A Note on Broken Sample Problem

By YING NIAN WU

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

yingnian@umich.edu

Summary

A broken sample (Degroot, Feder, and Goel, 1971) is a random sample observed for a two-component random variable (X, Y), but the links (or correspondences) between the X-components and the Y-components are broken (or missing). This note proposes a method for re-pairing the broken sample, as well as making inference about the parameter in the model for (X, Y). We also extend the broken sample formulation to study the situation where the X-components and the Y-components themselves are subject to missing. A potential area of application is file matching or record linkage, where, for the purpose of administration or statistical analysis, two files of records that are contaminated by errors and non-uniqueness need to be linked (or re-paired) such that the linked records are related to identical units.

Some key words: Broken sample; Capture-recapture; File matching; Metropolis-Hastings algorithm; MCMC; Re-pairing; Record linkage; Shuffling.

1. Introduction

The broken sample problem was first studied by Degroot, Feder, and Goel (1971). As a simple example, they considered pairing the photographs of N movie stars with N photographs of the same stars taken when they were babies, on the basis of r different facial measurements on the photograph of every star and s facial measurements on every baby photograph. In their formulation, the facial measurements on the star photograph and those on the baby photograph of a randomly selected star are modeled by a two-component random variable (X, Y), following a joint distribution $P(x, y \mid \theta)$, where X is for the star photograph, Y is for the baby photograph, and θ is the parameter. Then the problem is as follows: a random sample of size N is collected for (X, Y), but before the sample is observed, the links (or correspondences) between the X- and Y-components are broken (or missing), what we observe is a broken sample with $B_x = \{x_1, ..., x_N\}$ and $B_y = \{y_1, ..., y_N\}$, without any prior information concerning which x_i in B_x is originally linked to which y_j in B_y. The goal is to make inference about the original links based on the model $P(x, y \mid \theta)$, and also to make inference about θ if it is unknown.

As mentioned in Degroot, et al. (1971), a more realistic version of this problem is file matching or record linkage (e.g., Fellegi and Sunter, 1969), where the goal is to pair or link records in two files, such that the linked records represent identical units. Because a unique identifier such as social security number is usually not available, and the information in the records is often contaminated by various sources of errors, such as reporting errors and entry errors, statistical modeling and methods are needed for file matching. The purposes of file matching include: 1) Administration, see, e.g., Copas and Hilton (1990), for an immigration example. 2) Data analysis, where the variables concerned are recorded in two different files. See Winkler (1991) for a discussion of the analysis of computer linked files. 3) Estimation of the population size, where the two files can be viewed as two captures, and the matching is needed to obtain the number of recaptures. See, e.g., Hogan (1992), and Breiman (1994), for an example concerning the estimation of the population undercounts by the U.S. Bureau of Census.

Section 2 studies the broken sample problem of Degroot, et al. (1971), and proposes a shuffling algorithm for computation. Section 3 extends the broken sample formulation to address the more realistic situation of file matching, where each file covers only part of the population. Finally, Section 4 concludes with some remarks.

2. The Broken Sample Problem

2.1 Inference of unknown match and the shuffling algorithm

In the broken sample formulation of Degroot, et al. (1971), a mathematical representation of the unknown links is a one-to-one mapping ϕ from domain $\text{dom}(\phi) = \{1, ..., N\}$ onto image $\text{im}(\phi) = \{1, ..., N\}$, such that x_i is originally linked to $y_{\phi(i)}$ for $i \in \{1, ..., N\}$. We call ϕ a "match". Another representation is an $N \times N$ permutation matrix M, where $M_{ij} = 1$ if x_i is linked to y_j, and $M_{ij} = 0$ otherwise.

