Package ‘imputeFin’

July 11, 2020

Title  Imputation of Financial Time Series with Missing Values and/or Outliers

Version  0.1.1

Date  2020-07-11

Description  Missing values often occur in financial data due to a variety of reasons (errors in the collection process or in the processing stage, lack of asset liquidity, lack of reporting of funds, etc.). However, most data analysis methods expect complete data and cannot be employed with missing values. One convenient way to deal with this issue without having to redesign the data analysis method is to impute the missing values. This package provides an efficient way to impute the missing values based on modeling the time series with a random walk or an autoregressive (AR) model, convenient to model log-prices and log-volumes in financial data. In the current version, the imputation is univariate-based (so no asset correlation is used). In addition, outliers can be detected and removed.

The package is based on the paper:

Maintainer  Daniel P. Palomar <daniel.p.palomar@gmail.com>

URL  https://CRAN.R-project.org/package=imputeFin,
     https://github.com/dppalomar/imputeFin,
     https://www.danielppalomar.com,
     https://doi.org/10.1109/TSP.2019.2899816

BugReports  https://github.com/dppalomar/imputeFin/issues

License  GPL-3

Encoding  UTF-8

LazyData  true

RoxygenNote  7.1.1

Depends
**imputeFin-package**

**Imports**  MASS, zoo

**Suggests**  knitr, ggplot2, prettydoc, rmarkdown, R.rsp, testthat, xts

**VignetteBuilder**  knitr, rmarkdown, R.rsp

**NeedsCompilation**  no

**Author**  Daniel P. Palomar [cre, aut],
            Junyan Liu [aut]

**Repository**  CRAN

**Date/Publication**  2020-07-11 13:10:03 UTC

---

**R topics documented:**

- imputeFin-package .......................... 2
- fit_AR1_Gaussian .......................... 3
- fit_AR1_t ................................. 6
- impute_AR1_Gaussian ...................... 8
- impute_AR1_t .............................. 10
- plot_imputed .............................. 13
- ts_AR1_Gaussian ........................... 14
- ts_AR1_t ................................. 14

**Index** 16

---

**imputeFin-package**  *imputeFin: Imputation of Financial Time Series with Missing Values.*

**Description**

Missing values often occur in financial data due to a variety of reasons (errors in the collection process or in the processing stage, lack of asset liquidity, lack of reporting of funds, etc.). However, most data analysis methods expect complete data and cannot be employed with missing values. One convenient way to deal with this issue without having to redesign the data analysis method is to impute the missing values. This package provides an efficient way to impute the missing values based on modeling the time series with a random walk or an autoregressive (AR) model, convenient to model log-prices and log-volumes in financial data. In the current version, the imputation is univariate-based (so no asset correlation is used). In addition, outliers can be detected and removed.

**Functions**

- fit_AR1_Gaussian, impute_AR1_Gaussian, fit_AR1_t, impute_AR1_t, plot_imputed

**Data**

- ts_AR1_Gaussian, ts_AR1_t
fit_AR1_Gaussian

Help

For a quick help see the README file: GitHub-README.
For more details see the vignette: CRAN-vignette.

Author(s)

Junyan LIU and Daniel P. Palomar

References

J. Liu, S. Kumar, and D. P. Palomar, "Parameter estimation of heavy-tailed AR model with missing
April, 2019. <https://doi.org/10.1109/TSP.2019.2899816>

---

fit_AR1_Gaussian Fit Gaussian AR(1) model to time series with missing values and/or outliers

Description

Estimate the parameters of a univariate Gaussian AR(1) model to fit the given time series with
missing values and/or outliers. For multivariate time series, the function will perform a number of
individual univariate fittings without attempting to model the correlations among the time series. If
the time series does not contain missing values, the maximum likelihood (ML) estimation is done
in one shot. With missing values, the iterative EM algorithm is employed for the estimation until
converge is achieved.

