eco: R Package for Ecological Inference in 2×2 Tables*

Kosuke Imai† Ying Lu‡ Aaron Strauss§

Version 3.1-0
June 27, 2007

Abstract

eco is a publicly available R package that implements the Bayesian and likelihood methods proposed in Imai, Lu, and Strauss (2008b) for ecological inference in 2×2 tables as well as the method of bounds introduced by (Duncan and Davis, 1953). The package fits both parametric and nonparametric models using either the Expectation-Maximization algorithms (for likelihood models) or the Markov chain Monte Carlo algorithms (for Bayesian models). For all models, the individual-level data can be directly incorporated into the estimation whenever such data are available. Along with in-sample and out-of-sample predictions, the package also provides a functionality which allows one to quantify the effect of data aggregation on parameter estimation and hypothesis testing under the parametric likelihood models. This paper illustrates the usage of eco with several real data examples that are also part of the package.

1 Introduction

This paper illustrates how to use eco, a publicly available R package (R Development Core Team, 2007), to implement the Bayesian and likelihood methods proposed in Imai, Lu, and Strauss (2008b) for ecological inference in 2×2 tables. The package also implements the method of bounds introduced by (Duncan and Davis, 1953) for the analysis of general $R \times C$ tables. Ecological inference refers to the "inferences about individual behavior drawn from data about aggregates" (Freedman, 1999, p.4027). Such cross-level inferences are frequently conducted in epidemiology, political science, and sociology when only aggregate-level data are available (e.g.,

*This paper is the up-to-date version of Imai, Lu, and Strauss (In-press). Financial support from the National Science Foundation (SES–0550873) and Princeton University Committee on Research in the Humanities and Social Sciences is acknowledged.

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6610, Fax: 973–556–1929, Email: kimai@Princeton.Edu, URL: http://imai.princeton.edu/

‡Assistant Professor, Department of Sociology, University of Colorado at Boulder. Phone: 303–492–7030, Email:ying.lu@colorado.edu

§Ph.D. student, Department of Politics, Princeton University

Greenland and Robins, 1994; Achen and Shively, 1995; King, 1997; King *et al.*, 2004). Yet, the difficulty of ecological inference is that the observed correlation at the aggregate level does not necessarily imply the same individual-level relationship. Using an example of literacy rates across different racial groups, Robinson (1950) powerfully illustrated this "ecological fallacy."

Since Robinson's seminal article, various methods have been proposed for ecological inference. Duncan and Davis (1953) showed how to derive the bounds on unknown quantities of interest from aggregate data. We generalize and implement this method for $R \times C$ tables in `eco`. Goodman (1953, 1959) developed the regression-based approach to ecological inference, which gained popularity among applied researchers in the next several decades (e.g., Freedman *et al.*, 1991; Achen and Shively, 1995; Gelman *et al.*, 2001) – this approach can be easily implemented via `lm()` command in R, and hence is not implemented in `eco`. Recent years have witnessed a growing number of new methods based on modern statistical techniques (e.g., King *et al.*, 1999; Rosen *et al.*, 2001; Imai and King, 2004; Judge *et al.*, 2004; Wakefield, 2004) – some of these methods are available in R via `Zelig` (Imai, King, and Lau, 2008a) and `MCMCpack` (Martin and Quinn, 2006). At the same time, the appropriateness of the assumptions underlying some of these models is often disputed (e.g., Freedman *et al.*, 1998; Cho, 1998; King, 1999; Cho and Gaines, 2004).

In a recent paper, Imai, Lu, and Strauss (2008b) have proposed a theoretical framework for Bayesian and likelihood inference in 2×2 ecological tables. The framework is based on the theory of coarse data which is originally developed by Heitjan and Rubin (1991). We show that ecological inference can be formulated as a coarse data problem and that Bayesian and likelihood inference can be conducted within this coarse data framework. The main advantage of this framework is that it clarifies the modeling assumptions necessary for Bayesian and likelihood ecological inference. In particular, Imai, Lu, and Strauss (2008b) show that the ecological inference problem can be decomposed into three key factors: *distributional effects* which address the possible misspecification of parametric modeling assumptions about the unknown distribution of missing data, *contextual effects* which represent the possible correlation between missing data and observed variables, and *aggregation effects* which are directly related to the loss of information caused by data aggregation.

Furthermore, while Imai, Lu, and Strauss (2008b) propose statistical methods to address distributional and contextual effects, they also show that aggregation effects cannot be overcome by statistical adjustments. Instead, they demonstrate how to formally quantify the effect of data aggregation on parameter estimation and hypothesis testing. In this paper, we illustrate how to implement these proposed methods using an R package `eco`, which is freely available at the Comprehensive R Archive Network (`http://cran.r-project.org/`). In Section 2, we start our discussion by describing and generalizing the method of bounds (Duncan and Davis, 1953). We then outline the parametric and nonparametric models proposed by Imai, Lu, and Strauss (2008b), which respect the constraints imposed by the bounds. We also briefly review the method to formally quantify the aggregation effects. In Section 3, we illustrate the use of the `eco` package through the analysis of several real data examples. Finally, the appendices at the end of the paper provides the detailed references of commands and data sets that are a part of the `eco` package.

	black voters	white voters	
Voted	W_{i1}	W_{i2}	Y_i
Not Voted	$1 - W_{i1}$	$1 - W_{i2}$	$1 - Y_i$
	X_i	$1 - X_i$	

Table 1: 2×2 Ecological Table for the Racial Voting Example. X_i, Y_i, W_{i1}, and W_{i2} are proportions, and hence lie between 0 and 1. The unit of observation is typically a geographical unit and is denoted by i.

2 The Methodology

We use racial voting as a concrete example to describe ecological inference in 2×2 tables. Although this is a prominent example in political science, other problems in different disciplines may fit into the same framework. Table 1 presents a 2×2 ecological table of racial voting example. Suppose that from the census data we observe the fraction of registered white and black voters for each county, i.e., X_i and $1 - X_i$. The overall turnout rate Y_i can be obtained from the election returns for each county. However, the proportions of black and white voters who turned out, W_{i1} and W_{i2} respectively, are unknown.

The `eco` package implements the method of bounds and fits both parametric and nonparametric methods for such ecological data. The estimation is based on the Expectation-Maximization (EM) algorithms (Dempster, Laird, and Rubin, 1977) for likelihood models and on the Markov chain Monte Carlo (MCMC) algorithms for Bayesian models. These algorithms are described in Imai, Lu, and Strauss (2008b). Below, we briefly summarize each method and model. Note that although we do not discuss the issue of convergence of Markov chains in detail, users of **eco** should follow standard advice and conduct convergence diagnostics, perhaps using the **coda** package.

2.1 The Method of Bounds

Suppose that in a simple random sample of size n from a population, we observe the margins of Table 1 for each county i. The method of bounds is based on the following deterministic relationship,

$$Y_i \;=\; W_{i1}X_i + W_{i2}(1 - X_i), \quad \text{for} \quad i = 1, 2, \ldots, n \tag{1}$$

where $X_i, Y_i, W_{i1}, W_{i2} \in [0, 1]$. When Y_i is equal to either 0 or 1, W_{i1} and W_{i2} are completely known. If $X_i = 1$, then $W_{i1} = Y_i$ but W_{i2} does not exist. Similarly, if $X_i = 0$, then $W_{i2} = Y_i$ but W_{i1} does not exist. King (1997) called equation 1 a tomography line. For every i, this tomography line defines a *deterministic* relationship between the missing data, $W_i = (W_{i1}, W_{i2})$ and the observed data, (Y_i, X_i). Duncan and Davis (1953) first recognized that with equation 1, one can narrow the original bound of $[0, 1]$ for W_i to the following intervals,

$$W_{i1} \;\in\; \left[\max\left(0, \frac{X_i + Y_i - 1}{X_i}\right),\ \min\left(1, \frac{Y_i}{X_i}\right)\right], \tag{2}$$

$$W_{i2} \;\in\; \left[\max\left(0, \frac{Y_i - X_i}{1 - X_i}\right),\ \min\left(1, \frac{Y_i}{1 - X_i}\right)\right]. \tag{3}$$

Given these bounds for each i (e.g., a county), the analysis of larger units (e.g, a state) can be carried out by simply aggregating the upper and lower bounds with appropriate weights; $N_i X_i$ and $N_i(1 - X_i)$ for W_{i1} and W_{i2}, respectively, where N_i is the total number of voters in county i. When the resulting bounds are sufficiently narrow, researchers can draw reasonably informative conclusions about (in-sample) missing cells.

The bounds in equations 2 and 3 can be easily generalized to the situation of $R \times C$ ecological tables where $R \geq 2$ and $C \geq 2$. These generalized bounds can also be computed via the `eco` package. Suppose that we denote the observed row and column margins by Y_{ir} and X_{ic} for $c = 1, \ldots, C$, and $r = 1, \ldots, R$ where $\sum_{c=1}^{C} X_i = 1$ and $\sum_{r=1}^{R} Y_{ir} = 1$ for all i. Then, the unobserved proportion in the rth row and cth column can be defined as W_{irc}. The results in the statistical literature on contingency tables (Bonferroni, 1936; Fréchet, 1940; Hoeffding, 1940) imply that the bounds are given by,

$$\max\left\{0,\ \frac{X_{ic} + Y_{ir} - 1}{X_{ic}}\right\} \ \leq\ W_{irc}\ \leq\ \min\left\{1,\ \frac{Y_{ir}}{X_{ic}}\right\}. \tag{4}$$

Although applied researchers often find the bounds too wide for their purposes, the method of bounds shows the identifying power of the data without any statistical assumption. That is, the bounds imply the exact degree to which the data are informative about W_i. For this reason, statistical analysis that does not incorporate this deterministic relationship is likely to be sensitive to modeling assumptions. The `eco` package computes the bounds for the general $R \times C$ case as well as for the 2×2 case (see Section 3.1).

2.2 Parametric Models

Next, we describe the parametric models implemented by the package `eco`. Imai, Lu, and Strauss (2008b) propose the parametric models based on three assumptions. The simplest parametric model is based on the assumption of *coarsened at random* (CAR) and defined by,

$$W_i^* \mid \mu, \Sigma \ \overset{\text{i.i.d.}}{\sim}\ \mathcal{N}(\mu, \Sigma),$$

where $W_i^* = (\text{logit}(W_{i1}), \text{logit}(W_{i2}))$, μ represents a 2×1 vector of population means, and Σ is a 2×2 positive-definite variance matrix. The model, which is similar to the ones proposed by King (1997) and Wakefield (2004), assumes the independence between W_i and X_i and thus no contextual effect. The maximum likelihood (ML) estimates of μ and Σ can be computed via the EM algorithm. The Bayesian analysis, on the other hand, is based on the following conjugate prior distribution,

$$\mu \mid \Sigma \ \sim\ \mathcal{N}(\mu_0, \Sigma/\tau_0^2), \quad \text{and} \quad \Sigma \ \sim\ \text{InvWish}\,(\nu_0,\ S_0^{-1}),$$

where μ_0 denotes a (2×1) vector of the prior mean, τ_0 is a scalar, ν_0 is the prior degrees of freedom parameter, and S_0 represents a (2×2) positive definite prior scale matrix. The posterior inference can then be conducted by the MCMC algorithm.

The assumption of no contextual effect under the CAR model is unrealistic in many situations. Imai, Lu, and Strauss (2008b) consider two modeling strategies in order to relax this assumption.

