach to the Rasch model which relaxes a seemingly necessary linear constraint in the log-linear formulation and takes into account previously ignored moment inequality constraints. In Section 5, we apply both log-linear models with Rasch-like heterogeneity terms and our Bayesian hierarchical approach to the examples. We shall see that these approaches, applied separately, work reasonably well but seem lacking: the dependence structure in multiple-recapture census data is often similar to the Rasch model, with departures that reflect non-symmetric dependence between lists, or partial symmetry features that represent 'clumpy' heterogeneity of the objects being counted. 'Generalized Rasch' models that allow interactions between parameters that express heterogeneous catchability of objects, or non-symmetric dependences between lists, offer some hope of a more parsimonious representation, and hence smaller standard errors of estimation for the unobserved count. In Section 6, we explore the relationship between generalized Rasch log-linear models and our hierarchical Bayes formulation of the Rasch model.

1.1. *Heterogeneity among individuals and Bayesian approaches*

As we mentioned earlier, Sanathanan (1972, 1973) provided one of the early attempts to look at heterogeneity in the context of capture probabilities. She was interested in scanning experiments in particle physics and focused on a Rasch model in which either the individual or the list parameters are viewed as independent draws from a parametrically specified common distribution. Subsequently, Burnham and Overton (1978), Chao (1987, 1989), Chao *et al.* (1992) and Pollock (1991) built on Cormack's (1966) approach, incorporating heterogeneity into multiplicative models for the p_{ij}, e.g. $p_{ij} = \phi_i \psi_j$. Unfortunately, this part of the literature provides little recognition of the special statistical features that require attention when the number of parameters that are included for heterogeneity increases in direct proportion to the population size N.

Interest in the heterogeneity problem arose again in connection with discussions about the use of capture–recapture methods in the context of the 1990 US decennial census, and Darroch *et al.* (1993) presented a model for heterogeneity based on a log-linear representation of the Rasch model, which they then combined with log-linear models for dependence. Their approach had been anticipated in part by Cormack (1989) but without the link to the Rasch model framework, and then suggested separately by Agresti (1994). International Working Group for Disease Monitoring and Forecasting (1995a, b) provided a simple discussion of these approaches and their application to a problem of estimating the size of a population of diabetics. The introduction of the Rasch model representation for heterogeneity in this example, however, had apparently strange and not totally satisfactory consequences, and we return to their problem in Section 2 below.

Roberts (1967) presented an early Bayesian approach to the simple capture–recapture problem with two lists and with constant capture probabilities. Freeman (1972) introduced a Bayesian approach to sequential estimation and then later (Freeman, 1973) contrasted this with the capture–recapture model, using constant capture probabilities in both settings. He also introduced the role of different loss functions. Castledine (1981) generalized this approach to multiple-recapture studies and derived the marginal posterior distribution of N, assuming

independent prior distributions $p_{ij} \equiv p_j \sim \text{beta}(a, b)$ and $\pi(N) \propto 1$ or $\pi(N) \propto 1/N$ (the Jeffreys prior). Smith (1991) found the posterior distribution of N in this case by using both empirical Bayes and Bayes–empirical Bayes approaches. Garthwaite *et al.* (1995) extended these results to allow for random sample sizes and explored the sensitivity of the posterior for N to the prior specification. Smith (1988) used a Poisson approximation to the hypergeometric distribution of marked items in the sample, and an inverse gamma prior distribution for N, and identified estimators under several loss functions as equivalent to the Geiger–Werner–Schnabel multiple-recapture estimate.

George and Robert (1992) were the first to bring the modern Bayesian technology of Markov chain Monte Carlo (MCMC) estimation to bear on the capture–recapture problem. They built hierarchical Bayes structures beginning with Castledine's (1981) formulation and used Gibbs sampling methods, including the adaptive rejection sampling method of Gilks and Wild (1992), for simulating from the posterior distribution of N. Basu (1998) considered log-additive mixed effects models for the p_{ij} similar to the multiplicative models of Cormack (1966) and Pollock (1991) and gave both the catchability and the catch effort parameters discrete prior distributions. With this set-up Basu found complete conditional distributions for each parameter and hyperparameter, and implemented a Gibbs sampling scheme.

Madigan and York (1995, 1997) pursued the route of hierarchical Bayesian models for log-linear dependences for the multiple-recapture problem with covariates, using the subclass of decomposable graphical models. Instead of estimating N on the basis of a single model, they used Bayesian model averaging. Although we do not pursue the model averaging approach in this paper, it represents a sensible way to extend our approach to account for model uncertainty.

2. Three examples

In this paper, we explore different approaches to the multiple-recapture problem by using three examples. The first uses simulated data linked to the Bayesian hierarchical Rasch model described in Section 4 later. The other two examples arise in actual problems of population size estimation in public health and information sciences. Preliminary analyses of the data in these examples, using what are demonstrably inappropriate models, lead to erroneous inferences (see Hay (1997)).

2.1. Simulated data

Using the basic Rasch model (1), we randomly drew independent results for the presence of $N = 2000$ individuals from each of $J = 6$ lists. We simulated the values of the individual parameters θ_i, for $N = 2000$ subjects from an $N(0, 4)$ distribution, and their presence or absence from each of six lists according to list parameters $\beta = (-1, -0.5, -0.25, 0.25, 0.5, 1)$. The result was a 2000×6 array of 1s and 0s. We summarized this information according to the presence or absence of individuals in the six lists, yielding the 2^6 cross-classification of Table 1. When we analyse these data we shall treat the number of individuals in no lists as if it were unobserved and to be estimated.

In Table 2, we present the classical capture–recapture estimates for N for each pair of lists, using the Petersen estimator

$$\hat{N} = \left[\frac{n_{1+} n_{+1}}{n_{11}} \right], \tag{2}$$

Table 1. 2^6-table of 2000 individuals simulated from a Rasch model

						List 1							
						Yes				*No*			
						List 2				*List 2*			
						Yes		*No*		*Yes*		*No*	
						List 3		*List 3*		*List 3*		*List 3*	
						Yes	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
List 4	Yes	List 5	Yes	List 6	Yes	331	58	58	28	97	52	46	50
					No	14	6	6	7	15	10	17	37
			No	List 6	Yes	25	7	12	11	23	20	25	42
					No	4	1	1	8	8	9	16	37
	No	List 5	Yes	List 6	Yes	30	12	18	12	28	37	24	55
					No	3	6	9	6	8	15	16	70
			No	List 6	Yes	3	4	4	10	10	21	26	108
					No	2	1	5	11	10	21	30	**304**

Table 2. Traditional capture–recapture estimates for N using pairs of lists from Table 1

List	*List 1*	*List 2*	*List 3*	*List 4*	*List 5*
2	1253				
3	1254	1347			
4	1335	1416	1431		
5	1394	1457	1515	1534	
6	1472	1512	1564	1572	1623

where n_{1+} is the number of objects in list 1, n_{+1} is the number of objects in list 2, n_{11} is the number of objects in both lists and $[x]$ is the greatest integer contained in x. Note that all 15 estimates of N, which assume that the lists are pairwise independent and the objects homogeneous, lie below the true value of 2000, but more importantly below the observed number of objects in all six lists, i.e. $n = 2000 - 304 = 1696$. This is strong evidence regarding the positive dependence among the lists induced by the heterogeneity. One of the bench-marks of the analyses to come for these data will be the extent to which they adequately deal with this dependence in a parsimonious fashion.

2.2. *Multiple sources for diabetes ascertainment*

Bruno *et al.* (1994) used four sources to identify known cases of diabetes among the residents of the area of Casale Monferrato in northern Italy on October 1st, 1988: *clinics*, a list of all patients with a previous diagnosis of diabetes via clinics and/or family physicians; *hospitals*, a list of all patients discharged with a primary or secondary diagnosis of diabetes in all public and private hospitals in the region; *prescriptions*, a computerized database list of insulin and

Table 3. Data from prevalent cases of known *diabetes mellitus* for residents of Casale Monferrato, Italy, on October 1st, 1988, according to four sources of ascertainment

Prescriptions	*Reimbursements*	*Clinics*			
		Yes		*No*	
		Hospitals		*Hospitals*	
		Yes	*No*	*Yes*	*No*
Yes	Yes	58	46	14	8
Yes	No	157	650	20	182
No	Yes	18	12	7	10
No	No	104	709	74	?