Given ϕ and θ, the conditional distribution of the observed broken sample is

$$P(B_x, B_y \mid \phi, \theta) = \prod_{i=1}^{N} P(x_i, y_{\phi(i)} \mid \theta).$$

Under the permutation matrix representation, it becomes

$$P(B_x, B_y \mid M, \theta) = \prod_{i=1}^{N} \prod_{j=1}^{N} P(x_i, y_j \mid \theta)^{M_{ij}}.$$

Let's first study the inference of ϕ with known θ, i.e., re-pairing the broken sample. This can be accomplished using a Bayesian method, with a uniform prior distribution over all possible matches, i.e., $P(\phi) = 1/N!$. This is the only reasonable prior distribution for quantifying our ignorance of the missing correspondence. It is also justified by the fact that the only way to produce a broken sample or to make the links missing is to randomly permute the indices of the components of the original unbroken sample. In this sense, it is more appropriate to consider ϕ as missing data instead of an unknown parameter, and treat the uniform $P(\phi)$ as part of the model. The posterior distribution of ϕ (given θ) is

$$\pi(\phi) = P(\phi \mid B_x, B_y, \theta) \propto \prod_{i=1}^{N} P(x_i, y_{\phi(i)} \mid \theta).$$

$\pi(\phi)$ can be used to answer questions like what is the best re-pairing of the broken sample, how likely a pair of components arise from the same unit, what is the most likely Y-component that is originally linked to an X-component, etc. $\pi(\phi)$ can be summarized by median and posterior region if we impose a distance d on the space of matches. The median is given by

$$\phi_{\text{med}} = \arg\min_{\phi} \mathrm{E}_{\pi(\phi')}[d(\phi, \phi')] = \arg\min_{\phi} \sum_{\phi'} d(\phi, \phi')\pi(\phi'),$$

and the $1-\alpha$ posterior region can be summarized by a set $\{\phi : d(\phi_{\text{med}}, \phi) \le d_\alpha\}$ for some d_α, such that the posterior probability for this set is no less than $1-\alpha$. A typical choice of $d(\phi, \phi')$ is the number of mismatches between ϕ and ϕ', i.e., $d(\phi, \phi') = \sum_i \delta_i$, where $\delta_i = 1$ if $\phi(i) \neq \phi'(i)$, and $\delta_i = 0$ otherwise. Under this distance, ϕ_{med} is regarded as the optimal estimate of the true match with minimum expected number of mismatches under the posterior distribution, and d_α provides uncertainty about ϕ_{med} in terms of mismatches. Degroot, et al. (1971) propose to estimate ϕ using maximum likelihood, which is the Bayesian posterior mode, or posterior median with 0-1 distance. However, MLE does not address the uncertainty of the estimated match as well as related questions such as those mentioned above.

An interesting property of this Bayesian procedure is that the posterior statements for ϕ are calibrated with the frequency sampling statements. To be more specific, let $R(B_x, B_y)$ be the $1-\alpha$ posterior region constructed above, such that $P(\phi \in R(B_x, B_y) \mid B_x, B_y) \ge 1-\alpha$, then marginalizing over (B_x, B_y) gives $P(\phi \in R(B_x, B_y)) \ge 1-\alpha$. Meanwhile, the sampling coverage probability $P(\phi \in R(B_x, B_y) \mid \phi)$ is constant across all ϕ, because different ϕ only give different permutations of indices, and the posterior inference with the uniform prior distribution is invariant of the permutation of the indices. This pivotal argument leads to $P(\phi \in R(B_x, B_y) \mid \phi) \ge 1-\alpha$, i.e., the posterior region is also the confidence region.

Handling $\pi(\phi)$ analytically requires the ability to compute the normalizing constant like $\sum_\phi \prod_{i=1}^N P(x_i, y_{\phi(i)} \mid \theta)$, which is the permanent of the matrix $[P(x_i, y_j \mid \theta)]_{N\times N}$. As proved by Valiant (1979), the computation of permanents is NP-complete, so we have to resort to Markov chain Monte Carlo (MCMC) for computation. The following is a Metropolis-Hastings type of algorithm, which we call the shuffling algorithm, for simulating $\pi(\phi)$.