First, one may collect additional covariates Z_i and assume no contextual effect after conditioning on Z_i. Such a strategy is often employed in the literature (e.g., King, 1997; King *et al.*, 1999). Thus, we can extend our CAR model to the following CCAR (*conditionally coarsened at random*) model,

$$W_i^* \mid \beta, \Sigma, Z_i \overset{\text{indep.}}{\sim} \mathcal{N}(Z_i^\top \beta, \Sigma),$$

where β represents a $(k \times 1)$ vector of coefficients, and Z_i is a $(k \times 2)$ matrix of covariates. The ML estimates of β and Σ can be obtained by the ECM algorithm and the Bayesian analysis can be conducted by placing the semi-conjugate prior distribution,

$$\beta \mid \Sigma \;\sim\; \mathcal{N}(\beta_0, A_0^{-1}), \quad \text{and} \quad \Sigma \;\sim\; \text{InvWish}\,(\nu_0,\; S_0^{-1}),$$

where β_0 is a $(k \times 1)$ vector of prior means, and A_0 is a $(k \times 2)$ matrix of prior precision. The MCMC algorithm can be used to sample from the posterior distribution.

Finally, Imai, Lu, and Strauss (2008b) suggest an alternative approach where the contextual effects are directly modeled without additional covariates. This NCAR (*not coarsened at random*) model is formally defined as,

$$(W_i^*, X_i^*) \mid \eta, \Phi \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\eta, \Phi),$$

where $X_i^* = \text{logit} X_i$, η is a (3×1) vector of population means, and Φ is a (3×3) matrix of covariance. The ML estimates of η and Φ can be obtained by the EM algorithm, whereas the Bayesian analysis of the NCAR model can be conducted in the same way as under the CAR model except that the NCAR model relies upon the trivariate normal distribution rather than the bivariate normal distribution. An advantage of the NCAR model over the CCAR model is that the former does not require the availability of additional covariates to model the contextual effects. Indeed, under the NCAR model, one needs not specify the conditional expectation function of W_i^* given Z_i. The `eco` package implements all three models within the Bayesian or maximum likelihood framework (see Section 3.2).

2.3 Nonparametric Models

To address the *distributional effects*, Imai, Lu, and Strauss (2008b) propose Bayesian nonparametric models based on a Dirichlet process prior (e.g., Dey *et al.*, 1998). This model generalizes the CAR and NCAR parametric models to the case of the unknown distribution of W_i^*. For the CAR assumption, the Bayesian nonparametric model can be written as follows,

$$\begin{aligned} W_i^* \mid \mu_i, \Sigma_i &\sim \mathcal{N}\,(\mu_i,\; \Sigma_i), \\ \mu_i, \Sigma_i \mid G &\sim G, \\ G \mid \alpha &\sim \mathcal{D}(G_0,\; \alpha), \\ \alpha &\sim \text{Gamma}(a_0,\; b_0), \end{aligned}$$

where $\mathcal{D}(G_0, \alpha)$ represents the Dirichlet process prior with the base prior distribution G_0 and the scalar concentration parameter α. Under G_0, (μ_i, Σ_i) is distributed as,

$$\mu_i \mid \Sigma_i \;\sim\; \mathcal{N}\left(\mu_0,\; \frac{\Sigma_i}{\tau_0^2}\right), \quad \text{and} \quad \Sigma_i \;\sim\; \text{InvWish}\,(\nu_0,\; S_0^{-1}).$$

The MCMC algorithm summarized in Imai, Lu, and Strauss (2008b) can be used to sample from the posterior distribution of this model. Furthermore, the nonparametric NCAR model can be formulated in the same manner by using the parametric NCAR model as the base model and specifying the Dirichlet process prior distribution on (η_i, Φ_i), where η and Φ are now indexed by i. The package `eco` implements this Bayesian nonparametric model under both the CAR and NCAR assumptions (see Section 3.4).

2.4 Formal Assessment of Aggregation Effects

The fourth method we implement via the `eco` package is the formal assessment of aggregation effects under the parametric models. Imai, Lu, and Strauss (2008b) propose to measure the effect of data aggregation on parameter estimation and hypothesis testing by calculating the fraction of missing information. The idea is to quantify the amount of information the observed aggregate-level data provide in comparison with the information one would obtain if the individual-level data were available. In the context of parameter estimation, the fraction of missing information is defined as,

$$F_\theta \equiv \text{diag}\left(I - \mathcal{I}_{obs}(\hat{\theta})\mathcal{I}_{com}(\hat{\theta})^{-1}\right), \tag{5}$$

where $\mathcal{I}_{obs}$ is the observed Fisher information matrix and $\mathcal{I}_{com}$ represents the expected information matrix based on the complete-data log-likelihood function. Then, each element of the vector F_θ represents the fraction of missing information for each parameter. In the `eco` package, we use the Supplemented EM (SEM) algorithm (Meng and Rubin, 1991) and compute the fraction of missing information for the parametric CAR and NCAR models (see Section 3.3).

For the hypothesis testing, we follow the approach proposed by Kong, Meng, and Nicolae (2005) and compute the fraction of missing information against the null hypothesis $H_0 : \theta = \theta_0$, which is defined by,

$$F_H \equiv 1 - \frac{l_{obs}(\hat{\theta} \mid Y, X) - l_{obs}(\theta_0 \mid Y, X)}{E[\, l_{com}(\hat{\theta} \mid W, X) - l_{com}(\theta_0 \mid W, X) \mid Y, X; \hat{\theta}]}, \tag{6}$$

where $l_{obs}(\theta \mid Y, X)$ and $l_{com}(\theta \mid W, X)$ represent the observed-data log-likelihood and the complete-data log-likelihood functions, respectively. Moreover, $\hat{\theta}$ is the ML estimate of θ and the expectation is taken over the conditional distribution of W given (Y, X). Then, F_H equals one minus the logarithm of *the observed likelihood ratio statistic* divided by the logarithm of *the expected likelihood ratio statistic*. In the `eco` package, we use the SEM algorithm and compute F_H with the null hypothesis of the equal marginal means, i.e., $H_0 : E(W_1) = E(W_2)$, under the parametric CAR and NCAR models (see Section 3.3).

2.5 Additional Individual-Level Data

When bounds are not informative, ecological inference is difficult. The parametric inference will be sensitive to modeling assumptions, and the nonparametric model will not be able to recover the underlying distribution. Therefore, incorporating individual-level data may be helpful whenever

such additional information is available. For example, one might conduct a survey in randomly selected counties to obtain such information. Sometimes, a small scale survey can be conducted to get rough estimates of W_i for some counties, and incorporating such auxiliary information can also be helpful (Wakefield, 2004). In the `eco` package, it is straightforward to incorporate such information into the estimation of both parametric and nonparametric models (see Section 3.5).

3 Illustrative Examples

In this section, we illustrate how to implement the methods described in Section 2 via the package `eco` using some example data sets which also is a part of the package. The detailed references for the commands and data sets we use appear in Appendices B and C, respectively.

3.1 Computing the Bounds

We first consider the computation of the bounds described in Section 2.1 using the function `ecoBD()`. We illustrate the use of this function with the voter registration data from 275 counties of four Southern states in the United States: Florida, Louisiana, North Carolina, and South Carolina. The data set is taken from King (1997) and is available as `reg` as a part of the `eco` package. To load this data set, type at the R prompt (after loading the package via the `library(eco)` command),

```
> data(reg)
```

which stores the data frame as `reg` in the workspace. The data set can be summarized as,

```
> summary(reg)
       X                 Y               N                W1
 Min.   :0.00826   Min.   :0.297   Min.   :  1800   Min.   :0.000
 1st Qu.:0.13061   1st Qu.:0.678   1st Qu.:  9000   1st Qu.:0.459
 Median :0.24286   Median :0.783   Median : 15500   Median :0.571
 Mean   :0.25725   Mean   :0.777   Mean   : 32448   Mean   :0.562
 3rd Qu.:0.37143   3rd Qu.:0.910   3rd Qu.: 31350   3rd Qu.:0.692
 Max.   :0.73899   Max.   :1.000   Max.   :613000   Max.   :1.000
       W2
 Min.   :0.321
 1st Qu.:0.776
 Median :0.888
 Mean   :0.855
 3rd Qu.:1.000
 Max.   :1.000
```

where `X` is the fraction of black voters in each county, `Y` represents the fraction of registered voters, and `N` is the total number of voters in each county. In this data set, the registration rates are observed separately for blacks and whites, which are given by `W1` and `W2`, respectively.

To compute the bounds using the `reg` data set, we simply use the following syntax,

```
> res.BD <- ecoBD(Y ~ X, data = reg)
> print(resBD)

Call:
ecoBD(formula = Y ~ X, data = reg)

Aggregate Lower Bounds (Proportions):
    c1      c2
r1  0.3426  0.7047
r2  0.0154  0.0729

Aggregate Upper Bounds (Proportions):
    c1     c2
r1  0.985  0.927
r2  0.657  0.295
```

which prints out the aggregate lower and upper bounds. For example, the registration rate for blacks lies between 0.34 and 0.99, while that for whites is between 0.70 and 0.93. The actual registration rates for blacks and whites (which are usually unknown but in this case can be estimated using the sample means of `W1` and `W2` in the dataset `reg`) are 0.56 and 0.86, respectively.

The county-level bounds are also stored in the output object from `ecoBD()`. For example, the bound for the first county can be obtained by the following commands,

```
> res.BD$Wmin[1,,]
         c1       c2
r1 0.545455 0.850498
r2 0.000000 0.000000
> res.BD$Wmax[1,,]
         c1       c2
r1 1.000000 1.000000
r2 0.454545 0.149502
```

It is also possible to incorporate the information about the total number of eligible voters (`N` in the data set). The following commands accomplish this,

```
> res.BD1 <- ecoBD(Y ~ X, N = N, data = reg)
> print(res.BD1)

Call:
ecoBD(formula = Y ~ X, data = reg, N = N)

Aggregate Lower Bounds (Proportions):
    c1      c2
r1  0.2168  0.7055
r2  0.0244  0.0792
```

```
Aggregate Upper Bounds (Proportions):
    c1     c2
r1  0.976  0.921
r2  0.783  0.294

Aggregate Lower Bounds (Counts):
    c1       c2
r1   427500  4904400
r2    48200   550500

Aggregate Upper Bounds (Counts):
    c1       c2
r1  1923800  6400700
r2  1544500  2046800
```

The county-level bounds can be obtained from the output object. They are stored as `Nmin` and `Nmax`.

Finally, `ecoBD()` also computes the bounds for $R \times C$ ecological tables. The syntax is very similar to the 2×2 case. For example, `ecoBD(cbind(Y1, Y2, Y3) ~ X1 + X2 + X3 + X4, data = data` specifies 3×4 ecological tables.