Table 4. Traditional capture–recapture estimates for N using pairs of sources from Table 3

	Clinics	*Hospitals*	*Prescriptions*
Hospitals	2351		
Prescriptions	2185	2052	
Reimbursements	2262	803	1555

oral hypoglycemic prescriptions for 1988; *reimbursements*, a list of all residents of a region who requested a reimbursement for insulin and reagent strips.

We reproduce the data here as Table 3. Bruno *et al.* (1994) described a detailed analysis using log-linear models, including the use of stratification to reduce heterogeneity. Their best estimates for N remained in the neighbourhood of 2700, which is substantially in excess of the total observed number of cases in Table 3, i.e. $n = 2069$. When we look at sources in pairs and compute the standard capture–recapture estimates for N as we did in the previous example, we obtain the results in Table 4. Three of the six pairwise estimates fall below 2069 and the other three also lie well below the value reported by Bruno *et al.* (1994), which is quite unsatisfactory, indicative of the failure of the assumptions of independent lists and homogeneous objects. Unlike our simulation example, however, we have wide variation in the estimates of N and we may need to cope with both heterogeneity and dependence.

2.3. *The number of pages on the World Wide Web*

Lawrence and Giles (1998) studied the coverage and recency of six major and widely available WWW search engines by submitting 575 queries on various scientific topics.

These six search engines have built-in positive and negative associations with one another based on how they parse the queries, on how Web pages come to be in each search engine's database and on dependences that would be induced for example if some engines were to use other engines, or were to work from a common set of index pages, to develop part of the set of pages that match a particular query. Some of this information is proprietary with the search engine provider, and hence unknown to us, and therefore it cannot be directly incorporated into a statistical model for the multiple-recapture data. Indeed one of our central goals is to develop statistical models that are sufficiently flexible that they can capture dependence due to such hidden relations (if they exist), yet sufficiently parsimonious to produce useful estimates of the missing number.

Table 5. Multiple-list data for query 535, obtained from Lawrence and Giles (personal communication)

						Northern Light							
						Yes				*No*			
						Lycos				*Lycos*			
						Yes		*No*		*Yes*		*No*	
						Hot Bot		*Hot Bot*		*Hot Bot*		*Hot Bot*	
						Yes	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
Alta Vista	Yes	Infoseek	Yes	Excite	Yes	0	0	5	0	1	0	2	0
					No	1	2	5	2	0	1	5	0
			No	Excite	Yes	1	0	2	1	0	0	0	6
					No	4	0	14	7	1	1	7	33
	No	Infoseek	Yes	Excite	Yes	0	0	2	2	1	1	2	4
					No	0	0	5	0	1	0	11	8
			No	Excite	Yes	1	1	2	5	0	0	3	21
					No	3	2	13	35	0	2	79	?

Table 6. Traditional capture–recapture estimates for the total N Web pages matching query 535, using pairs of search engines

	Alta Vista	*Infoseek*	*Hot Bot*	*Northern Light*	*Excite*
Infoseek	256				
Hot Bot	359	254			
Northern Light	294	274	362		
Excite	353	192	489	309	
Lycos	202	183	293	172	252

It is also important to realize that, in principle, each of the 575 queries defines a different population of pages, some of which are observed in the corresponding $2^6 - 1$ layer of the full $575 \times (2^6 - 1)$ table. Thus, we tentatively think of query as a stratifying variable to go with our models. Keeping this stratification in mind, we restrict our attention for now to one layer of the $575 \times (2^6 - 1)$ table, corresponding to a single query, 535, listed in Table 5. Further analysis of these data, which is ongoing, could and should try to incorporate query as a multilevel stratifying variable in the models.

As with the other two examples, we begin our analyses by considering the lists in a pairwise fashion and computing the traditional capture–recapture estimates for the population size, which we give here in Table 6. The total number of observed pages across all six search engines is $n = 305$, and only five of the 15 estimates in Table 6 exceed this value. Thus there is evidence from these marginal calculations suggesting that 10 of the 15 pairs of search engines are positively dependent; there may be some counterbalancing negative dependence among the remaining five pairs. In any case, it seems unlikely that a model asserting joint independence for all six lists will produce very satisfactory population size estimates.

3. The Rasch model and quasi-symmetric log-linear models

We can think of the Rasch model, which we introduced in Section 1, as a mixed effects generalized linear model that allows for object heterogeneity and list heterogeneity. For object i, we model the probability of inclusion on list j, as in equation (1), where θ_i is the random catchability effect for object i, which is distributed as a random variable from F_Θ, and the β_j are fixed parameters representing the penetration of list j into the target population. The heterogeneity of capture probabilities across objects is therefore influenced by the distribution of θ, F_Θ.

Let $X_1, \ldots, X_J$ be the variables cross-classified in $2^J - 1$ tables like Tables 1 and 5. Mixed effects models for such tables with a random individual effect θ are a way of thinking about disaggregating the table according to values of θ, and then reaggregating. In the disaggregated table, let

$$P_j(\theta) = P(X_j = 1|\theta);$$

usually we assume that the lists are independent given θ, so that the probability of observing a count in cell $k_1 \ldots k_J$, given fixed θ, is

$$\pi_{k_1 \ldots k_J|\theta} = P(X_1 = k_1, \ldots, X_J = k_J|\theta) = \prod_{j=1}^{J} P_j(\theta)^{k_j}\{1 - P_j(\theta)\}^{1-k_j}. \tag{3}$$

Reaggregating into the $2^J - 1$ table is just integrating over θ; thus the marginal probability of observing a count in cell $k_1 \ldots k_J$ is simply

$$\pi_{k_1 \ldots k_J} = P(X_1 = k_1, \ldots, X_J = k_J) = \int \prod_{j=1}^{J} P_j(t)^{k_j}\{1 - P_j(t)\}^{1-k_j}\, \mathrm{d}F_\Theta(t). \tag{4}$$

The Rasch model specifies additive logits, $\log[P_j(\theta)/\{1 - P_j(\theta)\}] = \theta + \beta_j$, so

$$\pi_{k_1 \ldots k_J} = \int \exp\left\{\sum_{j=1}^{J} k_j(t + \beta_j)\right\} \prod_{j=1}^{J} \frac{1}{1 + \exp(t + \beta_j)}\, \mathrm{d}F_\Theta(t). \tag{5}$$

Integrating with respect to the distribution for θ in equations (4) and (5) turns the Rasch model into a log-linear model of the form

$$\log(\pi_{k_1 \ldots k_J}) = \alpha + k_1\beta_1 + \ldots + k_J\beta_K + \gamma(k_+) \tag{6}$$

where

$$\begin{aligned} k_+ &= \sum_{j=1}^{J} k_j, \\ \gamma(s) &= \log[E\{\exp(s\theta)|\mathbf{X} = 0\}] \end{aligned} \tag{7}$$

(Cressie and Holland, 1983; Fienberg and Meyer, 1983; Holland, 1990; Darroch *et al.*, 1993). The term $\gamma(k_+)$ models a specific kind of dependence in the $2^J - 1$ table cross-classifying the lists: this dependence is not due to associations between the lists, but rather it arises directly from aggregating across the strata indexed by θ in the model. The value of $\gamma(k_+)$ is not affected by permutations of $k_1, \ldots, k_J$ and hence we have a quasi-symmetry model (for example, see Bishop *et al.* (1975)), which we can fit to the $2^J - 1$ table of observed counts by using standard software for fitting log-linear and generalized linear models (GLMs). Unfortunately, this transformed version of the Rasch model ignores the moment inequalities that are implicit in equation (7) (for example, see Cressie and Holland (1983)), a point to which we return later.

For $J = 3$ lists, the quasi-symmetry model is equivalent (ignoring moment restrictions) to the constraints

$$\pi_{011}\pi_{100} = \pi_{101}\pi_{010} = \pi_{110}\pi_{001}. \tag{8}$$

These constraints do not relate the probability of the unobserved cell, π_{000}, to the other probabilities, and hence an additional assumption such as no J-way interaction is needed. Estimation can then be carried out using traditional methods for log-linear models, and N can be estimated by

$$\hat{N} = n + \hat{m}_{\text{odd}}/\hat{m}_{\text{even}}, \tag{9}$$

where $\hat{m}_{\text{odd}}$ is a product of estimated expected cell values over all cells whose subscripts sum to an odd value and $\hat{m}_{\text{even}}$ is a product of estimated expected cell values over all cells whose subscripts sum to an even value (for example, see Fienberg (1972)).

