

# The pan Package

May 14, 2008

**Version** 0.2-6

**Date** 2008-5-14

**Title** Multiple imputation for multivariate panel or clustered data

**Author** Original by Joseph L. Schafer <jls@stat.psu.edu>.

**Maintainer** Jing hua Zhao <jinghua.zhao@mrc-epid.cam.ac.uk>

**Description** Multiple imputation for multivariate panel or clustered data

**License** file LICENSE

**URL** <http://www.stat.psu.edu/~jls/misoftwa.html>

## R topics documented:

ecme . . . . .	1
pan . . . . .	4
pan.bd . . . . .	6

<b>Index</b>	<b>8</b>
--------------	----------

---

ecme	<i>ECME algorithm for general linear mixed model, as described by Schafer (1997)</i>
------	--

---

## Description

Performs maximum-likelihood estimation for generalized linear mixed models. The model, which is typically applied to longitudinal or clustered responses, is

$$y_i = X_i \beta + Z_i b_i + e_i, \quad i=1, \dots, m,$$

where

$y_i$  = ( $n_i \times 1$ ) response vector for subject or cluster  $i$ ;

$X_i$  = ( $n_i \times p$ ) matrix of covariates;

$Z_i = (n_i \times q)$  matrix of covariates;

$\beta = (p \times 1)$  vector of coefficients common to the population (fixed effects);

$b_i = (q \times 1)$  vector of coefficients specific to subject or cluster  $i$  (random effects); and

$e_i = (n_i \times 1)$  vector of residual errors.

The vector  $b_i$  is assumed to be normally distributed with mean zero and unstructured covariance matrix  $\psi$ ,

$b_i \sim N(0, \psi)$  independently for  $i=1, \dots, m$ .

The residual vector  $e_i$  is assumed to be

$e_i \sim N(0, \sigma^2 V_i)$

where  $V_i$  is a known  $(n_i \times n_i)$  matrix. In most applications,  $V_i$  is the identity matrix.

### Usage

```
ecme(y, subj, occ, pred, xcol, zcol=NULL, vmax, start,
     maxits=1000, eps=0.0001, random.effects=F)
```

### Arguments

<code>y</code>	vector of responses. This is simply the individual $y_i$ vectors stacked upon one another. Each element of $y$ represents the observed response for a particular subject-occasion, or for a particular unit within a cluster.
<code>subj</code>	vector of same length as $y$ , giving the subject (or cluster) indicators $i$ for the elements of $y$ . For example, suppose that $y$ is in fact $c(y_1, y_2, y_3, y_4)$ where $\text{length}(y_1)=2$ , $\text{length}(y_2)=3$ , $\text{length}(y_3)=2$ , and $\text{length}(y_4)=7$ . Then <code>subj</code> should be $c(1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4)$ .
<code>occ</code>	vector of same length as $y$ indicating the "occasions" for the elements of $y$ . In a longitudinal dataset where each individual is measured on at most $n_{\max}$ distinct occasions, each element of $y$ corresponds to one subject-occasion, and the elements of <code>occ</code> should be coded as $1, 2, \dots, n_{\max}$ to indicate these occasion labels. (You should label the occasions as $1, 2, \dots, n_{\max}$ even if they are not equally spaced in time; the actual times of measurement will be incorporated into the matrix "pred" below.) In a clustered dataset, the elements of <code>occ</code> label the units within each cluster $i$ , using the labels $1, 2, \dots, n_i$ .
<code>pred</code>	matrix of covariates used to predict $y$ . The number of rows should be $\text{length}(y)$ . The first column will typically be constant (one), and the remaining columns correspond to other variables appearing in $X_i$ and $Z_i$ .
<code>xcol</code>	vector of integers indicating which columns of <code>pred</code> will be used in $X_i$ . That is, <code>pred[,xcol]</code> is the $X_i$ matrices (stacked upon one another).
<code>zcol</code>	vector of integers indicating which columns of <code>pred</code> will be used in $Z_i$ . That is, <code>pred[,zcol]</code> is the $Z_i$ matrices (stacked upon one another). If <code>zcol=NULL</code> then the model is assumed to have no random effects; in that case the parameters are estimated noniteratively by generalized least squares.