3.2 Fitting the Parametric Models

In this section, we illustrate how to use the `eco` package to fit the parametric ecological inference models. First, we review the maximum likelihood (ML) estimation of the parametric models described in Section 2.2 using the function `ecoML()`. We then demonstrate the fitting of the Bayesian parametric model using the `eco()` function. The dataset used to illustrate these functions is the race and literacy dataset first collected by Robinson (1950) on the state level, and then refined to the county level by King (1997). For 1,040 counties, the marginal percentage of blacks (`X`), the marginal literacy rate (`Y`), and the population (`N`) is available, as well as the true cell-level values for literacy among blacks (`W1`) and whites (`W2`). As before, type the following in an R prompt to view summary statistics of the dataset,

```
> data(census)
> summary(census)
       Y                X                N                 W1
 Min.   :0.4055   Min.   :0.0508   Min.   :    798   Min.   :0.2012
 1st Qu.:0.7790   1st Qu.:0.1412   1st Qu.:   9900   1st Qu.:0.6251
 Median :0.8401   Median :0.3108   Median :  14428   Median :0.6897
 Mean   :0.8258   Mean   :0.3377   Mean   :  21710   Mean   :0.6845
 3rd Qu.:0.8912   3rd Qu.:0.4939   3rd Qu.:  20940   3rd Qu.:0.7513
 Max.   :0.9908   Max.   :0.9393   Max.   :1261132   Max.   :0.9665
       W2
 Min.   :0.5563
```

```
 1st Qu.:0.8936
 Median :0.9302
 Mean   :0.9189
 3rd Qu.:0.9599
 Max.   :0.9940
```

The Maximum Likelihood Estimation. We first demonstrate the ML estimation of the CAR model, which assumes no contextual effect, via the EM algorithm. Using the default values provided by the function (see the command reference in Appendix B), with the exception of suppressing the output, the following command fits the model and stores its output as `res.ML`,

```
> res.ML <- ecoML(Y ~ X, data = census, verbose = FALSE)
```

As this fitting process via the EM algorithm may take a long time on some computers, setting `verbose` to the value of `TRUE` (its default value) tracks the progress of the program,

```
> res.ML <- ecoML(Y ~ X, data = census, verbose = TRUE)
OPTIONS (flag: 4) Ncar: No; Fixed Rho: No; SEM: First run
cycle 1/1000: 0.000 0.000 1.000 1.000 0.000
cycle 2/1000: 1.223 1.886 0.698 1.057 0.010
cycle 3/1000: 1.234 2.151 0.602 0.877 -0.051 Prev LL: -1299.52
cycle 4/1000: 1.161 2.249 0.560 0.823 -0.070 Prev LL: -1230.91
cycle 5/1000: 1.090 2.320 0.524 0.807 -0.070 Prev LL: -1208.75
[output truncated for presentation purposes]
cycle 349/1000: 0.654 2.785 0.236 0.916 0.271 Prev LL: -1126.77
Final Theta: 0.654 2.785 0.236 0.916 0.271 Final LL: -1126.77
```

The `OPTIONS` output line is for verification purposes, and informs the user that there are no contextual effect to be modeled, the correlation parameter (ρ) is not fixed by the user, and that the fraction of missing information will be calculated via the SEM algorithm. The next lines of output display the iteration number of the EM loop, the parameter values for that iteration, and the observed log-likelihood for the previous set of parameter values. As showed by the `cycle 1` output line, the default starting parameter values for CAR are $\mu_0 = 1$, $\mu_2 = 0$, $\sigma_1^2 = 0$, $\sigma_1^2 = 1$, $\rho = 0$.

A few remarks about the convergence are worth mentioning although we refer readers to the relevant literature for the details (e.g., McLaughlan and Krishnan, 1997). When the absolute value difference between each of the parameter values from the previous iteration falls below the convergence threshold, `epsilon`, the algorithm stops. The last line of the output displays the final, converged parameter estimates. The default value for `epsilon` is 10^{-10}; this threshold can be changed by the user. The number of maximum iterations to cycle through before halting (`maxit`) can also be adjusted; the default value for `maxit` is $1,000$. The failure to set these inputs to appropriate values may result in inaccurate estimates.

Once the EM algorithm is completed, the SEM algorithm will begin automatically (as long as the `SEM` parameter is not set to `FALSE`). See Section 3.3 for details on executing the SEM algorithm. If the fraction of missing information is not desired, the user should set `SEM` to `FALSE` in the interest of computational time.

Summary statistics for the fitted model can be displayed simply by using the `summary()` function (a shorter summary is available via the `print()` function),

```
> summary(res.ML)

Call:  ecoML(formula = Y ~ X, data = census, verbose = TRUE)

*** Parameter Estimates ***

Original Model Parameters:
                  mu1     mu2  sigma1  sigma2   rho
ML est.       0.65354 2.78466 0.23574 0.91588 0.271
std. err.     0.03259 0.06440 0.02029 0.10265 0.093
frac. missing 0.62566 0.56690 0.66159 0.64286 0.772

*** Insample Predictions ***

Unweighted:
      mean std.dev   2.5 % 97.5 %
W1 0.65007 0.07564 0.48492  0.802
W2 0.91973 0.05908 0.79422  0.986

Weighted:
      mean std.dev   2.5 % 97.5 %
W1 0.64645 0.07891 0.49178  0.801
W2 0.92407 0.06796 0.79087  1.057


Log-likelihood: -1126.773
Number of Observations: 1040
Number of EM iterations: 350
Number of SEM iterations: 95
Convergence threshold for EM: 1e-10
```

where the "weighted" insample predictions are computed based on the weights proportional to the group-specific population size while assuming the overall population size is the same across counties (when the information about overall population size is available, this information can be used via the option `N` as shown in the next example below).

Under the CAR assumption, the point estimate for the unweighted, mean black literacy rate is 66% and for the unweighted, mean white literacy, 92%. The estimated unweighted standard deviations for the county-level black and white literacy rates are 7.6% and 5.9% respectively. The correlation between logit-transformed county literacy rates is estimated to be 0.271 (`rho`). In addition to the aggregate-level in-sample predictions, county-level estimates are available in

the $n \times 2$ matrix `res.ML$W`. Use the `names()` command to display all available elements of the output object returned by the `ecoML()` function.

Bayesian Estimation. Next, we illustrate how to fit the Bayesian parametric models via the MCMC algorithms using `eco`. Here, we use the NCAR model, which unlike the CAR model assumes the existence of contextual effect. First, the model can be fitted by the Gibbs sampler using the following syntax,

```
> res <- eco(Y ~ X, N = N, data = census, context = TRUE, parameter = TRUE,
             verbose = TRUE)

Starting Gibbs Sampler...
 10 percent done.
 20 percent done.
 30 percent done.
[output truncated for presentation purposes]
100 percent done.
```

where `context=TRUE` indicates that the NCAR ecological inference model is to be fitted, `parameter = TRUE` means that the posterior draws of the parameters will be saved in the output `res` in order to make out-of-sample predictions based on the fitted model, and `N` specifies the total number of individual-level observations within each aggregate unit so that both weighted and unweighted estimates can be obtained. The progress of the Gibbs sampler is printed to the screen if `verbose` is set to be `TRUE`. Moreover, one can specify the prior distribution for the parameters of the multivariate normal distribution in `eco()` (see Appendix B for details). The default is a non-informative prior, which is approximately uniform on W_i.

In this example, the other parameters are all set to be the default values. In particular, only $5,000$ Gibbs draws are taken, with no initial burn-in draws. In practice, to ensure proper convergence, the MCMC should be run for a longer period and the initial draws should also be discarded so that inference will be made based on draws from the target posterior distribution. We refer the readers to the standard texts on Bayesian data analysis for general advice on convergence (e.g., Gelman *et al.*, 2004). Here, we implement the following syntax, which fits the same NCAR model as above but discards the initial $20,000$ draws from a total of $50,000$ draws while saving every 10th draw (thus, using the thinning interval of 10),

```
> res <- eco(Y ~ X, N = N, data = census, context = TRUE, parameter=TRUE,
             n.draws = 50000, burnin = 20000, thin = 9, verbose = TRUE)
```

The summary of the fitted model can be viewed using the `summary()` function,

```
> summary(res)

Call: eco(formula = Y ~ X, data = census, N = N, context = TRUE,
parameter = TRUE, n.draws = 50000, burnin = 20000, thin = 9, verbose
= TRUE)
```

```
Parameter Estimates:
            mean  std.dev    2.5 % 97.5 %
mu1      0.83226  0.11702  0.63021  1.081
mu2      2.66263  0.16997  2.33865  2.975
mu3     -0.85978  0.03631 -0.93065 -0.790
Sigma11  0.29302  0.04199  0.22474  0.387
Sigma12  0.03482  0.04327 -0.05616  0.111
Sigma13 -0.27642  0.06776 -0.42183 -0.150
Sigma22  0.92734  0.10776  0.74044  1.152
Sigma23 -0.03642  0.11408 -0.25441  0.181
Sigma33  1.37280  0.06057  1.25952  1.495

*** Insample Predictions ***

Unweighted:
      mean std.dev   2.5 % 97.5 %
W1 0.65658 0.01714 0.62709  0.692
W2 0.91214 0.00874 0.89384  0.927

Weighted:
      mean std.dev   2.5 % 97.5 %
W1 0.67518 0.01655 0.64686  0.709
W2 0.93444 0.00874 0.92019  0.927

Number of Units: 1040

Number of Monte Carlo Draws: 3000
```

where the first part of the output summarizes the posterior distribution of the parameters of the trivariate normal distribution for the NCAR model – the ordinates of the distribution are (W_1^*, W_2^*, X^*), i.e., the logit-transformed black literacy rate, white literacy rate and black composition, respectively. Since `N` is specified in `eco()`, both "weighted" in-sample predictions aggregate estimates according to their actual group-specific population size. After controlling for the possible correlation between racial composition and literacy rate (i.e., contextual effect), the in-sample estimates, especially those of `W1`, are slightly improved compared to the CAR model (see the ML estimation of the CAR model earlier in this section).

In addition, the `predict()` function allows one to obtain the out-of-sample predictions of W_1 and W_2 based on their posterior predictive distributions using the posterior distribution of the model parameters. In the case of the NCAR model, the predictions will be based on the values of X_i which can be taken from another data set using the option `newdata` (the default, which we use here, is the data set used to fit the model). As before, the `summary()` function will summarize the results,

```
> out <- predict(res, verbose = TRUE)
```

```
 10 percent done.
 20 percent done.
 30 percent done.
 [output truncated for presentation purposes]
 100 percent done.

> summary(out)

Out-of-sample Prediction:
      mean std.dev   2.5 % 97.5 %
W1 0.68743 0.11380 0.43362  0.879
W2 0.90830 0.08170 0.69232  0.990
X  0.33622 0.21468 0.04528  0.818

Number of Monte Carlo Draws: 3000
```

The default number of Monte Carlo draws is the same as the number of MCMC draws stored in the object `res`, but this number can be changed by users via the `newdraw` option.

3.3 Quantifying the Aggregation Effects

We revisit the function `ecoML()` to outline the computation of the fraction of missing information calculation described in Section 2.4. As in Section 3.2, the `census` dataset is used. If the `SEM` option is left at its default value of `TRUE`, the `ecoML()` function proceeds from the SEM algorithm, once the SEM algorithm is completed. The transition is shown as follows.

```
> res.ML <- ecoML(Y ~ X, data = census, verbose = TRUE)
[output truncated for presentation purposes]
cycle 349/1000: 0.654 2.785 0.236 0.916 0.271 Prev LL: -1126.77
Final Theta: 0.654 2.785 0.236 0.916 0.271 Final LL: -1126.77
OPTIONS (flag: 4)   Ncar: No; Fixed Rho: No; SEM: Second run
cycle 1/1000: 0.000 0.000 1.000 1.000 0.000

R Matrix row 1 (Not done):   0.60  -0.75   0.17  -0.11  -0.16
R Matrix row 2 (Not done):  -0.20   0.40   0.01   0.10  -0.03
R Matrix row 3 (Not done):   0.01  -0.01   0.63  -0.14  -0.15
R Matrix row 4 (Not done):   0.00   0.10  -0.18   0.59  -0.20
R Matrix row 5 (Not done):  -0.07  -0.07  -0.26  -0.32   0.67

cycle 2/1000: 1.223 1.886 0.698 1.057 0.010
[output truncated for presentation purposes]
```

Note the final, converged parameter values (e.g., $\hat{\mu_1} = 0.654$). The SEM algorithm then begins to calculate the DM matrix, which is necessary for the computation of the asymptotic variance

matrix. Each row converges independently of the other rows, and the program output tracks this progress.