Incorporating two- and three-way interactions into a log-linear model may not be necessary, and so, as an alternative to the log-linear model (6), we can consider only lower order symmetries and simply set the higher order log-linear interaction terms equal to 0. Again, we can use standard log-linear model or GLM software to fit such models and then project the model to the missing cell by using equation (9).

We fitted all log-linear models in this paper with the `glm()` function in S-PLUS (Mathsoft, 1996). To construct the quasi-symmetry terms and to combine them with various linear constraints such as no highest order interaction, we used the software package `intersect` from the `S` archives at Carnegie Mellon's Statlib (`http://lib.stat.cmu.edu`) (see Darroch *et al.* (1993)). We computed interval estimates for the total number of cases N using the profile likelihood methods of Cormack (1992): the profile likelihood estimate of the $1 - \alpha$ confidence set for N is defined to be $\{N: G^2(N - n) - G^2(\hat{N} - n) < \chi^2_{(1),1-\alpha}\}$ where G^2 is the model deviance and $\chi^2_{(1),1-\alpha}$ is the $1 - \alpha$ quantile of a $\chi^2_{(1)}$-distribution. Because S-PLUS's `glm()` function estimates the capture–recapture model by using a Poisson likelihood, we approximated the multinomial deviance $\hat{G}^2$ from the Poisson fit by

$$\hat{G}^2(u) = D(u) - \log\left\{\frac{u^u(n+u)!}{(n+u)^{n+u}u!}\right\},$$

where $D(u)$ is the deviance for a log-linear Poisson model fit to the 2^J contingency table with u in the missing cell (for example, see Cormack (1992)).

4. Hierarchical Bayes formulation of the Rasch model

If one takes subject heterogeneity as modelled by equation (4) at all seriously, then the log-linear quasi-symmetry approach has two deficiencies. First, the moment constraints that are implicit in equations (7) are not easily incorporated into GLM fits and hence are usually ignored. Second, the need for and use of the 'no k-way interactions' assumption (for $2 < k \leqslant J$) does not translate into a natural condition on the conditional capture probabilities $P_j(\theta)$ or the catchability distribution $F_\Theta(t)$.

An alternative approach is to estimate the parameters β_j and any parameters of $F_\Theta(t)$ directly from the marginal likelihood (5), by maximum likelihood say, and to use the constraints that are implicit in this formulation to project an estimate onto the missing cell. Coull and Agresti (1999) did exactly this, for example, replacing $F_\Theta(t)$ with a discrete distribution motivated from Gaussian quadrature.

We prefer to work with a fully Bayesian hierarchical specification of the Rasch model. This allows us to lay out all the pieces of the model and to modify exactly those parts that need adjustment to reflect the dependence in the data. We can use MCMC computing methodology to give essentially exact inferences for remarkably complex models in which the log-linear and marginal maximum likelihood approaches become unwieldy.

4.1. Model formulation

We begin our formulation of the basic Rasch model as a hierarchical Bayes model as follows:

$$\left.\begin{aligned} X_{ij} &\overset{\text{indep}}{\sim} \text{Bernoulli}(p_j|\theta_i), \quad \log\left\{\frac{P_j(\theta_i)}{1-P_j(\theta_i)}\right\} = \theta_i + \beta_j, && i = 1, \ldots, N, \\ & && j = 1, \ldots, J, \\ \theta_i &\overset{\text{IID}}{\sim} F_\Theta(\theta_i), && i = 1, \ldots, N, \\ \beta_j &\overset{\text{IID}}{\sim} G_\beta(\beta_j), && j = 1, \ldots, J. \end{aligned}\right\} \tag{10}$$

Note that we only need to add prior distributions $G_\beta(\cdot)$ for the β_j to the development of the marginal–mixed effects model of equation (5).

When N is unknown and the objects in the $0 \ldots 0$ cell of the table are missing, we treat N as a parameter in the likelihood for the $2^J - 1$ cross-classifying the n observed objects

$$L(N, \theta, \beta; \mathbf{X}) = \binom{N}{n} \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{\frac{\exp(\theta_i + \beta_j)}{1+\exp(\theta_i + \beta_j)}\right\}^{x_{ij}} \left\{\frac{1}{1+\exp(\theta_i + \beta_j)}\right\}^{1-x_{ij}} \tag{11}$$

where $\mathbf{X}$ is an $N \times K$ matrix, with the (i, j) element $x_{ij} = 1$ if object i is on list j and 0 otherwise. We shall denote the n rows of observed data in $\mathbf{X}$ as $\mathbf{X}_1$. The remaining rows from $n+1$ to N of $\mathbf{X}$ are all $\mathbf{0}$s, vectors of 0s.

We assume in model (10) that the vector of list parameters, β, is independent of the vector (θ, σ^2, N) and that the list parameters are distributed as $\beta \sim N_K(\mu, \tau_b \cdot I_K)$. We also assume in model (10) that, conditionally on the variance σ^2 and the population size N, the catchability parameter vector is distributed as $\theta \sim N_N(\mathbf{0}, \sigma^2 I)$, and that $\sigma^2 \sim \Gamma^{-1}(\alpha, \eta)$.

As usual, the hyperparameters of the list vector β can be chosen to reflect any available prior information. For our analyses, we have chosen $\beta \sim N_K(\mathbf{0}, 10 I_K)$ because we have little prior information about the lists. Similarly we may fix the hyperparameters for the variance of the catchability parameters according to any prior knowledge that we may have solicited from the researcher; we have set $\alpha = \eta = 1$. Note, following the remarks after equation (1) in Section 1, that if the prior for σ^2 concentrates mass near 0 the model will perform similarly to the independent list model.

Finally, we assume that the prior distribution of N is

$$f_{\mathbb{N}}(N) \propto \frac{(N-l)!}{N!} I_{\{n \le N \le N_{\max}\}}.$$

This form of prior ensures that the conditional posterior for N is a truncated negative binomial; see equation (12) below. As l increases, the prior distribution for N is more concentrated near the observed number n. We have chosen to use $l = 1$, or the Jeffreys prior $f_{\mathbb{N}}(N) \propto 1/N$.

In principle, we can use any prior on N that leads to a proper posterior distribution, but we have found that restricting N to have finite support on the integers, say $[n, N_{\max}]$ for some value $N_{\max}$, is helpful in our MCMC runs. For the examples that we present below, we have

typically taken $N_{\max}$ to be 10000. In our examples, truncating at $N_{\max} = 10000$ gives up essentially no posterior probability. The specification of the shape of the prior distribution on N is somewhat more important and we shall return to it in Section 6.4.

4.2. Estimation by Markov chain Monte Carlo simulation

Given N and the complete 2^J-table, MCMC estimation of the posterior distributions for the parameters in the Rasch model is a straightforward exercise (Patz and Junker, 1999). Here we propose an extension to the MCMC procedure for the Rasch model for multiple-recapture population estimation. It is similar to the extensions of the binomial–logit model by George and Robert (1992), and the log-additive model by Basu (1998).

From the model specifications in Section 4.1, we can readily deduce that the complete conditional posterior distribution for each $\beta_1, \ldots, \beta_J$ is

$$f_{\mathrm{B}}(\beta_j|N, \theta, \mathbf{X}) \propto \frac{\exp(\beta_j x_{+j})}{\prod_{i=1}^{N} \{1 + \exp(\theta_i + \beta_j)\}} f_{\mathrm{B}}(\beta_j)$$

where $x_{+j} = \Sigma_{i=1}^{n} x_{ij}$, and $f_{\mathrm{B}}(\cdot)$ is the normal prior density indicated above. Similarly, for σ^2, the complete conditional posterior is

$$\sigma^2|N, \theta \sim \Gamma^{-1}\left(\alpha + n, \eta + \sum_{i=1}^{N} \theta_i^2 \Big/ 2\right).$$

Note that, given θ, σ^2 is drawn independently of the data $\mathbf{X}$.