<code>vmax</code>	optional matrix of dimension $c(\max(\text{occ}), \max(\text{occ}))$ from which the $V_i$ matrices will be extracted. In a longitudinal dataset, <code>vmax</code> would represent the $V_i$ matrix for an individual with responses at all possible occasions $1, 2, \dots, n_{\max} = \max(\text{occ})$ ; for individuals with responses at only a subset of these occasions, the $V_i$ will be obtained by extracting the rows and columns of <code>vmax</code> for those occasions. If no <code>vmax</code> is specified by the user, an identity matrix is used. In most applications of this model one will want to have $V_i = \text{identity}$ , so most of the time this argument can be omitted.
<code>start</code>	optional starting values of the parameters. If this argument is not given then <code>ecme()</code> chooses its own starting values. This argument should be a list of three elements named "beta", "psi", and "sigma2". Note that "beta" should be a vector of the same length as "xcol", "psi" should be a matrix of dimension $c(\text{length}(\text{zcol}), \text{length}(\text{zcol}))$ , and "sigma2" should be a scalar. This argument has no effect if <code>zcol=NULL</code> .
<code>maxits</code>	maximum number of cycles of ECME to be performed. The algorithm runs to convergence or until "maxits" iterations, whichever comes first.
<code>eps</code>	convergence criterion. The algorithm is considered to have converged if the relative differences in all parameters from one iteration to the next are less than <code>eps</code> —that is, if $\text{all}(\text{abs}(\text{new-old}) < \text{eps} * \text{abs}(\text{old}))$ .
<code>random.effects</code>	if TRUE, returns empirical Bayes estimates of all the random effects $b_i$ ( $i=1, 2, \dots, m$ ) and their estimated covariance matrices.

## Value

a list containing estimates of beta, sigma2, psi, an estimated covariance matrix for beta, the number of iterations actually performed, an indicator of whether the algorithm converged, and a vector of loglikelihood values at each iteration. If `random.effects=T`, also returns a matrix of estimated random effects (`bhat`) for individuals and an array of corresponding covariance matrices.

<code>beta</code>	vector of same length as "xcol" containing estimated fixed effects.
<code>sigma2</code>	estimate of error variance sigma2.
<code>psi</code>	matrix of dimension $c(\text{length}(\text{zcol}), \text{length}(\text{zcol}))$ containing the estimated covariance matrix psi.
<code>converged</code>	T if the algorithm converged, F if it did not
<code>iter</code>	number of iterations actually performed. Will be equal to "maxits" if <code>converged=F</code> .
<code>loglik</code>	vector of length "iter" reporting the value of the loglikelihood at each iteration.
<code>cov.beta</code>	matrix of dimension $c(\text{length}(\text{xcol}), \text{length}(\text{xcol}))$ containing estimated variances and covariances for elements of "beta".
<code>bhat</code>	if <code>random.effects=T</code> , a matrix with $\text{length}(\text{zcol})$ rows and $m$ columns, where <code>bhat[,i]</code> is an empirical Bayes estimate of $b_i$ .
<code>cov.b</code>	if <code>random.effects=T</code> , an array of dimension $\text{length}(\text{zcol})$ by $\text{length}(\text{zcol})$ by $m$ , where <code>cov.b[,i]</code> is an empirical Bayes estimate of the covariance matrix associated with $b_i$ .

## References

Schafer, J.L. (1997) Imputation of missing covariates under a multivariate linear mixed model. Technical report, Dept. of Statistics, The Pennsylvania State University.

## Examples

```
## Not run:
For a detailed example, see the file "ecmeex.R" distributed
with this function.
## End(Not run)
```

---

pan

*Imputation of multivariate panel or cluster data*

---

## Description

Gibbs sampler for the multivariate linear mixed model with incomplete data described by Schafer (1997). This function will typically be used to produce multiple imputations of missing data values in multivariate panel data or clustered data. The underlying model is

$$y_i = X_i \beta + Z_i b_i + e_i, \quad i=1, \dots, m,$$

where

$y_i$  = ( $n_i \times r$ ) matrix of incomplete multivariate data for subject or cluster  $i$ ;

$X_i$  = ( $n_i \times p$ ) matrix of covariates;

$Z_i$  = ( $n_i \times q$ ) matrix of covariates;

$\beta$  = ( $p \times r$ ) matrix of coefficients common to the population (fixed effects);

$b_i$  = ( $q \times r$ ) matrix of coefficients specific to subject or cluster  $i$  (random effects); and

$e_i$  = ( $n_i \times r$ ) matrix of residual errors.