Shuffling algorithm Start from an arbitrary match. Let ϕ_1 be the current match. Randomly pick $i \in \{1, ..., N\}$ and $j \in \{1, ..., N\}$, then there are only two different cases.

1) If $j = \phi_1(i)$, stay unchanged.

2) If $j \neq \phi_1(i)$, propose a move from ϕ_1 to ϕ_2, where ϕ_2 is different from ϕ_1 only in that $\phi_2(i) = j$ and $\phi_2(\phi_1^{-1}(j)) = \phi_1(i)$. Accept this move with probability $P_{\phi_1,\phi_2} = \min(1, \pi(\phi_2)/\pi(\phi_1))$, where

$$\frac{\pi(\phi_2)}{\pi(\phi_1)} = \frac{P(x_i, y_j \mid \theta) P(x_{\phi_1^{-1}(j)}, y_{\phi_1(i)} \mid \theta)}{P(x_i, y_{\phi(i)} \mid \theta) P(x_{\phi_1^{-1}(j)}, y_j \mid \theta)}.$$

If the move is accepted, change ϕ_1 into ϕ_2, i.e., the new ϕ_1 after the update is ϕ_2; otherwise, stay at ϕ_1. Repeat this process until convergence.

We call this transition *swapping*, see Picture 1 for an illustration. Wu (1998) proves that the algorithm converges in polynomial time. With the ability to simulate $\pi(\phi)$, the posterior median and posterior probability region can be approximated.

2.2 Inference of unknown parameter

If θ is unknown and is of interest, Degroot and Goel (1980) propose to estimate θ by maximizing the integrated likelihood,

$$\sum_\phi P(B_x, B_y \mid \phi, \theta) = \sum_\phi \prod_{i=1}^N P(x_i, y_{\phi(i)} \mid \theta),$$

where ϕ is treated as missing data and is integrated out. Theoretically, this integrated likelihood can be maximized by the EM algorithm (Dempster, Laird, and Rubin, 1977). Let $\theta^{(t)}$ be the estimate at the tth iteration, then at the $(t+1)$st iteration, the E-step computes the conditional expectation of the permutation matrix $\hat{M} = E[M \mid B_x, B_y, \theta^{(t)}]$, and the M-step find $\theta^{(t+1)}$ by maximizing the complete-data log-likelihood

$$\sum_{i=1}^N \sum_{j=1}^N \hat{M}_{ij} \log P(x_i, y_j \mid \theta).$$

Because $\hat{M}$ cannot be computed analytically, the Monte Carlo EM of Wei and Tannar (1991) has to be used, where $\hat{M}$ is approximated by sampling from $P(\phi \mid B_x, B_y, \theta^{(t)})$ using the shuffling algorithm. See Chan and Ledholter (1995) for a discussion of embedding MCMC into Monte Carlo EM.

A Bayesian approach is to put a prior distribution $P(\theta)$ on θ, and sample θ and ϕ jointly from the posterior distribution

$$P(\phi, \theta \mid B_x, B_y) \propto P(\theta) \prod_{i=1}^{N} P(x_i, y_{\phi(i)} \mid \theta).$$

using the data augmentation algorithm (Tannar and Wong, 1987), which iterates the following two steps. 1) Shuffling-step: Given the current value of θ, update ϕ by a number (e.g., N^2) of swapping transitions, starting from the ϕ generated by the previous iteration. 2) Fitting-step: Given the current match ϕ, draw θ from its conditional posterior distribution by fitting the model $P(x, y \mid \theta)$ under ϕ.