As Basu (1998) noted, a problem occurs when N is conditioned on θ since length$(\theta) = N$ would tell us N, and we would not have to estimate it. For this reason we follow Basu (1998) in computing a joint conditional posterior distribution $f_{\mathbb{N},\theta}(N, \theta|\mathbf{X}_1, \beta, \sigma)$ for (N, θ) together, and then breaking this apart as

$$f_{\mathbb{N},\theta}(N, \theta|\mathbf{X}_1, \beta, \sigma) = f_{\mathbb{N}}(N|\mathbf{X}_1, \beta, \sigma)\, f_{\theta|\mathbb{N}}(\theta|N, \mathbf{X}_1, \beta, \sigma)$$

for the simulation: we first draw N from $f_{\mathbb{N}}(N|\mathbf{X}_1, \beta, \sigma)$, and then we draw θ from

$$f_{\theta|\mathbb{N}}(\theta|N, \mathbf{X}, \beta, \sigma) = \prod_{i=1}^{N} f_{\theta_i}(\theta_i|\mathbf{X}, \beta, \sigma).$$

The (incomplete) conditional posterior for N is

$$\begin{aligned} f_{\mathbb{N}}(N|\mathbf{X}_1, \beta, \sigma) &\propto \binom{N}{n} f_{\mathbb{N}}(N) \prod_{i=n+1}^{N} \int \prod_{j=1}^{J} \frac{1}{1 + \exp(\theta_i + \beta_j)} f_{\theta_i}(\theta_i|\sigma)\, \mathrm{d}\theta_i \\ &\propto \binom{N}{n} f_{\mathbb{N}}(N)\, P(\mathbf{0}|\beta, \sigma)^{N-n}. \end{aligned} \tag{12}$$

We can determine the probability $P(\mathbf{0}|\beta, \sigma)$, which is the probability associated with the unobserved cell, i.e. $\mathbf{X} = \mathbf{0}$, analytically for some priors $f_\theta(\theta|\sigma)$, but in most cases we must approximate or compute it by numerical integration. Equation (12) also shows that the information in the observed data to estimate N is concentrated in $P(\mathbf{0}|\beta, \sigma)$, which is completely determined by data-based estimates of the β_j and σ. Thus we expect to find large posterior correlations between N and the β_j and σ, and that is indeed what happens when we estimate the model.

When $f_{\mathbb{N}}(N) \propto I_{\{N>n\}}$, $I_{\{N>n\}}/N$, $I_{\{N>n\}}/N(N-1)$, etc., $N-n$ has a truncated negative binomial conditional posterior. After we have simulated N, we simulate from the complete conditional posterior for θ:

$$f_{\theta_i|\mathbb{N}}(\theta_i|\beta, \sigma, N, \mathbf{X}) \propto \frac{\exp(-\theta_i^2/2\sigma^2 + \theta_i x_{i+})}{\prod_{j=1}^{J} \{1 + \exp(\theta_i + \beta_j)\}}$$

for $i = 1, \ldots, N$, where $x_{i+} = \Sigma_{j=1}^{J} x_{ij}$.

The time taken to complete a sufficiently long run for this algorithm is highly dependent on the data set. This is because each step of the Markov chain is $O(N)$ and N is being estimated. The algorithm yielded approximately 10000 simulations per hour for the WWW data set, on a Hewlett Packard 9000/770 UNIX workstation. In our analyses, we took 50000 simulations after a burn-in period of 10000 simulations. Long runs are necessary because the model parameters are highly correlated. For example, in the diabetes analysis, the posterior correlation between the population size N and a particular list parameter β_j is around -0.81 and the correlation between N and σ^2 is 0.82. For this reason the Markov chain is very slow to mix.

An alternative to the Basu (1998) approach is to treat the distribution of the observed data $P(\mathbf{X}_1|N, \beta, \sigma)$ implied by equation (11) as a different model for $\mathbf{X}_1$, for each different N. This leads to the formulation of the problem of selecting N as a model selection problem for $\mathbf{X}_1$, and the relevant MCMC technique is Green's (1995) 'reversible jump' approach for randomly selecting models from a well-defined model space. We have also implemented this approach; the results are quite similar to Basu's approach outlined above — and may be useful for more complex models — but, in the models that we have compared the approaches with, the Basu technique produces faster and more stable MCMC runs. Fortran programs implementing both the Basu and the Green approaches are available from us (contact `masjohns@stat.cmu.edu`).

5. Initial analyses of the three examples

5.1. The simulated data

We applied the basic log-linear models and Bayesian Rasch model to the simulated data for six lists that we presented earlier in Table 1. Table 7 contains various estimates for N, the size of the simulated population, and 95% intervals. For the classical models these are 95% profile-likelihood-based intervals (see Cormack (1992)), whereas for the Bayesian Rasch model it is a 95% equal-tailed posterior probability interval. We recall that the true total is $N = 2000$. The independence model fits the data poorly (as indicated by the value of the deviance), and it underestimates the true value substantially as well. All the quasi-symmetry log-linear models fit the data reasonably well but they also underestimate the true value. The quasi-symmetry model 95% confidence intervals illustrate somewhat erratic behaviour. The quasi-symmetry model with no second-order interaction is well behaved and has a relatively tight confidence interval which includes the true value. Allowing for a third-order interaction leads to a much lower estimate and an interval which does not include the true value. Finally the estimate with no fifth-order interaction is reasonable but the confidence interval explodes, suggesting some specification problem or a ridge in the likelihood function. See Section 6.3 for further examples of this sort of behaviour.

The Bayesian Rasch model yields a well-behaved posterior distribution, centred close to the true value with a reasonably tight 95% posterior interval. Table 8 contains the estimates

Table 7. Estimates of the population size for 2000 objects and six lists†

Model	*Degrees of freedom*	*Deviance*	*Point estimate*	*95% interval*
Independence	56	1335.44	1701	[1697, 1707]
Quasi-symmetry with no 3-way or higher interactions	55	50.16	1974	[1913, 2046]
Quasi-symmetry with no 4-way or higher interactions	54	42.05	1859	[1796, 1946]
Quasi-symmetry with no 5-way or higher interactions	53	41.46	1932	[1772, 2326]
Quasi-symmetry with no 6-way interactions	52	41.45	1904	[1697, 6536]
Bayesian Rasch model			Median 2019	[1939, 2128]

†Data simulated from a Rasch model (observed $n = 1696$).

Table 8. MCMC estimated posterior mean and quantiles for the list parameters $\{\beta_j\}$ and prior standard deviation σ on the random catchability effects $\{\theta_i\}$, based on 2000 objects simulated from the Rasch model, but where $n = 1696$†

	Mean	*2.5 percentile*	*Median*	*97.5 percentile*	*Actual*
List 1	−1.03	−1.27	−1.02	−0.81	−1.00
List 2	−0.40	−0.64	−0.40	−0.19	−0.50
List 3	−0.29	−0.53	−0.29	−0.08	−0.25
List 4	0.24	0.00	0.24	0.45	0.25
List 5	0.58	0.33	0.58	0.79	0.50
List 6	0.95	0.70	0.94	1.17	1.00
σ	2.10	1.90	2.10	2.32	2.00
N	2022	1939	2019	2128	2000

†Actual parameters used in the simulation of the data are given in the rightmost column.

of list parameters, or catch efforts, $\{\beta_j\}$ and the standard deviation of the random catchability effects $\{\theta_i\}$.

5.2. *The diabetes data*

International Working Group for Disease Monitoring and Forecasting (1995a) and Biggeri *et al.* (1999) have given detailed treatments of the estimation of the log-linear and quasi-symmetry log-linear models for the diabetes data. Thus, in the first block of Table 9, we simply provide some illustrative models in the class outlined in Section 3.

The independence model fits the data poorly, and the confidence bounds are tight and relatively close to the observed value of $n = 2069$. Both of the quasi-symmetry models, QS2 (the Rasch quasi-symmetry model with no three-way interactions) and QS3 (with no four-way interactions), improve substantially on the fit of the independence model, and the QS2 model produces an estimate of N which is reasonably close to the value of the Bayesian information criterion (BIC) model discussed below and has tighter intervals. But the fit of the QS3 model, *with* the second-order interactions included, seems to be 'off', producing a point estimate of N which is much like that of the independence model.