Once the SEM algorithm has converged, the final estimate for the DM matrix is displayed:

```
> res.ML <- ecoML(Y ~ X, data = census, verbose=TRUE)
[output truncated for presentation purposes]
cycle 94/1000: 0.654 2.785 0.236 0.917 0.270 Prev LL: -1126.77

R Matrix row 1 (    Done):   0.52  -0.73  -0.08  -0.13  -0.25
R Matrix row 2 (    Done):  -0.20   0.46   0.01   0.19  -0.02
R Matrix row 3 (    Done):  -0.00   0.01   0.61  -0.18  -0.16
R Matrix row 4 (    Done):   0.01   0.09  -0.19   0.58  -0.20
R Matrix row 5 (    Done):  -0.06  -0.08  -0.28  -0.35   0.69

Final Theta: 0.654 2.785 0.236 0.917 0.270 Final LL: -1126.77
```

To obtain an estimate of the fraction of missing data simply type (see Equation 5),

```
> res.ML$Fmis
[1] 0.6256639 0.5668982 0.6615895 0.6428587 0.7721757
```

For instance, the fraction of missing information for the calculation of black literacy is about 62.6%. The analogous quantity for white literacy is lower because in some counties, whites make up a very large proportion of the population, thus resulting in tighter bounds.

In addition to computing the fraction of missing data on parameter estimation, the `eco` package includes limited functionality for hypothesis testing (see Section 2.4). Currently, setting the `hyptest` parameter to `TRUE` for the `ecoML()` function, will calculate quantities of interest with parameters constrained to the null hypothesis: $\mu_1 = \mu_2$. Continuing with the census example, this null hypothesis would be that the mean black literacy rate is equal to the white literacy rate. To restrict the parameter space to this hypothesis, simply type,

```
> res.ML.HT <- ecoML(Y ~ X, data = census, verbose = FALSE, hyptest=TRUE)
```

With the results of the algorithm for both the constrained and unconstrained problem, the calculation of the fraction of missing information under this hypothesis is straightforward. First, calculate the observed log-likelihood ratio test statistic, then the complete log-likelihood ratio test statistic, and then subtract the ratio of those two quantities from 1 (see Equation 6).

```
> n <- dim(census)[1]
> obs.loglik.stat <- 2*(res.ML$loglik -  res.ML.HT$loglik)
> com.loglik.stat <- Qfun(res.ML$theta.em, res.ML$suff.stat, n) -
   Qfun(res.ML.HT$theta.em, res.ML.HT$suff.stat, n)
> frac.miss.data <- 1 - (obs.loglik.stat / com.loglik.stat)
> obs.loglik.stat
[1] 462.8226
> frac.miss.data
```

3.4 Fitting the Nonparametric Models

To avoid the distributional assumptions that are common to parametric models, the `eco` package also fits the nonparametric Bayesian models described in Section 2.3. Here, we use the voter registration dataset (see Section 2.1) to illustrate the use of the `ecoNP()` function, which fits the Bayesian nonparametric models via the MCMC algorithms. The following command fits the nonparametric model under the CAR assumption,

```
> res <- ecoNP(Y ~ X, data = reg, N = N, parameter = TRUE, n.draws = 50000,
               burnin = 20000, thin = 9, verbose = TRUE)

Starting Gibbs Sampler...
 10 percent done.
 20 percent done.
 30 percent done.
 [output truncated for presentation purposes]
100 percent done.
```

where the default prior specification places a diffuse distribution on the concentration parameter of the Dirichlet process prior. This default specification can be changed by the user (see Appendix B). Many of the inputs of `ecoNP()` are the same as those of `eco()`. For example, the nonparametric NCAR model can be fitted by the `context = TRUE` option.

Like the Bayesian parametric models, one can summarize the in-sample estimates of W using `summary()`. Unlike the parametric models, however, no parameter estimates are presented since the distribution of W is estimated nonparametrically based on a mixture of normal distributions,

```
> summary(res)

Call: ecoNP(formula = Y ~ X, data = reg, N = N, parameter = TRUE,
n.draws = 50000, burnin = 20000, thin = 9,verbose = TRUE)

*** Insample Predictions ***

Unweighted:
      mean std.dev   2.5 % 97.5 %
W1 0.58350 0.03377 0.51929  0.650
W2 0.84368 0.01170 0.82057  0.866

Weighted:
      mean std.dev   2.5 % 97.5 %
W1 0.53855 0.04835 0.44351  0.631
W2 0.82952 0.01372 0.80333  0.856

Number of Units: 275
Number of Monte Carlo Draws: 3000
```

Finally, out-of sample predictions can be made in the same way as done for the parametric Bayesian model. That is, we use the generic `predict()` and `summary()` functions,

```
> out <- predict(res, verbose = TRUE)
 10 percent done.
 20 percent done.
 30 percent done.
 [output truncated for presentation purposes]
100 percent done.
> summary(out)

Out-of-sample Prediction:
      mean std.dev   2.5 % 97.5 %
W1 0.59176 0.25188 0.05434  0.997
W2 0.82800 0.20838 0.25976  1.000

Number of Monte Carlo Draws: 825000
```

Since the distribution is estimated nonparametrically, the out-of-sample prediction is generated for each observation. Hence, the total number of Monte Carlo draws is $825,000 = 275 \times 3,000$.

3.5 Incorporating the Individual-level Data

If survey or other supplemental, individual-level data is available, this information can easily be incorporated into the models. Adding this data will alleviate the adverse aggregation effects endemic to ecological inference. In `eco` package, the three main functions, `eco()`, `ecoML()`, `ecoNP()`, all take supplemental individual level data. In this section we illustrate an example using `ecoML()` and `eco()`.

Continuing with the county literacy rates example, assume that the actual, within-county literacy for whites and blacks is collected for the first 100 counties (and only these counties). Simply, create an $n \times 2$ matrix of the supplemental data with the first column `W1` and the second column `W2`,

```
> survey.records<-1:100
> survey.data <- census[survey.records, c("W1","W2")]
```

Next, execute the `ecoML()` set the `supplement` parameter to the matrix of survey data as follows,

```
> res <- ecoML(Y ~ X, data=census[-survey.records,],
    supplement=survey.data,verbose=FALSE)
> res$theta.em
       u1        u2        s1        s2       r12
0.6907367 2.6884871 0.2355874 0.7316871 0.4019745
> res$Fmis
[1] 0.6194664 0.5708665 0.6396823 0.6209216 0.7701529
```

The estimated mean black literacy rate is about 65%, with the fraction of missing data being 61.9%, 0.6 percentage points less than without the supplemental data. Thus, adding true values for about 10% of the data points did not reduce information loss by much in this case.

Fitting the NCAR model is similar to the operation of the CAR model above. Since, no additional covariates are needed, simply adjust the `context` parameter,

```
> survey.records<-1:100
> survey.data <- census[survey.records, c("W1","W2","X")]
> res.ML.NCAR <- ecoML(Y ~ X, data=census[-survey.records,],
    supplement=survey.data,context=TRUE,verbose=FALSE)
> res.ML.NCAR$theta.em
         ux          u1          u2          sx          s1          s2
-0.86201800  0.90322700  2.57480315  1.35928401  0.32312537  0.54261008
        r1x         r2x         r12
-0.60889544  0.08569204  0.38889276
> res.ML.NCAR$Fmis
        ux         u1         u2         sx         s1         s2        r1x
-0.6157753  0.7211584  0.4009365  0.8480136  0.8445779 -0.4313306  0.8206274
       r2x        r12
 0.8461927  0.8343209
```

Under the NCAR model, the mean black literacy rate is estimated to be 71%, which is higher than estimated under the CAR model with the same survey data provided. The average white literacy rate is adjusted slightly downward in this more general model, from 94% to 93%. The change in the fraction of missing data is not monotonic when switching models. Information loss due to aggregation is larger for estimating the mean black literacy rate under NCAR, but is smaller when estimating mean white literacy.

Similarly, the `eco` package can fit Bayesian models with supplemented individual level data. Below, we fit the parametric CAR and NCAR models via `eco()` using the same census data with the first 100 counties's data revealed.

```
> survey.records<-1:100
> survey.data <- census[survey.records, c("W1","W2")]
> res.CAR <- eco(Y ~ X, data = census[-survey.records, ], N = N,
    supplement = survey.data, parameter = TRUE, n.draws = 50000,
    burnin = 20000, thin = 9, verbose = FALSE)

> survey.data <- census[survey.records, c("W1","W2","X")]
> res.NCAR <- eco(Y ~ X, data = census[-survey.records, ], N = N,
    supplement = survey.data, context=TRUE, parameter = TRUE, n.draws = 50000,
    burnin = 20000, thin = 9, verbose = FALSE)
```

Then, the posterior means of the parameters can be computed directly from the output objects (or via the `summary()` function).

```
> colMeans(res.CAR$mu)
```

```
      mu1       mu2
0.6193476 2.8169714
> colMeans(res.CAR$Sigma)
  Sigma11   Sigma12   Sigma22
0.2303096 0.1415191 0.9134931
> colMeans(res.NCAR$mu)
       mu1        mu2        mu3
 0.7822204  2.6813690 -0.8607085
> colMeans(res.NCAR$sigma)
    Sigma11     Sigma12     Sigma13     Sigma22     Sigma23     Sigma33
 0.25194916  0.13371959 -0.26838902  0.73719227  0.02965051  1.36985765
```

Under the Bayesian CAR model, the mean black literacy rate is estimated to be 64.3% and white literacy rate 92.4%. Under the NCAR model, these two estimates are 68% and 91.8%, respectively. Comparing to the sample estimates (68.4% and 91.9%), `ecoNP()` performs very well in this particular data set with the aid of 100 individual-level observations.

References

Achen, C. H. and Shively, W. P. (1995). *Cross-Level Inference*. University of Chicago Press, Chicago.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Publicazioni del R Instituto Superiore di Scienze Economiche e Commerciali di Fienze* **8**, 3–62.

Cho, W. K. T. (1998). Iff the assumptions fits...:a comment on the King ecological inference solution. *Political Analysis* **7**, 143–163.

Cho, W. K. T. and Gaines, B. J. (2004). The limits of ecological inference: The case of split-ticket voting. *American Journal of Political Science* **48**, 1, 152–171.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–37.

Dey, D., Müller, P., and Sinha, D., eds. (1998). *Practical nonparametric and semiparametric Bayesian statistics*. Springer-Verlag Inc, New York.

Duncan, O. D. and Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review* **18**, 6, 665–666.

Fréchet, M. (1940). *Les Probabilitiés, Associées a un Système d'Événments Compatibles et Dépendants*, vol. Premiere Partie. Hermann & Cie, Paris.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy. In N. Smelser and P. Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences*, vol. 6, 4027–4030. Elsevier.