2.3 A numerical example

Degroot and Goel (1980) studied the inference of the correlation coefficient ρ when the model $P(x, y \mid \theta)$ is a bivariate normal distribution $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and they derived many interesting theoretical results. In their paper, numerical calculation was done only for small sample size such as $N = 5$. Here, we shall illustrate our method by simulating the posterior distribution of ρ with a much larger sample size, for example, we take $N = 100$. We use the non-informative prior distribution $P(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-3/2}$. The fitting-step can be accomplished by first drawing $\boldsymbol{\Sigma}$ from an inverse-Wishart distribution, then drawing $\boldsymbol{\mu}$ given $\boldsymbol{\Sigma}$ from a bivariate normal distribution, and hence ρ is computed.

We did three experiments with the observed broken sample simulated from a bivariate normal distribution with $\mu_X = \mu_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, and $\rho = 0, 0.9$, and -0.9 respectively. Figure 1 shows the posterior histograms of ρ for the three simulated data sets. For large N, $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ can be estimated accurately, and the posterior distributions of ρ exhibit the properties derived by Degroot and Goel (1980) for the integrated likelihoods of ρ with $\mu_X, \mu_X, \sigma_X^2, \sigma_Y^2$ fixed at MLE, e.g., $\rho = 0$ is a local minimum, and there is at least one local maximum in each of $(-1, 0)$ and $(0, 1)$.

3. Extended Broken Sample Formulation and Algorithms

3.1 Extended broken sample formulation

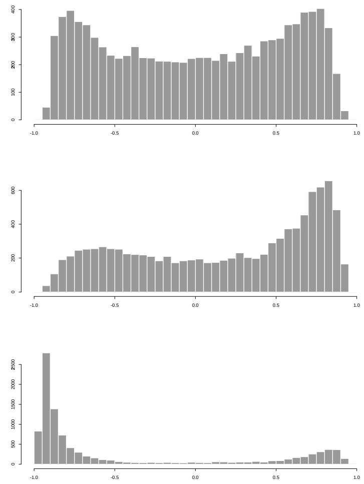


Figure 1: Empirical posterior distributions of ρ approximated by 10,000 draws for the three simulated data sets with $\rho = 0$, .9, and -.9 respectively. The top figure is for $\rho = 0$, the middle one is for $\rho = .9$, and the bottom one is for $\rho = -.9$.

In file matching, it is almost always the case that each of the two files only covers part of the underlying population, especially in the capture-recapture estimation of the population size. Therefore, the model should include not only the mechanism for generating the record values, but also the mechanism for including the records in the files. So we introduce two inclusion indicators, U and V, where U is the indicator for X — if $U = 1$ then X is observed (or included in the file), otherwise X is missing; V is the indicator for Y with similar interpretation. With these two indicators, the model becomes $P(x, y; u, v \mid \theta)$ with parameter θ. This model can be decomposed into the record generating mechanism and the record inclusion mechanism in the following two ways

$$\begin{aligned} P(x, y; u, v \mid \theta) &= P_{uv}(u, v \mid \theta)\, P_{xy|uv}(x, y \mid u, v, \theta) && (1) \\ &= P_{xy}(x, y \mid \theta)\, P_{uv|xy}(u, v \mid x, y, \theta). && (2) \end{aligned}$$

The notation P_{xy}, $P_{uv|xy}$, P_{uv}, and $P_{xy|uv}$ follows standard interpretations. (2) is more directly relevant than (1) in the survey context, where the probability for a unit to be included in the survey can depend on the background variables. If it is reasonable to assume ignorable missing (Rubin, 1976) or ignorable inclusion, i.e., the inclusion of the records does not depend on the values of the records, then $P(x, y; u, v \mid \theta) = P_{xy}(x, y \mid \theta) P_{uv}(u, v \mid \theta)$. Furthermore, in the traditional capture-recapture context, U and V are assumed to follow independent Bernoulli distributions with unknown probabilities.