The fourth model, labelled 'BIC', involves all first-order interactions except the interaction between reimbursements and clinics, and was chosen on the basis of a stepwise procedure using the BIC (for example, see Kass and Wasserman (1995)); it provides an extremely good fit to the data. The results for this model are similar to those for the 'best' model reported by International Working Group for Disease Monitoring and Forecasting (1995b) and Bruno *et al.* (1994).

In contrast, the saturated model produces an estimate of the total population that is twice

Table 9. Estimates of the number of *diabetes mellitus* cases in Casale Monferrato, Italy, on October 1st, 1988, using various methods

Model	*Degrees of freedom*	*Deviance*	*Point estimate*	*95% interval*
Independence	10	217.48	2250	[2216, 2288]
QS2†	9	105.64	2669	[2522, 2846]
QS3	8	93.95	2239	[2142, 2425]
BIC‡	5	7.62	2771	[2533, 3112]
Saturated	—	—	5367	—
Bayesian Rasch	—	—	2693§	[2567, 2906]
QS2 + BIC§§	6	8.32	2752	[2531, 3060]
QS3 + BIC*	5	2.04	4152	[2843, 7388]
QS2 + k_+ × prescriptions**	6	67.39	4367	[3111, 6893]
QS2 + k_+ × clinic	6	66.48	7796	[3793, 12068]
QS2 + k_+ × reimbursements	6	77.03	2411	[2306, 2552]
QS2 + k_+ × hospitals	6	65.86	2381	[2257, 2568]
Bayesian k_+ × prescriptions	—	—	2743§	[2558, 2993]
Bayesian k_+ × reimbursements	—	—	2556§	[2446, 2750]
Bayesian k_+ × clinic	—	—	2537§	[2426, 2831]
Bayesian k_+ × hospitals	—	—	2845§	[2664, 3210]

†QS2 indicates the Rasch quasi-symmetry model with no three- or higher way interactions. Similarly QS3 indicates Rasch quasi-symmetry with no four- or higher way interactions.
‡Stepwise BIC selects independence + reimbursements:hospitals + prescriptions:reimbursements + prescriptions:clinic + prescriptions:hospitals + clinic:hospitals.
§Posterior mode.
§§Stepwise BIC starts with QS2 and adds prescriptions:reimbursements + prescriptions:clinic + reimbursements:hospitals.
*Stepwise BIC starts with QS3 and adds prescriptions:reimbursements + prescriptions:clinic + reimbursements:hospitals.
**k_+ × prescriptions indicates one list-by-total interaction involving the prescriptions list, and similarly for the other k_+ × list interaction models shown.

as large as any of the simpler quasi-symmetry models. This behaviour is indicative of overfitting of the saturated model to near-zero margins in the $(2^4 - 1)$ × (total captures) table of counts cross-classifying capture patterns among the four captures with the total number of captures; such fitted near-zero margins contribute to the expected values in the numerator and denominator of the last term in equation (9), producing severe underestimates or overestimates of the number missing.

Finally, the Bayesian Rasch model produces results that are remarkably close to those from the best fitting log-linear model but with a much more parsimonious model. The posterior 95% probability interval is in fact much tighter than the corresponding classical confidence interval for the best fitting log-linear model.

5.3. The World Wide Web data

The classical multiple-recapture independence model fits the data in our WWW example for query 535 quite poorly, and it projects only an additional 75 unseen Web pages, as seen in the first block of Table 10. The quasi-symmetry model restricted to two-way interactions, QS2, provides a more reasonable fit to the data, with estimates of N that are close to those from the BIC-based 'best' standard log-linear model, discussed below. In contrast, the quasi-symmetry models with higher order interactions mostly blow up again; the overfitting of near-zero

Table 10. Estimates of the number of Web pages of type 'query 535' using various estimation methods

Model	*Degrees of freedom*	*Deviance*	*Point estimate*	*95% interval*
Independence	56	148.73	373	[350, 400]
QS2†	55	88.61	614	[498, 778]
QS3	54	83.33	1266	[624, 3296]
QS4	53	82.43	508	[309, 5778]
QS5	52	81.71	861882	[306, ∞]
BIC‡	46	65.62	602	[481, 793]
Bayesian Rasch	—	—	685§	[554, 1525]
QS2 + BIC§§	44	60.69	588	[468, 778]
QS3 + BIC*	44	55.99	1370	[650, 3829]
QS4 + BIC**	43	55.09	526	[309, 6570]
QS2 + k_+ × Alta Vista††	50	81.96	660	[474, 1003]
QS2 + k_+ × Infoseek	50	83.29	810	[570, 1227]
QS2 + k_+ × Excite	50	76.74	687	[514, 978]
QS2 + k_+ × Hot Bot	50	80.70	591	[409, 993]
QS2 + k_+ × Lycos	50	74.23	737	[558, 1021]
QS2 + k_+ × Northern Light	50	81.04	739	[507, 1175]
QS2 + k_+ × Infoseek + k_+ × Excite	46	68.09	857	[566, 1436]
QS2 + k_+ × Excite + k_+ × Lycos	46	68.35	756	[531, 1180]
QS2 + k_+ × Infoseek + k_+ × Lycos	46	64.36	829	[559, 1344]
QS2 + k_+ × Infoseek + k_+ × Hot Bot + k_+ × Lycos	42	63.14	634	[393, 1364]
Bayesian k_+ × Infoseek	—	—	685§	[531, 996]
Bayesian k_+ × Lycos	—	—	695§	[526, 1071]
Bayesian k_+ × Excite	—	—	666§	[559, 1319]

†QS2 indicates the Rasch quasi-symmetry model with no three- or higher way interactions. Similarly QS3 indicates Rasch quasi-symmetry with no four- or higher way interactions, etc.
‡Stepwise BIC selects Alta Vista:Infoseek + Alta Vista:Hot Bot + Alta Vista:Lycos + Alta Vista: Northern Light + Infoseek:Excite + Infoseek:Hot Bot + Excite:Northern Light + Lycos:Northern Light.
§Posterior mode.
§§Stepwise BIC starts with QS2 and adds Alta Vista:Infoseek + Alta Vista:Excite + Alta Vista:Hot Bot + Alta Vista:Northern Light + Infoseek:Lycos + Infoseek:Northern Light + Excite:Hot Bot + Excite:Lycos + Excite:Northern Light + Hot Bot:Lycos + Hot Bot:Northern Light.
*Stepwise BIC starts with QS3 and adds Alta Vista:Infoseek + Alta Vista:Excite + Alta Vista:Hot Bot + Alta Vista:Northern Light + Infoseek:Northern Light + Excite: Hot Bot + Excite:Lycos + Excite:Northern Light + Hot Bot:Lycos + Hot Bot:Northern Light.
**Stepwise BIC starts with QS4 and adds Alta Vista:Infoseek + Alta Vista:Excite + Alta Vista:Hot Bot + Alta Vista:Northern Light + Infoseek:Northern Light + Excite:Hot Bot + Excite:Lycos + Excite: Northern Light + Hot Bot:Lycos + Hot Bot:Northern Light.
††k_+ × Alta Vista indicates one list-by-total interaction involving the Alta Vista list, and similarly for the other k_+ × list interaction models shown.

margins in the $(2^6 - 1) \times$ (total captures) table is especially evident in the QS5 model. We interpret this as evidence that the likelihood function is not well behaved, and it may also provide diagnostic information to suggest that the positive dependence presumed by the underlying Rasch model is not well satisfied by the data.

We chose the model labelled BIC in the first block of Table 10 using a stepwise procedure and the BIC criterion as in the diabetes example above. It includes eight first-order interaction terms:

$$\text{AV:Is} + \text{AV:HB} + \text{AV:Ly} + \text{AV:NL} + \text{Is:Ex} + \text{Is:HB} + \text{Ex:NL} + \text{Ly:NL},$$

where AV denotes Alta Vista, Is denotes Infoseek, HB denotes Hot Bot, Ly denotes Lycos,

NL denotes Northern Light and Ex denotes Excite. Not surprisingly, this model fits the data considerably better, although the goodness-of-fit improvement is not as striking as was the case in the diabetes example. What is especially interesting here is that, whereas we observe only 305 Web pages in total for the six search engines combined, our best estimate for the total population size is 602, with a fair amount of variability about this value.