Freedman, D. A., Klein, S. P., Sacks, J., Smyth, C. A., and Everett, C. G. (1991). Ecological regression and voting rights (with discussion). *Evaluation Review* **15**, 673–816.

Freedman, D. A., Ostland, M., Roberts, M. R., and Klein, S. P. (1998). "Review of 'A Solution to the Ecological Inference Problem' ". *Journal of the American Statistical Association* **93**, 1518–1522.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall, London, 2nd edn.

Gelman, A., Park, D. K., Ansolabehere, S., Price, P. N., and Minnite, L. C. (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society, Series A* **164**, 101–118.

Goodman, L. (1953). Ecological regressions and behavior of individuals. *American Sociological Review* **18**, 663–666.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *The American Journal of Sociology* **64**, 610–624.

Greenland, S. and Robins, J. M. (1994). Ecologic studies: Biases, misconceptions, and counterexamples. *American Journal of Epidemiology* **139**, 747–760.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics* **19**, 2244–2253.

Hoeffding, W. (1940). Masstabinvariate korrelationstheorie. *Schriften des Matematischen Instituts und des Instituts für Angewandte Matematik de Universität Berlin* **5**, 179–233. Reprinted as Scale-invariant correlation theory. in Fishedr, N.I. and Sen, P.K., editors (1994). *The Collected Works of Wassily Hoeffding*, pp. 57–107. Springer, New York.

Imai, K. and King, G. (2004). Did illegal overseas absentee ballots decide the 2000 U.S. presidential election? *Perspectives on Politics* **2**, 3, 537–549.

Imai, K., King, G., and Lau, O. (2008a). Toward a common framework of statistical analysis and development. *Journal of Computational and Graphical Statistics* **17**, 4, 892–913.

Imai, K., Lu, Y., and Strauss, A. (2008b). Bayesian and likelihood inference for 2×2 ecological tables: An incomplete data approach. *Political Analysis* **16**, 1, 41–69.

Imai, K., Lu, Y., and Strauss, A. (In-press). eco: R package for ecological inference in 2 x 2 tables. *Journal of Statistical Software* Abstract reprinted in *Journal of the Computational and Graphical Statistics*. Software available at The Comprehensive R Archive Network, `http://cran.r-project.org/`.

Judge, G. G., Miller, D. J., and Cho, W. K. T. (2004). *Ecological Inference: New Methodological Strategies (eds. G. King, O. Rosen, and M. Tanner)*, chap. An Information Theoretic Approach to Ecological Estimation and Inference, 162–187. Cambridge University Press, Cambridge.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton, NJ.

King, G. (1999). Comment on "Review of 'A Solution to the Ecological Inference Problem' ". *Journal of the American Statistical Association* **94**, 352–355.

King, G., Rosen, O., and Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research* **28**, 61–90.

King, G., Rosen, O., and Tanner, M. A., eds. (2004). *Ecological Inferece: New Methodological Strategies*. Cambridge University Press.

Kong, A., Meng, X.-L., and Nicolae, D. L. (2005). Quantifying relative incomplete information for hypothesis testing in statistical and genetic studies. *Unpublished Manuscript* Department of Statistics, Harvard University.

Martin, A. D. and Quinn, K. M. (2006). *MCMCpack: Markov chain Monte Carlo (MCMC) Package*.

McLaughlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, New York.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* **15**, 3, 351–357.

Rosen, O., Jiang, W., King, G., and Tanner, M. A. (2001). Bayesian and frequentist inference for ecological inference: The R×C case. *Statistica Neerlandica* **55**, 2, 134–156.

Wakefield, J. (2004). Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* **167**, 385–445.

Appendices

A What's New?

This section summarizes the history of all prior changes that are made to the `eco` package.

version	date	changes
$3.1-2$	01.29.09	minor documentation fixes
$3.1-1$	06.27.07	some minor improvements; final version for JSS publication
$3.0-2$	01.11.07	made it comparible with the Windows; a bug fix in `summary.ecoML()`
$3.0-1$	12.27.06	a major revision; added ML estimation, calculation of fraction of missing information, stable release for R-2.4.1
$2.2-2$	09.23.06	changed due to updates in R
$2.2-1$	09.28.05	nonparametric model with contextual effects added
$2.1-1$	07.06.05	a major revision; added bounds and prediction; added/updated other functionalities
$1.1-1$	06.15.05	add the Metropolis algorithm to sample W
$1.0-1$	12.21.04	first official version; submitted to CRAN
$0.9-1$	09.07.04	first beta version

B Main Command References

`eco` *Fitting the Parametric Bayesian Model of Ecological Inference in 2x2 Tables*

Description

`eco` is used to fit the parametric Bayesian model (based on a Normal/Inverse-Wishart prior) for ecological inference in 2×2 tables via Markov chain Monte Carlo. It gives the in-sample predictions as well as the estimates of the model parameters. The model and algorithm are described in Imai, Lu and Strauss (2008, Forthcoming).

Usage

```
eco(formula, data = parent.frame(), N = NULL, supplement = NULL,
    context = FALSE, mu0 = 0, tau0 = 2, nu0 = 4, S0 = 10,
    mu.start = 0, Sigma.start = 10, parameter = TRUE,
    grid = FALSE, n.draws = 5000, burnin = 0, thin = 0,
    verbose = FALSE)
```

Arguments

`formula`	A symbolic description of the model to be fit, specifying the column and row margins of 2×2 ecological tables. `Y ~ X` specifies `Y` as the column margin (e.g., turnout) and `X` as the row margin (e.g., percent African-American). Details and specific examples are given below.
`data`	An optional data frame in which to interpret the variables in `formula`. The default is the environment in which `eco` is called.
`N`	An optional variable representing the size of the unit; e.g., the total number of voters.
`supplement`	An optional matrix of supplemental data. The matrix has two columns, which contain additional individual-level data such as survey data for W_1 and W_2, respectively. If `NULL`, no additional individual-level data are included in the model. The default is `NULL`.
`context`	Logical. If `TRUE`, the contextual effect is also modeled, that is to assume the row margin X and the unknown W_1 and W_2 are correlated. See Imai, Lu and Strauss (2008, Forthcoming) for details. The default is `FALSE`.
`mu0`	A scalar or a numeric vector that specifies the prior mean for the mean parameter μ for (W_1, W_2) (or for (W_1, W_2, X) if `context=TRUE`). When the input of `mu0` is a scalar, its value will be repeated to yield a vector of the length of μ, otherwise, it needs to be a vector of same length as μ. When `context=TRUE`, the length of μ is 3, otherwise it is 2. The default is `0`.
`tau0`	A positive integer representing the scale parameter of the Normal-Inverse Wishart prior for the mean and variance parameter (μ, Σ). The default is `2`.
`nu0`	A positive integer representing the prior degrees of freedom of the Normal-Inverse Wishart prior for the mean and variance parameter (μ, Σ). The default is `4`.
`S0`	A positive scalar or a positive definite matrix that specifies the prior scale matrix of the Normal-Inverse Wishart prior for the mean and variance parameter (μ, Σ) . If it is a scalar, then the prior scale matrix will be a diagonal matrix with the same dimensions as Σ and the diagonal elements all take value of `S0`, otherwise `S0` needs to have same dimensions as Σ. When `context=TRUE`, Σ is a 3×3 matrix, otherwise, it is 2×2. The default is `10`.
`mu.start`	A scalar or a numeric vector that specifies the starting values of the mean parameter μ. If it is a scalar, then its value will be repeated to yield a vector of the length of μ, otherwise, it needs to be a vector of same length as μ. When `context=FALSE`, the length of μ is 2, otherwise it is 3. The default is `0`.
`Sigma.start`	A scalar or a positive definite matrix that specified the starting value of the variance matrix Σ. If it is a scalar, then the prior scale matrix will be a diagonal matrix with the same dimensions as Σ and the diagonal elements all

take value of `S0`, otherwise `S0` needs to have same dimensions as Σ. When `context=TRUE`, Σ is a 3×3 matrix, otherwise, it is 2×2. The default is `10`.

`parameter` Logical. If `TRUE`, the Gibbs draws of the population parameters, μ and Σ, are returned in addition to the in-sample predictions of the missing internal cells, W. The default is `TRUE`.

`grid` Logical. If `TRUE`, the grid method is used to sample W in the Gibbs sampler. If `FALSE`, the Metropolis algorithm is used where candidate draws are sampled from the uniform distribution on the tomography line for each unit. Note that the grid method is significantly slower than the Metropolis algorithm. The default is `FALSE`.

`n.draws` A positive integer. The number of MCMC draws. The default is `5000`.

`burnin` A positive integer. The burnin interval for the Markov chain; i.e. the number of initial draws that should not be stored. The default is `0`.

`thin` A positive integer. The thinning interval for the Markov chain; i.e. the number of Gibbs draws between the recorded values that are skipped. The default is `0`.

`verbose` Logical. If `TRUE`, the progress of the Gibbs sampler is printed to the screen. The default is `FALSE`.

Details

An example of 2×2 ecological table for racial voting is given below:

	black voters	white voters	
vote	W_{1i}	W_{2i}	Y_i
not vote	$1 - W_{1i}$	$1 - W_{2i}$	$1 - Y_i$
	X_i	$1 - X_i$	

where Y_i and X_i represent the observed margins, and W_1 and W_2 are unknown variables. In this exmaple, Y_i is the turnout rate in the ith precint, X_i is the proproption of African American in the ith precinct. The unknowns W_{1i} an dW_{2i} are the black and white turnout, respectively. All variables are proportions and hence bounded between 0 and 1. For each i, the following deterministic relationship holds, $Y_i = X_i W_{1i} + (1 - X_i) W_{2i}$.

Value

An object of class `eco` containing the following elements:

`call` The matched call.

`X` The row margin, X.

`Y` The column margin, Y.

`N` The size of each table, N.

`burnin`	The number of initial burnin draws.
`thin`	The thinning interval.
`nu0`	The prior degrees of freedom.
`tau0`	The prior scale parameter.
`mu0`	The prior mean.
`S0`	The prior scale matrix.
`W`	A three dimensional array storing the posterior in-sample predictions of W. The first dimension indexes the Monte Carlo draws, the second dimension indexes the columns of the table, and the third dimension represents the observations.
`Wmin`	A numeric matrix storing the lower bounds of W.
`Wmax`	A numeric matrix storing the upper bounds of W.
`mu`	The posterior draws of the population mean parameter, μ.
`Sigma`	The posterior draws of the population variance matrix, Σ.