The matrix  $b_i$ , when stacked into a single column, is assumed to be normally distributed with mean zero and unstructured covariance matrix  $\psi$ , and the rows of  $e_i$  are assumed to be independently normal with mean zero and unstructured covariance matrix  $\sigma$ . Missing values may appear in  $y_i$  in any pattern.

In most applications of this model, the first columns of  $X_i$  and  $Z_i$  will be constant (one) and  $Z_i$  will contain a subset of the columns of  $X_i$ .

## Usage

```
pan(y, subj, pred, xcol, zcol, prior, seed, iter=1, start)
```

**Arguments**

<code>y</code>	matrix of responses. This is simply the individual $y_i$ matrices stacked upon one another. Each column of <code>y</code> corresponds to a response variable. Each row of <code>y</code> corresponds to a single subject-occasion, or to a single subject within a cluster. Missing values (NA) may occur in any pattern.
<code>subj</code>	vector of length <code>nrow(y)</code> giving the subject (or cluster) indicators $i$ for the rows of <code>y</code> . For example, suppose that <code>y</code> is in fact <code>rbind(y1,y2,y3,y4)</code> where <code>nrow(y1)=2</code> , <code>nrow(y2)=3</code> , <code>nrow(y3)=2</code> , and <code>nrow(y4)=7</code> . Then <code>subj</code> should be <code>c(1,1,2,2,2,3,3,4,4,4,4,4,4)</code> .
<code>pred</code>	matrix of covariates used to predict <code>y</code> . This should have the same number of rows as <code>y</code> . The first column will typically be constant (one), and the remaining columns correspond to other variables appearing in $X_i$ and $Z_i$ .
<code>xcol</code>	vector of integers indicating which columns of <code>pred</code> will be used in $X_i$ . That is, <code>pred[,xcol]</code> is the $X_i$ matrices (stacked upon one another).
<code>zcol</code>	vector of integers indicating which columns of <code>pred</code> will be used in $Z_i$ . That is, <code>pred[,zcol]</code> is the $Z_i$ matrices (stacked upon one another).
<code>prior</code>	a list with four components (whose names are <code>a</code> , <code>Binv</code> , <code>c</code> , and <code>Dinv</code> , respectively) specifying the hyperparameters of the prior distributions for $\psi$ and $\sigma$ . For information on how to specify and interpret these hyperparameters, see Schafer (1997) and the example command file "panex.R" distributed with this package. Note: This is a slight departure from the notation in Schafer (1997), where <code>a</code> and <code>Binv</code> were denoted by "nu1" and "Lambdainv1", and <code>c</code> and <code>Dinv</code> were "nu2" and "Lambdainv2".
<code>seed</code>	integer seed for initializing <code>pan()</code> 's internal random number generator. This argument should be a positive integer.
<code>iter</code>	total number of iterations or cycles of the Gibbs sampler to be carried out.
<code>start</code>	optional list of quantities to specify the initial state of the Gibbs sampler. This list has the same form as "last" (described below), one of the components returned by <code>pan()</code> . This argument allows the Gibbs sampler to be restarted from the final state of a previous run. If "start" is omitted then <code>pan()</code> chooses its own initial state.

**Details**

The Gibbs sampler algorithm used in `pan()` is described in detail by Schafer (1997).

**Value**

A list containing the following components. Note that when you are using `pan()` to produce multiple imputations, you will be primarily interested in the component "y" which contains the imputed data; the arrays "beta", "sigma", and "psi" will be used primarily for diagnostics (e.g. time-series plots) to assess the convergence behavior of the Gibbs sampler.