Once again, the mode for the Bayesian Rasch model is close to that of the best classical log-linear model estimate. Clearly the data are generally not as informative for the missing count as we might have expected. The dramatically greater width of the credible interval for N (compared with that for the confidence interval from the log-linear model) hints at inadequacies of this basic Rasch model. However, we have a clear standard against which to calibrate our search for alternative models: the basic log-linear model with a selection of first-order terms. A careful perusal of the fits of the other models in Table 10, discussed in Section 6, makes clear that adding higher order terms in a standard log-linear fashion is unlikely to make major gains in the fitting and estimation of N. It is only through something akin to a generalization of the hierarchical Rasch model that we could hope to reduce the parameterization and to improve the fit simultaneously.

6. List and latent variable interactions

6.1. List-by-list interactions

A comparison of the model fits and estimated totals in the first blocks of Tables 9 and 10 suggests that a generalization of the quasi-symmetry models allowing us to free some of the two-way, list-by-list, interactions might help in improving the fit and the estimated totals from these models. Models such as

$$\log(\pi_{k_1 \ldots k_J}) = \alpha + k_1\beta_1 + \ldots + k_J\beta_J + \gamma(k_+) + \sum_{j_1 \neq j_2} \sum \beta_{j_1 j_2} k_{j_1} k_{j_2}, \tag{13}$$

as well as those freeing higher order interactions among the lists, were first considered in detail by Kelderman (1984), under the name 'generalized Rasch models', whose interest in them was to develop hierarchically nested alternatives to the null hypothesis that the data follow the log-linear Rasch model of equation (6) for model fitting investigations in educational testing. Cormack (1989) provided an independent alternative development of these ideas. They have also proven useful in extending the log-linear Rasch model to accommodate dependence in the table $\{n_{k_1 \ldots k_J}\}$ that is Rasch like but more general than the 'exchangeable higher moments' structure of the Rasch model (Darroch and McCloud, 1990; Carriquiry and Fienberg, 1998; Biggeri *et al.*, 1999). In particular, the list-by-list interactions allow for negative dependence between some pairs of lists that are excluded by the basic Rasch model's assertion of equal positive associations among all the lists. For example, Jannarone (1986) and Jannarone *et al.* (1990) have developed extensions of the Rasch model that are similar to this to model non-exchangeable dependence between examination items in educational testing. In Section 6.3 we explore some models of this type, using the BIC to select list-by-list interactions to add to the basic quasi-symmetry models.

6.2. List-by-total interactions

Another extension, suggested by the exploration of the log-linear Rasch model for census work by Darroch *et al.* (1993) as well as by Carriquiry and Fienberg's (1998) examination of

the work of Darroch and McCloud (1990), Biggeri *et al.* (1999) and others, is to allow a list × total number of captures model, along the lines of

$$\log(\pi_{k_1 \ldots k_J}) = \alpha + k_1\beta_1 + \ldots + k_J\beta_J + \gamma(k_+) + \sum_l \gamma'(k_l, k_+), \tag{14}$$

or more generally

$$\log(\pi_{k_1 \ldots k_J}) = \alpha + k_1\beta_1 + \ldots + k_J\beta_J + \gamma(k_+^{(1)}, k_+^{(2)}), \tag{15}$$

where $k_+^{(1)}$ and $k_+^{(2)}$ are total numbers of captures in different subsets of the lists. This structure allows us to model partial quasi-symmetry (e.g. partial exchangeability if the βs are all equal) among the lists, which may be appropriate if the method of constructing some lists is very different from that of other lists (Darroch *et al.*, 1993), so that the catchability of objects is quite different for different lists.

Model (15) may be obtained as the marginal distribution of the data after integrating θ out of the basic likelihood (3) in Section 3 as follows. We begin by supposing that θ is multi-dimensional, i.e. $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$. Moreover, we suppose that different sets of lists depend on different θ_i through the Rasch model. For example, suppose that $\theta = (\theta_1, \theta_2)$ and we can partition the lists into I lists that depend only on θ_1 and $J - I$ lists that depend only on θ_2. Then, after we permute the list indices, the likelihood given θ becomes

$$\pi_{k_1 \ldots k_J|\theta_1\theta_2} = \prod_{j=1}^{I} P_j(\theta_1)^{k_j}\{1 - P_j(\theta_1)\}^{1-k_j} \prod_{j=I+1}^{J} P_j(\theta_2)^{k_j}\{1 - P_j(\theta_2)\}^{1-k_j}. \tag{16}$$

This sort of structure was employed by Darroch *et al.* (1993) to model the different visibility of people in administrative lists, compared with their visibility in US census lists and in a post-enumeration survey also conducted by the US Bureau of the Census. The latent variable formulation thus provides a behavioural motivation for models such as equation (15) that is different from those provided by Darroch and McCloud (1990), or Cormack (1994), for similar models.

If, as would usually seem reasonable, the density $f(\theta_1, \theta_2)$ does not factor, then a derivation similar to that leading from equation (4) to equation (6) in Section 3 now leads us to model (15), where $k_+^{(1)}$ is the number of captures in the first I lists and $k_+^{(2)}$ is the number of captures in the remaining lists. Models with a single list-by-total interaction,

$$\log(\pi_{k_1 \ldots k_J}) = \alpha + k_1\beta_1 + \ldots + k_J\beta_J + \gamma(k_+) + \gamma'(k_l, k_+),$$

can be obtained with some algebra from equation (15), after setting $I = 1$ in that model. Such general partial quasi-symmetry terms are not usually considered in log-linear modelling of multiple-recapture data, and they suggest new ways to expand the basic Rasch quasi-symmetry log-linear model (6) to account for 'extra-Rasch' variability in the catchability random effect.

6.3. *Some illustrative fits for the examples*

We fitted these models in two ways. First we fitted log-linear models of the form (15), by analogy with the basic log-linear quasi-symmetry and quasi-symmetry plus list × list models. Second, we fitted the models directly from their Bayesian latent variable formulation using MCMC methods. In the case of the two-dimensional Rasch model (16) underlying the list-by-total interaction submodel (15), we defined θ_1 to underlie $J - 1$ lists and θ_2 to underlie the list for which we want a list-by-total interaction. The prior distribution for (θ_1, θ_2) was taken to be independent and identically distributed $N_2(\mathbf{0}, \Sigma)$, with $\Sigma \sim$ inverse Wishart$_\nu(\Sigma_0)$ and $\nu = 1$

and $\Sigma_0 = I$, the identity matrix. The population size N is assumed to be independent of Σ, the covariance matrix of the catchability distributions. The prior distributions on N and on the list parameters β_j were the same as in the unidimensional model described in expression (10). The MCMC development is exactly analogous to the development in Section 4, except that the inverse gamma complete conditional distribution for σ^2 is replaced by a corresponding inverse Wishart distribution for Σ. The Bayesian–MCMC fits have the advantage, as with the simple Rasch model, that higher order list interactions are constrained more weakly by various inequality restrictions analogous to those in equations (7), rather than by strict linear restrictions. Also as with the simple Rasch model, we shall see later that the Bayesian versions of list × total models often provide better estimates of total populations than do the log-linear versions of these models.

6.3.1. Diabetes

In Section 5.2 we examined the first block in Table 9 and saw that the quasi-symmetry models by themselves did not provide an adequate fit to the data although the quasi-symmetry with no second-order interactions had a reasonable estimate of N, In the second block of Table 9, we show what happens when we combine the QS2 and QS3 models with list-by-list interactions selected by the BIC. The QS2 plus BIC-selected interactions model produces a result that is essentially the same as the best log-linear model selected by using the BIC alone. The QS3 plus BIC-selected interactions model, although fitting the data extremely well, blows up, producing an estimate of the number missing that is double that of any of the simpler quasi-symmetry models. This behaviour, which is similar to what we observed with the more complex models in Sections 5.2 and 5.3, is indicative that the model overfits a near-zero margin of the $(2^4 - 1) \times$ (total captures) table of counts cross-classifying capture patterns across the four lists with the total number of captures.