Author(s)

Kosuke Imai, Department of Politics, Princeton University, ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Department of Sociology, University of Colorado at Boulder, ⟨ying.lu@Colorado.Edu⟩

References

Imai, Kosuke, Ying Lu and Aaron Strauss. (Forthcoming). "eco: R Package for Ecological Inference in 2x2 Tables" Journal of Statistical Software, available at `http://imai.princeton.edu/research/eco.html`

Imai, Kosuke, Ying Lu and Aaron Strauss. (2008). "Bayesian and Likelihood Inference for 2 x 2 Ecological Tables: An Incomplete Data Approach" Political Analysis, Vol. 16, No. 1 (Winter), pp. 41-69. available at `http://imai.princeton.edu/research/eiall.html`

See Also

`ecoML`, `ecoNP`, `predict.eco`, `summary.eco`

Examples

```
## load the registration data
data(reg)

## NOTE: convergence has not been properly assessed for the following
## examples. See Imai, Lu and Strauss (2008, Forthcoming) for more
```

```
## complete analyses.

## fit the parametric model with the default prior specification
res <- eco(Y ~ X, data = reg, verbose = TRUE)
## summarize the results
summary(res)

## obtain out-of-sample prediction
out <- predict(res, verbose = TRUE)
## summarize the results
summary(out)

## load the Robinson's census data
data(census)

## fit the parametric model with contextual effects and N
## using the default prior specification
res1 <- eco(Y ~ X, N = N, context = TRUE, data = census, verbose = TRUE)
## summarize the results
summary(res1)

## obtain out-of-sample prediction
out1 <- predict(res1, verbose = TRUE)
## summarize the results
summary(out1)
```

`ecoBD` *Calculating the Bounds for Ecological Inference in RxC Tables*

Description

`ecoBD` is used to calculate the bounds for missing internal cells of $R \times C$ ecological table. The data can be entered either in the form of counts or proportions.

Usage

```
ecoBD(formula, data = parent.frame(), N = NULL)
```

Arguments

`formula`	A symbolic description of ecological table to be used, specifying the column and row margins of $R \times C$ ecological tables. Details and specific examples are given below.

`data` An optional data frame in which to interpret the variables in `formula`. The default is the environment in which `ecoBD` is called.

`N` An optional variable representing the size of the unit; e.g., the total number of voters. If `formula` is entered as counts and the last row and/or column is omitted, this input is necessary.

Details

The data may be entered either in the form of counts or proportions. If proportions are used, `formula` may omit the last row and/or column of tables, which can be calculated from the remaining margins. For example, `Y ~ X` specifies `Y` as the first column margin and `X` as the first row margin in 2×2 tables. If counts are used, `formula` may omit the last row and/or column margin of the table only if `N` is supplied. In this example, the columns will be labeled as `X` and `not X`, and the rows will be labeled as `Y` and `not Y`.

For larger tables, one can use `cbind()` and `+`. For example, `cbind(Y1, Y2, Y3) ~ X1 + X2 + X3 + X4)` specifies 3×4 tables.

An $R \times C$ ecological table in the form of counts:

$$\begin{array}{ccccc} n_{i11} & n_{i12} & \ldots & n_{i1C} & n_{i1.} \\ n_{i21} & n_{i22} & \ldots & n_{i2C} & n_{i2.} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ n_{iR1} & n_{iR2} & \ldots & n_{iRC} & n_{iR.} \\ n_{i.1} & n_{i.2} & \ldots & n_{i.C} & N_i \end{array}$$

where $n_{nr.}$ and $n_{i.c}$ represent the observed margins, N_i represents the size of the table, and n_{irc} are unknown variables. Note that for each i, the following deterministic relationships hold; $n_{ir.} = \sum_{c=1}^{C} n_{irc}$ for $r = 1, \ldots, R$, and $n_{i.c} = \sum_{r=1}^{R} n_{irc}$ for $c = 1, \ldots, C$. Then, each of the unknown inner cells can be bounded in the following manner,

$$\max(0, n_{ir.} + n_{i.c} - N_i) \leq n_{irc} \leq \min(n_{ir.}, n_{i.c}).$$

If the size of tables, `N`, is provided,

An $R \times C$ ecological table in the form of proportions:

$$\begin{array}{ccccc} W_{i11} & W_{i12} & \ldots & W_{i1C} & Y_{i1} \\ W_{i21} & W_{i22} & \ldots & W_{i2C} & Y_{i2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ W_{iR1} & W_{iR2} & \ldots & W_{iRC} & Y_{iR} \\ X_{i1} & X_{i2} & \ldots & X_{iC} & \end{array}$$

where Y_{ir} and X_{ic} represent the observed margins, and W_{irc} are unknown variables. Note that for each i, the following deterministic relationships hold; $Y_{ir} = \sum_{c=1}^{C} X_{ic} W_{irc}$ for $r = 1, \ldots, R$, and $\sum_{r=1}^{R} W_{irc} = 1$ for $c = 1, \ldots, C$. Then, each of the inner cells of the table can be bounded

in the following manner,

$$\max(0, (X_{ic} + Y_{ir} - 1)/X_{ic}) \leq W_{irc} \leq \min(1, Y_{ir}/X_{ir}).$$

Value

An object of class `ecoBD` containing the following elements (When three dimensional arrays are used, the first dimension indexes the observations, the second dimension indexes the row numbers, and the third dimension indexes the column numbers):

`call`	The matched call.
`X`	A matrix of the observed row margin, X.
`Y`	A matrix of the observed column margin, Y.
`N`	A vector of the size of ecological tables, N.
`aggWmin`	A three dimensional array of aggregate lower bounds for proportions.
`aggWmax`	A three dimensional array of aggregate upper bounds for proportions.
`Wmin`	A three dimensional array of lower bounds for proportions.
`Wmax`	A three dimensional array of upper bounds for proportions.
`Nmin`	A three dimensional array of lower bounds for counts.
`Nmax`	A three dimensional array of upper bounds for counts.

The object can be printed through `print.ecoBD`.

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu/`; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

References

Imai, Kosuke, Ying Lu and Aaron Strauss. (Forthcoming) "eco: R Package for Ecological Inference in 2x2 Tables" Journal of Statistical Software, available at `http://imai.princeton.edu/research/eco.html`

Imai, Kosuke, Ying Lu and Aaron Strauss. (Forthcoming) "Bayesian and Likelihood Inference for 2 x 2 Ecological Tables: An Incomplete Data Approach" Political Analysis, available at `http://imai.princeton.edu/research/eiall.html`

See Also

`eco`, `ecoNP`

Examples

```
## load the registration data
data(reg)

## calculate the bounds
res <- ecoBD(Y ~ X, N = N, data = reg)
## print the results
print(res)
```

`ecoML` *Fitting Parametric Models and Quantifying Missing Information for Ecological Inference in 2x2 Tables*

Description

`ecoML` is used to fit parametric models for ecological inference in 2×2 tables via Expectation Maximization (EM) algorithms. The data is specified in proportions. At it's most basic setting, the algorithm assumes that the individual-level proportions (i.e., W_1 and W_2) and distributed bivariate normally (after logit transformations). The function calculates point estimates of the parameters for models based on different assumptions. The standard errors of the point estimates are also computed via Supplemented EM algorithms. Moreover, `ecoML` quantifies the amount of missing information associated with each parameter and allows researcher to examine the impact of missing information on parameter estimation in ecological inference. The models and algorithms are described in Imai, Lu and Strauss (Forthcoming).

Usage

```
ecoML(formula, data = parent.frame(), N = NULL, supplement = NULL,
      theta.start = c(0,0,1,1,0), fix.rho = FALSE,
      context = FALSE, sem = TRUE, epsilon = 10^(-10),
  maxit = 1000, loglik = TRUE, hyptest = FALSE, verbose = FALSE)
```

Arguments

`formula`	A symbolic description of the model to be fit, specifying the column and row margins of 2×2 ecological tables. `Y ~ X` specifies `Y` as the column margin (e.g., turnout) and `X` (e.g., percent African-American) as the row margin. Details and specific examples are given below.
`data`	An optional data frame in which to interpret the variables in `formula`. The default is the environment in which `ecoML` is called.

`N` An optional variable representing the size of the unit; e.g., the total number of voters.

`supplement` An optional matrix of supplemental data. The matrix has two columns, which contain additional individual-level data such as survey data for W_1 and W_2, respectively. If `NULL`, no additional individual-level data are included in the model. The default is `NULL`.

`fix.rho` Logical. If `TRUE`, the correlation (when `context=TRUE`) or the partial correlation (when `context=FALSE`) between W_1 and W_2 is fixed through the estimation. For details, see Imai, Lu and Strauss(2006). The default is `FALSE`.

`context` Logical. If `TRUE`, the contextual effect is also modeled. In this case, the row margin (i.e., X) and the individual-level rates (i.e., W_1 and W_2) are assumed to be distributed tri-variate normally (after logit transformations). See Imai, Lu and Strauss (2006) for details. The default is `FALSE`.

`sem` Logical. If `TRUE`, the standard errors of parameter estimates are estimated via SEM algorithm, as well as the fraction of missing data. The default is `TRUE`.

`theta.start` A numeric vector that specifies the starting values for the mean, variance, and covariance. When `context = FALSE`, the elements of `theta.start` correspond to ($E(W_1)$, $E(W_2)$, $var(W_1)$, $var(W_2)$, $cor(W_1, W_2)$). When `context = TRUE`, the elements of `theta.start` correspond to ($E(W_1)$, $E(W_2)$, $var(W_1)$, $var(W_2)$, $corr(W_1, X)$, $corr(W_2, X)$, $corr(W_1, W_2)$). Moreover, when `fix.rho=TRUE`, $corr(W_1, W_2)$ is set to be the correlation between W_1 and W_2 when `context = FALSE`, and the partial correlation between W_1 and W_2 given X when `context = FALSE`. The default is `c(0,0,1,1,0)`.

`epsilon` A positive number that specifies the convergence criterion for EM algorithm. The square root of `epsilon` is the convergence criterion for SEM algorithm. The default is `10^(-10)`.

`maxit` A positive integer specifies the maximum number of iterations before the convergence criterion is met. The default is `1000`.

`loglik` Logical. If `TRUE`, the value of the log-likelihood function at each iteration of EM is saved. The default is `TRUE`.

`hyptest` Logical. If `TRUE`, model is estimated under the null hypothesis that means of $W1$ and $W2$ are the same. The default is `FALSE`.

`verbose` Logical. If `TRUE`, the progress of the EM and SEM algorithms is printed to the screen. The default is `FALSE`.

Details

When `SEM` is `TRUE`, `ecoML` computes the observed-data information matrix for the parameters of interest based on Supplemented-EM algorithm. The inverse of the observed-data information matrix can be used to estimate the variance-covariance matrix for the parameters estimated from EM algorithms. In addition, it also computes the expected complete-data

information matrix. Based on these two measures, one can further calculate the fraction of missing information associated with each parameter. See Imai, Lu and Strauss (2006) for more details about fraction of missing information.

Moreover, when `hytest=TRUE`, `ecoML` allows to estimate the parametric model under the null hypothesis that `mu_1=mu_2`. One can then construct the likelihood ratio test to assess the hypothesis of equal means. The associated fraction of missing information for the test statistic can be also calculated. For details, see Imai, Lu and Strauss (2006) for details.

Value

An object of class `ecoML` containing the following elements:

`call` The matched call.

`X` The row margin, X.

`Y` The column margin, Y.

`N` The size of each table, N.

`context` The assumption under which model is estimated. If `context = FALSE`, CAR assumption is adopted and no contextual effect is modeled. If `context = TRUE`, NCAR assumption is adopted, and contextual effect is modeled.

`sem` Whether SEM algorithm is used to estimate the standard errors and observed information matrix for the parameter estimates.

`fix.rho` Whether the correlation or the partial correlation between W_1 an W_2 is fixed in the estimation.

`r12` If `fix.rho = TRUE`, the value that $corr(W_1, W_2)$ is fixed to.

`epsilon` The precision criterion for EM convergence. $\sqrt{\epsilon}$ is the precision criterion for SEM convergence.

`theta.sem` The ML estimates of $E(W_1)$,$E(W_2)$, $var(W_1)$,$var(W_2)$, and $cov(W_1, W_2)$. If `context = TRUE`, $E(X)$,$cov(W_1, X)$, $cov(W_2, X)$ are also reported.