In the extended broken sample formulation, the data is generated via the following three steps. 1) A random sample of size N is drawn for the four-component random variable $(X,Y;U,V) \sim P(x,y;u,v \mid \theta)$. 2) If $U=0$, then X is discarded, and if $V=0$, then Y is discarded. 3) The remaining X-components are then randomly shuffled to produce $B_x=(x_1,...,x_{|B_x|})$, and the remaining Y-components are randomly shuffled to produce $B_y=(y_1,...,y_{|B_y|})$, with $|B_x|\le N$ and $|B_y|\le N$ being the sizes of B_x and B_y respectively. We call those x_i the X-records, y_i the Y-records, and B_x and B_y files. The goal is to pair the X-records in B_x to the Y-records in B_y, and to draw inference about θ and N if they are unknown.

The unknown match can still be represented by a mapping ϕ from $\{1,...,|B_x|\}$ into $\{1,...,|B_y|\}$, such that $\mathrm{dom}(\phi)\subset\{1,...,|B_x|\}$, $\mathrm{im}(\phi)\subset\{1,...,|B_y|\}$, and if $i\in\mathrm{dom}(\phi)$, then x_i is originally linked to $y_{\phi(i)}$; if $i\notin\mathrm{dom}(\phi)$, then the Y-record that is originally linked to x_i is missing in B_y; and if $j\notin\mathrm{im}(\phi)$, then the X-record originally linked to y_j is missing in B_x. We use $|\phi|$ to denote the number of matched pairs in ϕ, i.e., the size of $\mathrm{dom}(\phi)$ or $\mathrm{im}(\phi)$.

The distribution of the observed files and the unknown match can be derived as follows. Given N and θ,

$$\begin{aligned} P(|B_x|,|B_y|,|\phi| \mid N,\theta) &= \frac{N!}{|\phi|!(|B_x|-|\phi|)!(|B_y|-|\phi|)!(N-|B_x|-|B_y|+|\phi|)!} \\ &\times P_{uv}(1,1\mid\theta)^{|\phi|}P_{uv}(1,0\mid\theta)^{|B_x|-|\phi|} \\ &\times P_{uv}(0,1\mid\theta)^{|B_y|-|\phi|}P_{uv}(0,0\mid\theta)^{N-|B_x|-|B_y|+|\phi|}, \end{aligned} \quad (3)$$

which is a multinomial distribution on the 2 × 2 table classified by U and V.

Given $(|B_x|,|B_y|,|\phi|,N,\theta)$, ϕ follows a uniform distribution

$$P(\phi \mid |B_x|,|B_y|,|\phi|,N,\theta)=\left[\binom{|B_x|}{|\phi|}\binom{|B_y|}{|\phi|}|\phi|!\right]^{-1}. \quad (4)$$

Given $(|B_x|,|B_y|,\phi,N,\theta)$, the conditional distribution of the observed record values is

$$\begin{aligned} P(B_x,B_y \mid |B_x|,|B_y|,\phi,N,\theta) &= \prod_{i\in\mathrm{dom}(\phi)} P_{xy|uv}(x_i,y_{\phi(i)}\mid 1,1,\theta) \\ & \prod_{i\notin\mathrm{dom}(\phi)} P_{x|uv}(x_i\mid 1,0,\theta)\prod_{j\notin\mathrm{im}(\phi)} P_{y|uv}(y_j|0,1,\theta). \end{aligned} \quad (5)$$

Combining (3), (4), and (5), we have

$$P(B_x,B_y,\phi\mid\theta,N) = \frac{N!}{(N-|B_x|-|B_y|+|\phi|)!}\prod_{i\in\mathrm{dom}(\phi)}P(x_i,y_{\phi(i)};1,1\mid\theta)$$

$$\begin{aligned} &\times \prod_{i \notin \mathrm{dom}(\phi)} P_{xuv}(x_i; 1, 0 \mid \theta) \prod_{j \notin \mathrm{im}(\phi)} P_{yuv}(y_j; 0, 1 \mid \theta) \\ &\times \; P_{uv}(0, 0 \mid \theta)^{N - |B_x| - |B_y| + |\phi|}, \end{aligned}$$

with $|B_x| \leq N$, $|B_y| \leq N$, and $|B_x| + |B_y| - |\phi| \leq N$.