<code>beta</code>	array of dimension <code>c(length(xcol),ncol(y),iter) = (p x r x number of Gibbs cycles)</code> containing the simulated values of beta from all cycles. That is, <code>beta[,T]</code> is the $(p \times r)$ matrix of simulated fixed effects at cycle $T$ .
-------------------	--

sigma	array of dimension $c(ncol(y), ncol(y), iter) = (r \times r \times \text{number of Gibbs cycles})$ containing the simulated values of sigma from all cycles. That is, $\text{sigma}[., T]$ is the simulated version of the model's sigma at cycle T.
psi	array of dimension $c(\text{length}(zcol) * ncol(y), \text{length}(zcol) * ncol(y), iter) = (q * r \times q * r \times \text{number of Gibbs cycles})$ containing the simulated values of psi from all cycles. That is, $\text{psi}[., T]$ is the simulated version of the model's psi at cycle T.
y	matrix of imputed data from the final cycle of the Gibbs sampler. Identical to the input argument y except that the missing values (NA) have been replaced by imputed values. If "iter" has been set large enough (which can be determined by examining time-series plots, etc. of "beta", "sigma", and "psi") then this is a proper draw from the posterior predictive distribution of the complete data.
last	a list of four components characterizing the final state of the Gibbs sampler. The four components are: "beta", "sigma", "psi", and "y", which are the simulated values of the corresponding model quantities from the final cycle of Gibbs. This information is already contained in the other components returned by pan(); we are providing this list merely as a convenience, to allow the user to start future runs of the Gibbs sampler at this state.

### Note

This function assumes that the rows of y (and thus the rows of subj and pred) have been sorted by subject number. That is, we assume that  $\text{subj} = \text{sort}(\text{subj})$ ,  $y = y[\text{order}(\text{subj}), ]$ , and  $\text{pred} = \text{pred}[\text{order}(\text{subj}), ]$ . If the matrix y is created by stacking  $y_i$ ,  $i = 1, \dots, m$  then this will automatically be the case.

### References

Schafer, J.L. (1997) Imputation of missing covariates under a multivariate linear mixed model. Technical report, Dept. of Statistics, The Pennsylvania State University.

### Examples

```
## Not run:
For a detailed example, see the file "panex.R" distributed
with this function. Here is a simple example of how pan()
might be used to produce three imputations.

# run Gibbs for 1000 cycles
result <- pan(y, subj, pred, xcol, zcol, prior, seed=9565, iter=1000)
# first imputation
imp1 <- result$y
# another 1000 cycles
result <- pan(y, subj, pred, xcol, zcol, prior, seed=54324, iter=1000, start=result$last)
# second imputation
imp2 <- result$y
# another 1000 cycles
result <- pan(y, subj, pred, xcol, zcol, prior, seed=698212, iter=1000, start=result$last)
# third imputation
imp3 <- result$y
## End(Not run)
```

pan.bd

*Imputation of multivariate panel or cluster data***Description**

Implementation of pan() that restricts the covariance matrix for the random effects to be block-diagonal. This function is identical to pan() in every way except that psi is now characterized by a set of  $r$  matrices of dimension  $q \times q$ .

**Usage**

```
pan.bd(y, subj, pred, xcol, zcol, prior, seed, iter=1, start)
```

**Arguments**

y	See description for pan().
subj	See description for pan().
pred	See description for pan().
xcol	See description for pan().
zcol	See description for pan().
prior	Same as for pan() except that the hyperparameters for psi have new dimensions. The hyperparameter $c$ is now a vector of length $r$ , where $c[j]$ contains the prior degrees of freedom for the $j$ th block portion of psi ( $j=1, \dots, r$ ). The hyperparameter $Dinv$ is now an array of dimension $c(q,q,r)$ , where $Dinv[,j]$ contains the prior scale matrix for the $j$ th block portion of psi ( $j=1, \dots, r$ ).
seed	See description for pan().
iter	See description for pan().
start	See description for pan().

**Value**

A list with the same components as that from pan(), with two minor differences: the dimension of "psi" is now  $(q \times q \times r \times \text{"iter"})$ , and the dimension of "last\$psi" is now  $(q \times q \times r)$ .

# Index

## \*Topic **models**

ecme, [1](#)

pan, [4](#)

pan.bd, [6](#)

ecme, [1](#)

pan, [4](#)

pan.bd, [6](#)