`W` In-sample estimation of W_1 and W_2.

`suff.stat` The sufficient statistics for `theta.em`.

`iters.em` Number of EM iterations before convergence is achieved.

`iters.sem` Number of SEM iterations before convergence is achieved.

`loglik` The log-likelihood of the model when convergence is achieved.

`loglik.log.em`
A vector saving the value of the log-likelihood function at each iteration of the EM algorithm.

`mu.log.em` A matrix saving the unweighted mean estimation of the logit-transformed individual-level proportions (i.e., W_1 and W_2) at each iteration of the EM process.

`Sigma.log.em`
: A matrix saving the log of the variance estimation of the logit-transformed individual-level proportions (i.e., W_1 and W_2) at each iteration of EM process. Note, non-transformed variances are displayed on the screen (when `verbose = TRUE`).

`rho.fisher.em`
: A matrix saving the fisher transformation of the estimation of the correlations between the logit-transformed individual-level proportions (i.e., W_1 and W_2) at each iteration of EM process. Note, non-transformed correlations are displayed on the screen (when `verbose = TRUE`).

`DM`
: The matrix characterizing the rates of convergence of the EM algorithms. Such information is also used to calculate the observed-data information matrix

`Icom`
: The (expected) complete data information matrix estimated via SEM algorithm. When `context=FALSE, fix.rho=TRUE`, `Icom` is 4 by 4. When `context=FALSE, fix.rho=FALSE`, `Icom` is 5 by 5. When `context=TRUE`, `Icom` is 9 by 9.

`Iobs`
: The observed information matrix. The dimension of `Iobs` is same as `Icom`.

`Imiss`
: The difference between `Icom` and `Iobs`. The dimension of `Imiss` is same as `miss`.

`Vobs`
: The (symmetrized) variance-covariance matrix of the ML parameter estimates. The dimension of `Vobs` is same as `Icom`.

`Iobs`
: The (expected) complete-data variance-covariance matrix. The dimension of `Iobs` is same as `Icom`.

`Vobs.original`
: The estimated variance-covariance matrix of the ML parameter estimates. The dimension of `Vobs` is same as `Icom`.

`Fmis`
: The fraction of missing information associated with each parameter estimation.

`VFmis`
: The proportion of increased variance associated with each parameter estimation due to observed data.

`Ieigen`
: The largest eigen value of `Imiss`.

`Icom.trans`
: The complete data information matrix for the fisher transformed parameters.

`Iobs.trans`
: The observed data information matrix for the fisher transformed parameters.

`Fmis.trans`
: The fractions of missing information associated with the fisher transformed parameters.

Author(s)

Kosuke Imai, Department of Politics, Princeton University, ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Department of Sociology, University of Colorado at Boulder, ⟨ying.lu@Colorado.Edu⟩; Aaron Strauss, Department of Politics, Princeton University, ⟨abstraus@Princeton.Edu⟩.

References

Imai, Kosuke, Ying Lu and Aaron Strauss. (Forthcoming). "eco: R Package for Ecological Inference in 2x2 Tables" Journal of Statistical Software, available at `http://imai.princeton.edu/research/eco.html`

Imai, Kosuke, Ying Lu and Aaron Strauss. (Forthcoming). "Bayesian and Likelihood Inference for 2 x 2 Ecological Tables: An Incomplete Data Approach" Political Analysis, available at `http://imai.princeton.edu/research/eiall.html`

See Also

`eco`, `ecoNP`, `summary.ecoML`

Examples

```
## load the census data
data(census)

## NOTE: convergence has not been properly assessed for the following
## examples. See Imai, Lu and Strauss (2006) for more complete analyses.
## In the first example below, in the interest of time, only part of the
## data set is analyzed and the convergence requirement is less stringent
## than the default setting.

## In the second example, the program is arbitrarily halted 100 iterations
## into the simulation, before convergence.

## load the Robinson's census data
data(census)

## fit the parametric model with the default model specifications
## Not run: res <- ecoML(Y ~ X, data = census[1:100,],N=census[1:100,3],epsilon=10^(-6), verb
## summarize the results
## Not run: summary(res)

## obtain out-of-sample prediction
## Not run: out <- predict(res, verbose = TRUE)
## summarize the results
## Not run: summary(out)

## fit the parametric model with some individual
## level data using the default prior specification
surv <- 1:600
## Not run:
res1 <- ecoML(Y ~ X, context = TRUE, data = census[-surv,],
```

```
                    supplement = census[surv,c(4:5,1)], maxit=100, verbose = TRUE)
## End(Not run)
## summarize the results
## Not run: summary(res1)
```

`ecoNP` *Fitting the Nonparametric Bayesian Models of Ecological Inference in 2x2 Tables*

Description

`ecoNP` is used to fit the nonparametric Bayesian model (based on a Dirichlet process prior) for ecological inference in 2×2 tables via Markov chain Monte Carlo. It gives the in-sample predictions as well as out-of-sample predictions for population inference. The models and algorithms are described in Imai, Lu and Strauss (2008, Forthcoming).

Usage

```
ecoNP(formula, data = parent.frame(), N = NULL, supplement = NULL,
      context = FALSE, mu0 = 0, tau0 = 2, nu0 = 4, S0 = 10,
      alpha = NULL, a0 = 1, b0 = 0.1, parameter = FALSE,
      grid = FALSE, n.draws = 5000, burnin = 0, thin = 0,
      verbose = FALSE)
```

Arguments

`formula`	A symbolic description of the model to be fit, specifying the column and row margins of 2×2 ecological tables. `Y ~ X` specifies `Y` as the column margin (e.g., turnout) and `X` as the row margin (e.g., percent African-American). Details and specific examples are given below.
`data`	An optional data frame in which to interpret the variables in `formula`. The default is the environment in which `ecoNP` is called.
`N`	An optional variable representing the size of the unit; e.g., the total number of voters.
`supplement`	An optional matrix of supplemental data. The matrix has two columns, which contain additional individual-level data such as survey data for W_1 and W_2, respectively. If `NULL`, no additional individual-level data are included in the model. The default is `NULL`.
`context`	Logical. If `TRUE`, the contextual effect is also modeled, that is to assume the row margin X and the unknown W_1 and W_2 are correlated. See Imai, Lu and Strauss (2008, Forthcoming) for details. The default is `FALSE`.

`mu0` A scalar or a numeric vector that specifies the prior mean for the mean parameter μ of the base prior distribution G_0 (see Imai, Lu and Strauss (2008, Forthcoming) for detailed descriptions of Dirichlete prior and the normal base prior distribution) . If it is a scalar, then its value will be repeated to yield a vector of the length of μ, otherwise, it needs to be a vector of same length as μ. When `context=TRUE` , the length of μ is 3, otherwise it is 2. The default is `0`.

`tau0` A positive integer representing the scale parameter of the Normal-Inverse Wishart prior for the mean and variance parameter (μ_i, Σ_i) of each observation. The default is `2`.

`nu0` A positive integer representing the prior degrees of freedom of the variance matrix Σ_i. the default is `4`.

`S0` A positive scalar or a positive definite matrix that specifies the prior scale matrix for the variance matrix Σ_i. If it is a scalar, then the prior scale matrix will be a diagonal matrix with the same dimensions as Σ_i and the diagonal elements all take value of `S0`, otherwise `S0` needs to have same dimensions as Σ_i. When `context=TRUE`, Σ is a 3×3 matrix, otherwise, it is 2×2. The default is `10`.

`alpha` A positive scalar representing a user-specified fixed value of the concentration parameter, α. If `NULL`, α will be updated at each Gibbs draw, and its prior parameters `a0` and `b0` need to be specified. The default is `NULL`.

`a0` A positive integer representing the value of shape parameter of the gamma prior distribution for α. The default is `1`.

`b0` A positive integer representing the value of the scale parameter of the gamma prior distribution for α. The default is `0.1`.

`parameter` Logical. If `TRUE`, the Gibbs draws of the population parameters, μ and Σ, are returned in addition to the in-sample predictions of the missing internal cells, W. The default is `FALSE`. This needs to be set to `TRUE` if one wishes to make population inferences through `predict.eco`. See an example below.

`grid` Logical. If `TRUE`, the grid method is used to sample W in the Gibbs sampler. If `FALSE`, the Metropolis algorithm is used where candidate draws are sampled from the uniform distribution on the tomography line for each unit. Note that the grid method is significantly slower than the Metropolis algorithm.

`n.draws` A positive integer. The number of MCMC draws. The default is `5000`.

`burnin` A positive integer. The burnin interval for the Markov chain; i.e. the number of initial draws that should not be stored. The default is `0`.

`thin` A positive integer. The thinning interval for the Markov chain; i.e. the number of Gibbs draws between the recorded values that are skipped. The default is `0`.

`verbose` Logical. If `TRUE`, the progress of the Gibbs sampler is printed to the screen. The default is `FALSE`.

Value

An object of class `ecoNP` containing the following elements:

`call`	The matched call.
`X`	The row margin, X.
`Y`	The column margin, Y.
`burnin`	The number of initial burnin draws.
`thin`	The thinning interval.
`nu0`	The prior degrees of freedom.
`tau0`	The prior scale parameter.
`mu0`	The prior mean.
`S0`	The prior scale matrix.
`a0`	The prior shape parameter.
`b0`	The prior scale parameter.
`W`	A three dimensional array storing the posterior in-sample predictions of W. The first dimension indexes the Monte Carlo draws, the second dimension indexes the columns of the table, and the third dimension represents the observations.
`Wmin`	A numeric matrix storing the lower bounds of W.
`Wmax`	A numeric matrix storing the upper bounds of W.
`mu`	A three dimensional array storing the posterior draws of the population mean parameter, μ. The first dimension indexes the Monte Carlo draws, the second dimension indexes the columns of the table, and the third dimension represents the observations.
`Sigma`	A three dimensional array storing the posterior draws of the population variance matrix, Σ. The first dimension indexes the Monte Carlo draws, the second dimension indexes the parameters, and the third dimension represents the observations.
`alpha`	The posterior draws of α.
`nstar`	The number of clusters at each Gibbs draw.

Author(s)

Kosuke Imai, Department of Politics, Princeton University, ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Department of Sociology, University of Colorado at Boulder, ⟨ying.lu@Colorado.Edu⟩

References

Imai, Kosuke, Ying Lu and Aaron Strauss. (Forthcoming). "eco: R Package for Ecological Inference in 2x2 Tables" Journal of Statistical Software, available at `http://imai.princeton.edu/research/eco.html`

Imai, Kosuke, Ying Lu and Aaron Strauss. (2008). "Bayesian and Likelihood Inference for 2 x 2 Ecological Tables: An Incomplete Data Approach" Political Analysis, Vol. 16, No. 1 (Winter), pp. 41-69. available at `http://imai.princeton.edu/research/eiall.html`

See Also

`eco`, `ecoML`, `predict.eco`, `summary.ecoNP`

Examples

```
## load the registration data
data(reg)

## NOTE: We set the number of MCMC draws to be a very small number in
## the following examples; i.e., convergence has not been properly
## assessed. See Imai, Lu and Strauss (2006) for more complete examples.

## fit the nonparametric model to give in-sample predictions
## store the parameters to make population inference later
res <- ecoNP(Y ~ X, data = reg, n.draws = 50, param = TRUE, verbose = TRUE)

##summarize the results
summary(res)

## obtain out-of-sample prediction
out <- predict(res, verbose = TRUE)

## summarize the results
summary(out)

## density plots of the out-of-sample predictions
par(mfrow=c(2,1))
plot(density(out[,1]), main = "W1")
plot(density(out[,2]), main = "W2")

## load the Robinson's census data
data(census)

## fit the parametric model with contextual effects and N
## using the default prior specification
```

```
res1 <- ecoNP(Y ~ X, N = N, context = TRUE, param = TRUE, data = census,
n.draws = 25, verbose = TRUE)

## summarize the results
summary(res1)

## out-of sample prediction
pres1 <- predict(res1)
summary(pres1)
```

`predict.eco` *Out-of-Sample Posterior Prediction under the Parametric Bayesian Model for Ecological Inference in 2x2 Tables*

Description

Obtains out-of-sample posterior predictions under the fitted parametric Bayesian model for ecological inference. `predict` method for class `eco` and `ecoX`.