The model $P(B_x, B_y, \phi \mid \theta, N)$ provides a basis for statistical inference, where θ and N are unknown parameters and ϕ the missing data. If the purpose is record linkage, then ϕ is of major concern. If the purpose is data analysis, then θ is of main interest. If the purpose is the capture-recapture estimation of the population size, then N is what we care about (see Smith, 1991, and the references therein for the capture-recapture model). Because of the three connected sources of uncertainties in ϕ, θ, and N, the most straightforward method for statistical inference is to put prior distributions on θ and N, and then use MCMC to simulated the joint posterior distribution of (ϕ, θ, N).

3.2 Extension of algorithms

Let's first consider the situation where N and θ are known, and let $\pi(\phi)$ be the conditional distribution of ϕ given B_x, B_y, θ, and N, which is proportional to $P(B_x, B_y, \phi \mid \theta, N)$. Then the record linkage can be based on $\pi(\phi)$, which defines a monomer-dimer system (Heilmann and Lieb, 1972), where a linked pair of records is a dimer, and an unlinked record is a monomer. $\pi(\phi)$ can be simulated by the following extended version of the shuffling algorithm, each iteration of which is described as follows.

The extended shuffling algorithm: Let ϕ_1 be the current simulated match. Randomly pick $i \in \{1, ..., |B_x|\}$ and $j \in \{1, ..., |B_y|\}$, then there are five different cases.

1) If $i \in \mathrm{dom}(\phi_1)$, $j \in \mathrm{im}(\phi_1)$, and $j \neq \phi_1(i)$, then swap.

2) If $i \in \mathrm{dom}(\phi_1)$ and $j \notin \mathrm{im}(\phi_1)$, then propose a move from ϕ_1 to ϕ_2, where ϕ_2 is different from ϕ_1 only in that $\mathrm{im}(\phi_2) = \mathrm{im}(\phi_1) - \{\phi_1(i)\} + \{j\}$, and $\phi_2(i) = j$. Accept this move with probability $P_{\phi_1,\phi_2} = \min(1, \pi(\phi_2)/\pi(\phi_1))$. We call this transition as *X-switching.* See Picture 2.

3) If $i \notin \mathrm{dom}(\phi_1)$ and $j \in \mathrm{im}(\phi_1)$, then propose a move from ϕ_1 to ϕ_2, where ϕ_2 is different from ϕ_1 only in that $\mathrm{dom}(\phi_2) = \mathrm{dom}(\phi_1) - \{\phi_1^{-1}(j)\} + \{i\}$ and $\phi_2(i) = j$. Accept this move with probability $P_{\phi_1,\phi_2} = \min(1, \pi(\phi_2)/\pi(\phi_1))$. We call this transition as *Y-switching.* See Picture 3.

4) If $i \notin \mathrm{dom}(\phi_1)$, and $j \notin \mathrm{im}(\phi_1)$, propose a move from ϕ_1 to ϕ_2, where ϕ_2 is different from ϕ_1 only in that $\mathrm{dom}(\phi_2) = \mathrm{dom}(\phi_1) + \{i\}$, $\mathrm{im}(\phi_2) = \mathrm{im}(\phi_1) + \{j\}$, and $\phi_2(i) = j$. Accept

this move with probability $P_{\phi_1,\phi_2} = \min(1, \pi(\phi_2)/\pi(\phi_1))$. We call this transition as *linking*. See Picture 4.

5) If $i \in \mathrm{dom}(\phi_1)$, $j \in \mathrm{im}(\phi_1)$ and $j = \phi_1(i)$, and $N - |B_x| - |B_y| + |\phi| \geq 1$, propose a move from ϕ_1 to ϕ_2, where ϕ_2 is different from ϕ_1 only in that $\mathrm{dom}(\phi_2) = \mathrm{dom}(\phi_1) - \{i\}$ and $\mathrm{im}(\phi_2) = \mathrm{im}(\phi_1) - \{j\}$. Accept this move with probability $P_{\phi_1,\phi_2} = \min(1, \pi(\phi_2)/\pi(\phi_1))$. We call this transition as *unlinking*. See Picture 5.