We also compare models that add a single list-by-total interaction to the basic QS2 model, in the third panel of Table 9. These models provide substantially better fits than QS2 alone, for example, but they do not fit well relative to the saturated model. Moreover the point estimates are not particularly consistent with one another, and all the interval estimates seem wrong in that they do not contain the base-line BIC estimate of 2771. Adding more than one list-by-total term to the basic QS2 model produces marginally better fits but leads to high point estimates indicative of overfitting, and in some cases a loss of degrees of freedom due to the sparseness of the table.

In the last block of Table 9 we have provided estimates for the Bayesian form of the 'Rasch plus list × total' models outlined at the end of Section 6.2, again for all four lists. It can be seen that the point estimates and interval estimates are much more stable and mutually consistent, and the corresponding interval estimates all contain the BIC point estimate of 2771, except for the 'reimbursements × total' model, which just misses it.

This example suggests that the Bayesian Rasch model alone does quite well, and the Bayesian Rasch plus list × total models do a better job of smoothing over the sparse table than do the corresponding log-linear Rasch plus list × total models. There seems to be little in the data to cause us to prefer one model over the other, but it does not matter since these models all give answers that are consistent with one another, and consistent with our best log-linear model, the base-line BIC model.

The parsimony of the Bayesian Rasch and Rasch plus list × total models is attractive. In addition, we can examine and compare posterior information on the catchability effects θ in these models. In the simple unidimensional Bayesian Rasch model for example, the posterior

mode for the variance of θ is approximately 1.24, with an equal-tailed 95% credible interval running from 1.04 to 1.45. When $\theta = (\theta_1, \theta_2)$ is two dimensional, the variance of θ_1 (the catchability parameter underlying $J - 1$ lists) increases slightly to the range 1.38–1.51 (depending on the particular model), with interval estimates that are similar to those of the unidimensional case. The posterior mode for the variance of θ_2 (the catchability parameter underlying the list in the list-by-total interaction) ranges from approximately 3.39 to 11.36 (depending on the model), with 95% intervals ranging from a lower bound of approximately 1.95 to 4.09 to an upper bound of approximately 20.06 to 34.16 (again depending on the model). The correlation between θ_1 and θ_2 has a posterior mode ranging from 0.64 to 0.75, with 95% intervals bounded below by about 0.4. Clearly not much information is available to identify the distribution of θ_2 separately in the model, but including θ_2 does provide the model with a little more freedom to fit the observed table of counts, while still smoothing over sparse margins in the $(2^4 - 1) \times$ (total captures) table.

6.3.2. *World Wide Web*

We now compare our earlier analyses of the WWW data from Section 5.3, in the first block of Table 10, with models in the second block of Table 10 that also contain list-by-list and list-by-total interactions. What is clear from a perusal of Table 10 is that the log-linear models are very unstable when any interactions higher than first order (two-way interactions) are included in the model. However, all the two-way interaction models perform similarly, suggesting a population total estimate of approximately 600 with a confidence interval that runs from approximately 500 to 800.

The middle block in Table 10 also shows some QS2 models with two or even three list-by-total terms included. We see some moderate increase in the point estimates and in the range of the interval estimates, as can be expected when more of the positive dependence in the lists is modelled, but these results are broadly supportive of our general impression about the point and interval estimates for the population total garnered from the two-way interaction models above.

Comparing the three Bayesian Rasch plus list $\times$ total models included in the last block of Table 10 with the simple Bayesian Rasch model from the first block of Table 10 shows that the right list-by-total interaction helps the interval estimate to settle down to something reasonable, but there is little hint to guide us in selecting the list-by-total interaction that we want. Again, sparseness in the table seems to be the culprit. With 305 observations spread across six lists there is an average of fewer than five observations per cell, and in fact there are many empty cells in the table. This produces severe non-smoothness in the $2^6 - 1$ table — let alone the $(2^6 - 1) \times$ (total captures) table — and higher order models tend to track this non-smoothness; they are consequently misled in their estimation of dependence in the table. The hierarchical Bayesian model fits moments of all orders, restricted by the inequalities displayed in equations (7), and it may be that unless further natural restrictions are placed on these interactions the Baysian model will be similarly confused by the relatively sparse table for query 535.

In the simple unidimensional Bayesian Rasch model, the posterior mode for the variance of θ is approximately 1.65, with an equal-tailed 95% credible interval running approximately from 1.20 to 2.23. The posterior mode of the variance for θ_1 (the catchability parameter for $J - 1$ lists in the two-dimensional model) increases to the range 1.97–2.75 (depending on the model) and for θ_2 ranges from 6.15 to 12.35, with correspondingly wide 95% intervals. The posterior mode of the correlation between θ_1 and θ_2 ranges from 0.63 to 0.88, with 95%

intervals bounded below by approximately 0.39. In the Web data therefore, there was little difference between the simple Bayesian Rasch model and the two-dimensional models underlying the Baysian Rasch plus list × total models. This explains the similarity in the point estimates among the various Bayesian Rasch models listed at the end of Table 10.

6.4. Sensitivity of inferences to prior specifications

In the Bayesian Rasch models that we have explored, the posterior 95% intervals for the population size N are not greatly affected by reasonable variations in prior choices. The effects are generally larger for the WWW problem than for the diabetes problem, so we focus our discussion on that problem.

For example, in the simple Bayesian Rasch model for the WWW problem, changing the prior distribution on N from a truncated uniform, $\pi(N) \propto 1$, to a truncated Jeffreys prior, $\pi(N) \propto 1/N$, keeping the σ^2-prior fixed at $\Gamma^{-1}(1, 1)$, changes the 95% posterior interval from [569, 1662] to [554, 1525]. These changes are not large relative to the sizes of the estimates involved; increasing the truncation point $N_{\max}$ above 10000 also did not have a substantial effect on the posterior distribution for N. However, as the prior distribution on σ^2 concentrates near 0, the estimate for N decreases towards the independent lists estimate, as expected. For example, keeping the truncated Jeffreys prior for N and changing the prior on σ^2 from $\Gamma^{-1}(1, 1)$ to $\Gamma^{-1}(5, 0.05)$ reduces the interval estimate to [464, 961].

The prior distribution that we chose for the β_j, $\beta_j \sim N(0, 100)$, accommodates the range of values that main effects and interactions parameters typically attain in fitted log-linear or logistic models. When we relaxed the β-priors to be improper uniform priors, we saw about a 50% increase in the posterior mode for N, with a corresponding change in the upper end point of the posterior 95% confidence interval. Such a relaxation, however, allows parameter values that we do not believe are realistic for such models; restricting the β-priors to be near $N(0, 100)$ leads to much smaller changes in the estimate of N.

As we mentioned earlier, the diabetes problem seems to be less sensitive to these changes; there were virtually no changes in the posterior credible interval estimates for N under similar modifications of the priors on N and σ^2, and when the priors on the βs were relaxed to be uniform the 95% posterior credible interval for N changed only from [2567, 2906] to [2547, 2921].

7. Discussion

In this paper, we have reviewed and extended the by now long history of statistical modelling of multiple-recapture or multiple-list census data for estimating population totals. When there is heterogeneity in the lists' penetration into the target population of objects to be counted, as well as heterogeneity in the catchability of individual objects, modelling is inherently multilevel: there is a level of fixed effects for lists and a level of random effects for objects. We have argued that the Rasch model, borrowed from the educational testing literature, provides a natural starting place for modelling the multilevel structure. We may also incorporate additional multilevel structure into the Rasch model based on observed object or list covariates.

Only recently have we begun to understand how to modify the Rasch model to accommodate list-by-list dependence and/or list-by-total interactions. We can incorporate these interactions directly into the likelihood that relates capture history to the random catchability effect which we view as a latent variable, or we may interpret them as a kind of stratification of the latent variable into multidimensional components by particular captures or lists. Biggeri

et al. (1999) have also tried to interpret list-by-total interactions as manifestations of a stratification of the latent variable by one or more captures, and this remains an interesting and active area of research. We have shown here how to convert these models into extensions of the log-linear quasi-symmetry model associated with the Rasch model, and we have suggested that frequentist analyses of these models using relatively standard GLM programs provide a useful first approximation to a fully Bayesian approach.