Usage

```
## S3 method for class 'eco':
predict(object, newdraw = NULL, subset = NULL,
                  verbose = FALSE, ...)
## S3 method for class 'ecoX':
predict(object, newdraw = NULL, subset = NULL,
                  newdata = NULL, cond = FALSE, verbose = FALSE, ...)
```

Arguments

`object`	An output object from `eco` or `ecoNP`.
`newdraw`	An optional list containing two matrices (or three dimensional arrays for the nonparametric model) of MCMC draws of μ and Σ. Those elements should be named as `mu` and `Sigma`, respectively. The default is the original MCMC draws stored in `object`.
`newdata`	An optional data frame containing a new data set for which posterior predictions will be made. The new data set must have the same variable names as those in the original data.
`subset`	A scalar or numerical vector specifying the row number(s) of `mu` and `Sigma` in the output object from `eco`. If specified, the posterior draws of parameters for those rows are used for posterior prediction. The default is `NULL` where all the posterior draws are used.

`cond`	logical. If `TRUE`, then the conditional prediction will made for the parametric model with contextual effects. The default is `FALSE`.
`verbose`	logical. If `TRUE`, helpful messages along with a progress report on the Monte Carlo sampling from the posterior predictive distributions are printed on the screen. The default is `FALSE`.
`...`	further arguments passed to or from other methods.

Details

The posterior predictive values are computed using the Monte Carlo sample stored in the `eco` output (or other sample if `newdraw` is specified). Given each Monte Carlo sample of the parameters, we sample the vector-valued latent variable from the appropriate multivariate Normal distribution. Then, we apply the inverse logit transformation to obtain the predictive values of proportions, W. The computation may be slow (especially for the nonparametric model) if a large Monte Carlo sample of the model parameters is used. In either case, setting `verbose = TRUE` may be helpful in monitoring the progress of the code.

Value

`predict.eco` yields a matrix of class `predict.eco` containing the Monte Carlo sample from the posterior predictive distribution of inner cells of ecological tables. `summary.predict.eco` will summarize the output, and `print.summary.predict.eco` will print the summary.

Author(s)

Kosuke Imai, Department of Politics, Princeton University, ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Department of Sociology, University of Colorado at Boulder, ⟨ying.lu@Colorado.Edu⟩

See Also

`eco`, `predict.ecoNP`

---

predict.ecoNP *Out-of-Sample Posterior Prediction under the Nonparametric Bayesian Model for Ecological Inference in 2x2 Tables*

Description

Obtains out-of-sample posterior predictions under the fitted nonparametric Bayesian model for ecological inference. `predict` method for class `ecoNP` and `ecoNPX`.

Usage

```
## S3 method for class 'ecoNP':
predict(object, newdraw = NULL, subset = NULL, obs = NULL,
                  verbose = FALSE, ...)
## S3 method for class 'ecoNPX':
predict(object, newdraw = NULL, subset = NULL, obs = NULL,
                  cond = FALSE, verbose = FALSE, ...)
```

Arguments

`object`	An output object from `ecoNP`.
`newdraw`	An optional list containing two matrices (or three dimensional arrays for the nonparametric model) of MCMC draws of μ and Σ. Those elements should be named as `mu` and `Sigma`, respectively. The default is the original MCMC draws stored in `object`.
`subset`	A scalar or numerical vector specifying the row number(s) of `mu` and `Sigma` in the output object from `eco`. If specified, the posterior draws of parameters for those rows are used for posterior prediction. The default is `NULL` where all the posterior draws are used.
`obs`	An integer or vector of integers specifying the observation number(s) whose posterior draws will be used for predictions. The default is `NULL` where all the observations in the data set are selected.
`cond`	logical. If `TRUE`, then the conditional prediction will made for the parametric model with contextual effects. The default is `FALSE`.
`verbose`	logical. If `TRUE`, helpful messages along with a progress report on the Monte Carlo sampling from the posterior predictive distributions are printed on the screen. The default is `FALSE`.
`...`	further arguments passed to or from other methods.

Details

The posterior predictive values are computed using the Monte Carlo sample stored in the `eco` or `ecoNP` output (or other sample if `newdraw` is specified). Given each Monte Carlo sample of the parameters, we sample the vector-valued latent variable from the appropriate multivariate Normal distribution. Then, we apply the inverse logit transformation to obtain the predictive values of proportions, W. The computation may be slow (especially for the nonparametric model) if a large Monte Carlo sample of the model parameters is used. In either case, setting `verbose = TRUE` may be helpful in monitoring the progress of the code.

Value

`predict.eco` yields a matrix of class `predict.eco` containing the Monte Carlo sample from the posterior predictive distribution of inner cells of ecological tables. `summary.predict.eco`

will summarize the output, and `print.summary.predict.eco` will print the summary.

Author(s)

Kosuke Imai, Department of Politics, Princeton University, ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Department of Sociology, University of Colorado at Boulder, ⟨ying.lu@Colorado.Edu⟩

See Also

`eco`, `ecoNP`, `summary.eco`, `summary.ecoNP`

`summary.eco` *Summarizing the Results for the Bayesian Parametric Model for Ecological Inference in 2x2 Tables*

Description

`summary` method for class `eco`.

Usage

```
## S3 method for class 'eco':
summary(object, CI = c(2.5, 97.5), param = TRUE,
        units = FALSE, subset = NULL, ...)

## S3 method for class 'summary.eco':
print(x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

`object`	An output object from `eco`.
`CI`	A vector of lower and upper bounds for the Bayesian credible intervals used to summarize the results. The default is the equal tail 95 percent credible interval.
`x`	An object of class `summary.eco`.
`digits`	the number of significant digits to use when printing.
`param`	Logical. If `TRUE`, the posterior estimates of the population parameters will be provided. The default value is `TRUE`.
`units`	Logical. If `TRUE`, the in-sample predictions for each unit or for a subset of units will be provided. The default value is `FALSE`.

`subset`	A numeric vector indicating the subset of the units whose in-sample predications to be provided when `units` is `TRUE`. The default value is `NULL` where the in-sample predictions for each unit will be provided.
`...`	further arguments passed to or from other methods.

Value

`summary.eco` yields an object of class `summary.eco` containing the following elements:

`call`	The call from `eco`.
`n.obs`	The number of units.
`n.draws`	The number of Monte Carlo samples.
`agg.table`	Aggregate posterior estimates of the marginal means of W_1 and W_2 using X and N as weights.
`param.table`	Posterior estimates of model parameters: population mean estimates of W_1 and W_2 and their logit transformations.
`W1.table`	Unit-level posterior estimates for W_1.
`W2.table`	Unit-level posterior estimates for W_2.

This object can be printed by `print.summary.eco`

Author(s)

Kosuke Imai, Department of Politics, Princeton University, ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Department of Sociology, University of Colorado at Boulder, ⟨ying.lu@Colorado.Edu⟩

See Also

`eco`, `predict.eco`

`summary.ecoNP` *Summarizing the Results for the Bayesian Nonparametric Model for Ecological Inference in 2x2 Tables*

Description

`summary` method for class `ecoNP`.

Usage

```
## S3 method for class 'ecoNP':
summary(object, CI = c(2.5, 97.5), param = FALSE,
        units = FALSE, subset = NULL, ...)

## S3 method for class 'summary.ecoNP':
print(x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

`object`	An output object from `ecoNP`.
`CI`	A vector of lower and upper bounds for the Bayesian credible intervals used to summarize the results. The default is the equal tail 95 percent credible interval.
`x`	An object of class `summary.ecoNP`.
`digits`	the number of significant digits to use when printing.
`param`	Logical. If `TRUE`, the posterior estimates of the population parameters will be provided. The default value is `FALSE`.
`units`	Logical. If `TRUE`, the in-sample predictions for each unit or for a subset of units will be provided. The default value is `FALSE`.
`subset`	A numeric vector indicating the subset of the units whose in-sample predications to be provided when `units` is `TRUE`. The default value is `NULL` where the in-sample predictions for each unit will be provided.
`...`	further arguments passed to or from other methods.

Value

`summary.ecoNP` yields an object of class `summary.ecoNP` containing the following elements:

`call`	The call from `ecoNP`.
`n.obs`	The number of units.
`n.draws`	The number of Monte Carlo samples.
`agg.table`	Aggregate posterior estimates of the marginal means of W_1 and W_2 using X and N as weights.
`param.table`	Posterior estimates of model parameters: population mean estimates of W_1 and W_2. If `subset` is specified, only a subset of the population parameters are included.
`W1.table`	Unit-level posterior estimates for W_1.
`W2.table`	Unit-level posterior estimates for W_2.

This object can be printed by `print.summary.ecoNP`

Author(s)

Kosuke Imai, Department of Politics, Princeton University, ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Department of Sociology, University of Colorado at Boulder, ⟨ying.lu@Colorado.Edu⟩

See Also

`ecoNP`, `predict.eco`

C Selected Data References

`reg` *Voter Registration in US Southern States*

Description

This data set contains the racial composition, the registration rate, the number of eligible voters as well as the actual observed racial registration rates for every county in four US southern states: Florida, Louisiana, North Carolina, and South Carolina.

Usage

```
data(reg)
```

Format

A data frame containing 5 variables and 275 observations

X	numeric	the fraction of Black voters
Y	numeric	the fraction of voters who registered themselves
N	numeric	the total number of voters in each county
W1	numeric	the actual fraction of Black voters who registered themselves
W2	numeric	the actual fraction of White voters who registered themselves

References

King, G. (1997). "A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data". Princeton University Press, Princeton, NJ.

census *Black Illiteracy Rates in 1910 US Census*

Description

This data set contains the proportion of the residents who are black, the proportion of those who can read, the total population as well as the actual black literacy rate and white literacy rate for 1040 counties in the US. The dataset was originally analyzed by Robinson (1950) at the state level. King (1997) recoded the 1910 census at county level. The data set only includes those who are older than 10 years of age.

Usage

```
data(census)
```

Format

A data frame containing 5 variables and 1040 observations

X	numeric	the proportion of Black residents in each county
Y	numeric	the overall literacy rates in each county
N	numeric	the total number of residents in each county
W1	numeric	the actual Black literacy rate
W2	numeric	the actual White literacy rate

References

Robinson, W.S. (1950). "Ecological Correlations and the Behavior of Individuals." *American Sociological Review*, vol. 15, pp.351-357.

King, G. (1997). "A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data". Princeton University Press, Princeton, NJ.