The algorithm can be summarized as follows. 1) Randomly pick up a pair of records. 2) Flip the linkage status of this pair, i.e., if they are linked under the current match, unlink them, otherwise, link them. 3) Make appropriate arrangement if the two records are previously linked to any other records to make the resulting configuration a legitimate match. 4) Accept this move with an appropriate probability. The calculation of P_{ϕ_1,ϕ_2} only involves the records being picked up, as well as the records they may be currently linked to, and hence is very simple.

If θ and N are unknown, which is usually the case, the joint posterior distribution of (θ, N, ϕ) can be simulated by an extended version of the shuffling-fitting algorithm, each iteration of which consists of three steps:

1) Given θ and N, shuffling ϕ a number (e.g., $|B_x| \times |B_y|$) of times.

2) Given ϕ and N, draw θ from its conditional posterior distribution.

3) Given ϕ and θ, simulate N by running the Metropolis algorithm a number of steps, where the proposed move is +1 or -1 with equal probabilities.

See Wu (1998) for a numerical example.

4. Concluding Remarks

The broken sample formulation studied in this note has the advantage that the three closely connected themes, record linkage, data analysis, and estimation of the population size are treated together. In record linkage, the overall match is estimated, therefore, the dependence among the linkage status of individual pairs is taken into account. Like Degroot, et al. (1971), this note is only at a theoretical level, and future work is needed to apply the method to real file matching problems.

Acknowledgment

The author thanks Prof. Persi Diaconis, Prof. Jun Liu, Prof. Donald Rubin, and Prof. Alan Zaslavsky for helpful discussions.

Reference

Breiman, L. (1994). The 1991 census adjustment: undercount or bad data? *Statistical Science*, **9**, 458-537.

Chan, K. S. and Ledholter, L. (1995) Monte Carlo EM estimation for time series models involving counts. *JASA*, **90**, 242-252.

Copas, J. B. and Hilton, F. J. (1990). Record linkage: statistical models for matching computer records (with comments). *JRSS-A*, **153**, 3, 320.

Degroot, M. H., Feder, P. I., and Goel, P. K. (1971). Matchmaking. *Ann. Math. Stat.*, **42**, 578-593.

Degroot, M. H. and Goel, P. K. (1980). Estimation of the correlation coefficient from a broken random sample. *Ann. Stat.*, **8**, 264-278.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *JRSS-B*, **39**, 1-38.

Diaconis, P. and Shahshahani, M. (1987). Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete*, **57**, 159-180.

Fellegi, I. P. and Sunter, A. B. (1969). A theory of record linkage. *JASA*, **40**, 1183-1210.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Heilmann, O. J. and Lieb, E. H. (1972). Theory of monomer-dimer systems. *Comm. Math. Physics.*, **25**, 190-232.

Hogan, H. (1992). The 1990 Post Enumeration Survey: an overview. *American Statistician,* **40**, 261-269.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-92.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63** , 581-590.

Smith, P. (1991). Bayesian analysis for a multiple capture-recapture model. *Biometrika*, **78**, 399-407.

Tanner, M. A. (1993) *Tools for Statistical Inference*, Springer-Verlag, New York.