When the basic log-linear quasi-symmetry model holds, we have illustrated that the fully Bayesian hierarchical formulation of the Rasch model provides at least as good a population total estimate, and does so more parsimoniously (exploiting a few hyperprior parameters rather than a full set of quasi-symmetry terms in the log-linear model). An important open question in comparing these two approaches is understanding the interplay between the Bayes model's relaxation of the no highest order interaction assumption that is needed in the log-linear model to project an estimate onto the missing cell count in the $2^J - 1$ table cross-classifying list membership for all objects, and the Bayes model's imposition of moment constraints on the quasi-symmetry terms in the log-linear model that are usually not imposed in frequentist GLM fits of the model.

When the basic log-linear quasi-symmetry model does not hold, adding list-by-list or list-by-total interactions, as outlined in the previous section, can greatly improve the log-linear model fit and the population total estimates based on the log-linear models. These are naturally seen as log-linear manifestations of an underlying hierarchical Bayes model, and we demonstrated how to derive them as such. These models lent some additional flexibility to the Bayesian analyses of our examples and produced somewhat more stable estimates than the basic Bayesian Rasch model did.

Acknowledgements

This work was supported in part by grants REC-9720374 and DMS-9705032 from the National Science Foundation to Carnegie Mellon University. We are indebted to Steve Lawrence and Lee Giles for providing us with unpublished data for analysis. Alan Agresti, S. Basu, Annibale Biggeri, Richard Cormack, Jon Forster, Ed George, David R. Jones and Steve Lawrence all provided helpful input of various sorts as we were working on different parts of the paper, and the Editors and reviewers made valuable suggestions for revisions.

References

Agresti, A. (1994) Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, **50**, 494–500.

Basu, S. (1998) Bayesian estimation of the number of undetected errors when both reviewers and errors are heterogeneous. In *Frontiers in Reliability Analysis* (eds A. P. Basu, S. K. Basu and S. Mukhopadhyay), pp. 19–36. Singapore: World Scientific.

Biggeri, A., Stanghellini, E., Merletti, M. and Marchi, M. (1999) Latent class models for varying catchability and correlation among sources in capture-recapture estimation of the size of a human population. *Statist. Appl.*, to be published.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: Massachusetts Institute of Technology Press.

Bruno, G., Laporte, R., Merletti, E., Biggeri, A., McCarty, D. and Pagano, C. (1994) National diabetes programs: application of capture-recapture to "count" diabetes. *Diab. Care*, **17**, 548–556.

Burnham, K. P. and Overton, W. S. (1978) Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, **65**, 625–633.

Carriquiry, A. and Fienberg, S. E. (1998) *Encyclopedia of Biostatistics*, vol. 5, pp. 3724–3730. New York: Wiley.

Castledine, B. J. (1981) A bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, **67**, 197–210.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.
——— (1989) Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, **45**, 427–438.
Chao, A., Lee, S.-M. and Jeng, S.-L. (1992) Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, **48**, 201–216.
Cormack, R. M. (1966) A test for equal catchability. *Biometrics*, **22**, 330–342.
——— (1989) Log-linear models for capture-recapture. *Biometrics*, **45**, 395–413.
——— (1992) Interval estimates for mark-recapture studies of closed populations. *Biometrics*, **48**, 567–576.
——— (1994) *Statistics in Ecology and Environmental Monitoring*, pp. 19–32. Dunedin: University of Otago Press.
Coull, B. A. and Agresti, A. (1999) The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, **55**, 294–301.
Cressie, N. and Holland, P. W. (1983) Characterizing the manifest probabilities of latent trait models. *Psychometrika*, **48**, 129–141.
Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993) A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J. Am. Statist. Ass.*, **88**, 1137–1148.
Darroch, J. N. and McCloud, P. I. (1990) Separating two sources of dependence in repeated influenza outbreaks. *Biometrika*, **77**, 237–243.
Fienberg, S. E. (1972) Multiple-recapture census for closed populations and incomplete contingency tables. *Biometrika*, **59**, 591–603.
——— (1992) Bibliography on capture-recapture modelling with application to census undercount adjustment. *Surv. Methodol.*, **18**, 143–154.
Fienberg, S. E. and Meyer, M. M. (1983) Loglinear models and categorical data analysis with psychometric and econometric applications. *J. Econometr.*, **22**, 191–214.
Freeman, P. R. (1972) Sequential estimation of the size of a population. *Biometrika*, **59**, 9–18.
——— (1973) A numerical comparison between sequential tagging and sequential recapture. *Biometrika*, **60**, 499–508.
Garthwaite, P. H., Yu, K. and Hope, P. B. (1995) Bayesian analysis of a multiple-recapture model. *Communs Statist. Theory Meth.*, **24**, 2229–2247.
Geiger, H. and Werner, A. (1924) Die Zahl der Ion Radium Ausgesandsena-teilchen. *Z. Phys.*, **21**, 187–201.
George, E. I. and Robert, C. P. (1992) Capture-recapture estimation via gibbs sampling. *Biometrika*, **79**, 677–683.
Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.
Green, P. J. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711–732.
Hay, G. (1997) The selection from multiple data sources in epidemiological capture–recapture studies. *Statistician*, **46**, 515–520.
Holland, P. W. (1990) On the sampling theory foundations of item response theory models. *Psychometrika*, **55**, 577–601.
International Working Group for Disease Monitoring and Forecasting (1995a) Mark-recapture and multiple-record systems: I, History and theoretical development. *Am. J. Epidem.*, **142**, 1047–1058.
——— (1995b) Mark-recapture and multiple-record systems: II, Applications in human diseases. *Am. J. Epidem.*, **142**, 1059–1068.
Jannarone, R. J. (1986) Conjunctive item response theory kernels. *Psychometrika*, **51**, 357–373.
Jannarone, R. J., Yu, K. F. and Laughlin, J. E. (1990) Easy bayes estimation for rasch-type models. *Psychometrika*, **55**, 449–460.
Johnson, M. S., Cohen, W. M. and Junker, B. W. (1998) Measuring appropriability in research and development with item response models. *Technical Report 690*. Department of Statistics, Carnegie Mellon University, Pittsburgh. (Available from `http://www.stat.cmu.edu/www/cmu-stats/tr/`.)
Kass, R. E. and Wasserman, L. (1995) A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *J. Am. Statist. Ass.*, **98**, 928–934.
Kelderman, H. (1984) Loglinear rasch model tests. *Psychometrika*, **49**, 223–245.
Lawrence, S. and Giles, C. L. (1998) Searching the world wide web. *Science*, **280**, 98–100.
Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, 215–232.
——— (1997) Bayesian methods for estimating the size of a closed population. *Biometrika*, **84**, 19–31.
Mathsoft (1996) *S-plus Version 3.4 Release 1 for HP 9000 Series*. Seattle: Mathsoft.
Patz, R. J. and Junker, B. W. (1999) A straightforward approach to markov chain monte carlo methods for item response models. *J. Educ. Behav. Statist.*, **24**, 146–177.
Petersen, C. J. G. (1896) The yearly immigration of young plaice into the limfjord from the german sea. *Rep. Dan. Biol. Stn Min. Fish.*, **6**, 1–48.
Pollock, K. H. (1991) Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *J. Am. Statist. Ass.*, **86**, 225–238.
Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.
Roberts, H. V. (1967) Informative stopping rules and inferences about population size. *J. Am. Statist. Ass.*, **62**, 763–775.

Sanathanan, L. (1972) Models and methods in visual scanning experiments. *Technometrics*, **14**, 813–830.
———(1973) A comparison of some models in visual scanning experiments. *Technometrics*, **15**, 67–78.
Schnabel, Z. (1938) The estimation of the total fish population of a lake. *Am. Math. Mnthly*, **45**, 348–352.
Smith, P. J. (1988) Bayesian methods for multiple capture-recapture surveys. *Biometrics*, **44**, 1177–1189.
———(1991) Bayesian analyses for a multiple capture-recapture model. *Biometrika*, **78**, 399–407.
Wu, M. L., Adams, R. J. and Wilson, M. R. (1997) *Conquest: Generalized Item Response Modeling Software*. Camberwell: Australian Council on Educational Research.