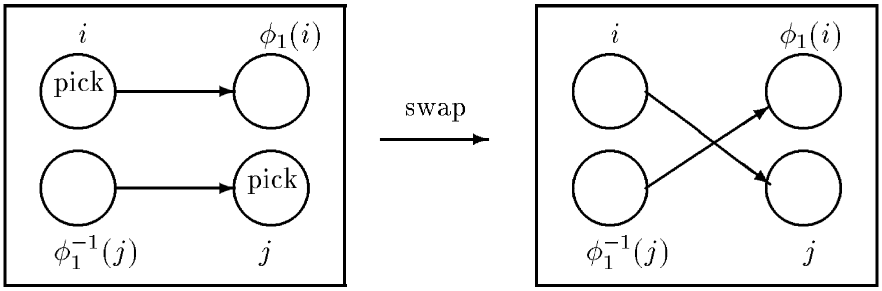
Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distribution by data augmentation (with discussion), *JASA*, **82**, 528-550.

Wei, G. C. G. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *JASA*, **85**, 699-704.

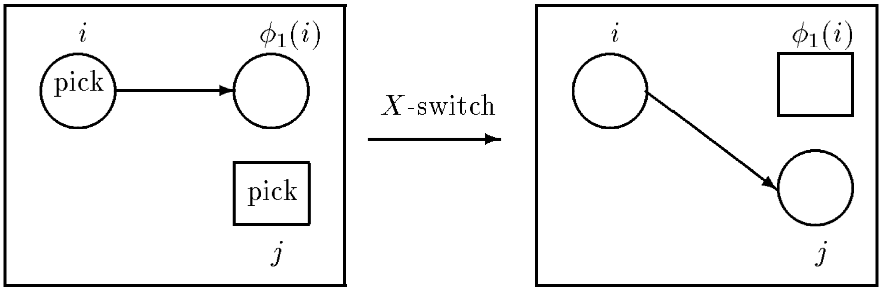
Winkler, W. E. (1991). Error model for analysis of computer linked files, *ASA Proc. of Survey Rsch. Methods Sect.*, 472-477.

Wu, Y. (1998) A note on broken sample. Technical Report. Dept. of Statistics, Univ. of Michigan.

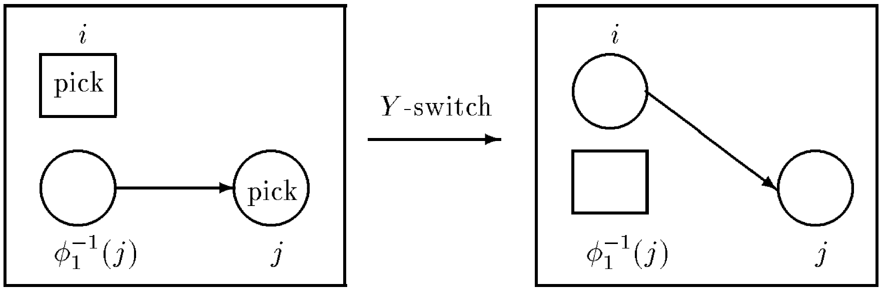
Picture 1: Swap.



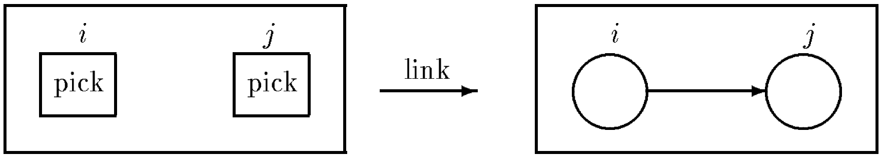
Picture 2: X-switch.



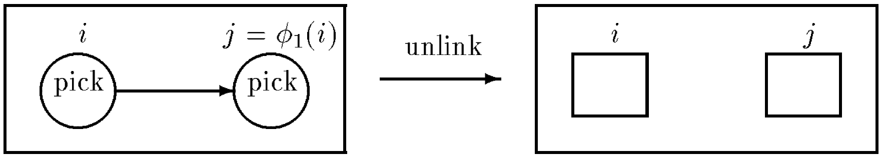
Picture 3: Y-switch.



Picture 4: Link.



Picture 5: Unlink.


i
$j = \phi_1(i)$
pick
pick
unlink
i
j