Usage

```r
fit_AR1_Gaussian(
  y,
  random_walk = FALSE,
  zero_mean = FALSE,
  remove_outliers = FALSE,
  outlier_prob_th = 0.001,
  verbose = TRUE,
  return_iterates = FALSE,
  return_condMeanCov = FALSE,
  tol = 1e-08,
  maxiter = 100
)
```
Arguments

- **y**: Time series object coercible to either a numeric vector or numeric matrix (e.g., zoo or xts) with missing values denoted by NA.
- **random_walk**: Logical value indicating if the time series is assumed to be a random walk so that \( \phi_1 = 1 \) (default is FALSE).
- **zero_mean**: Logical value indicating if the time series is assumed zero-mean so that \( \phi_0 = 0 \) (default is FALSE).
- **remove_outliers**: Logical value indicating whether to detect and remove outliers.
- **outlier_prob_th**: Threshold of probability of observation to declare an outlier (default is \( 1e^{-3} \)).
- **verbose**: Logical value indicating whether to output messages (default is TRUE).
- **return_iterates**: Logical value indicating if the iterates are to be returned (default is FALSE).
- **return_condMeanCov**: Logical value indicating if the conditional mean and covariance matrix of the time series (excluding the leading and trailing missing values) given the observed data are to be returned (default is FALSE).
- **tol**: Positive number denoting the relative tolerance used as stopping criterion (default is \( 1e^{-8} \)).
- **maxiter**: Positive integer indicating the maximum number of iterations allowed (default is 100).

Value

If the argument `y` is a univariate time series (i.e., coercible to a numeric vector), then this function will return a list with the following elements:

- **phi0**: The estimate for \( \phi_0 \) (real number).
- **phi1**: The estimate for \( \phi_1 \) (real number).
- **sigma2**: The estimate for \( \sigma^2 \) (positive number).
- **phi0_iterates**: Numeric vector with the estimates for \( \phi_0 \) at each iteration (returned only when `return_iterates = TRUE`).
- **phi1_iterates**: Numeric vector with the estimates for \( \phi_1 \) at each iteration (returned only when `return_iterates = TRUE`).
- **sigma2_iterates**: Numeric vector with the estimates for \( \sigma^2 \) at each iteration (returned only when `return_iterates = TRUE`).
- **f_iterates**: Numeric vector with the objective values at each iteration (returned only when `return_iterates = TRUE`).
- **cond_mean_y**: Numeric vector (of same length as argument `y`) with the conditional mean of the time series (excluding the leading and trailing missing values) given the observed data (returned only when `return_condMeanCov = TRUE`).
fit_AR1_Gaussian

cond_cov_y  Numeric matrix (with number of columns/rows equal to the length of the argument y) with the conditional covariance matrix of the time series (excluding the leading and trailing missing values) given the observed data (returned only when return_condMeanCov = TRUE).

index_miss  Indices of missing values imputed.

index_outliers  Indices of outliers detected/corrected.

If the argument y is a multivariate time series (i.e., with multiple columns and coercible to a numeric matrix), then this function will return a list with each element as in the case of univariate y corresponding to each of the columns (i.e., one list element per column of y), with the following additional elements that combine the estimated values in a convenient vector form:

phi0_vct  Numeric vector (with length equal to the number of columns of y) with the estimates for phi0 for each of the univariate time series.

phi1_vct  Numeric vector (with length equal to the number of columns of y) with the estimates for phi1 for each of the univariate time series.

sigma2_vct  Numeric vector (with length equal to the number of columns of y) with the estimates for sigma2 for each of the univariate time series.

Author(s)

Junyan Liu and Daniel P. Palomar

References


See Also

impute_AR1_Gaussian, fit_AR1_t

Examples

library(imputeFin)
data(ts_AR1_Gaussian)
y_missing <- ts_AR1_Gaussian$y_missing
fitted <- fit_AR1_Gaussian(y_missing)
fit_AR1_t  

Fit Student's $t$ AR(1) model to time series with missing values and/or outliers

Description

Estimate the parameters of a univariate Student’s $t$ AR(1) model to fit the given time series with missing values and/or outliers. For multivariate time series, the function will perform a number of individual univariate fittings without attempting to model the correlations among the time series. If the time series does not contain missing values, the maximum likelihood (ML) estimation is done via the iterative EM algorithm until converge is achieved. With missing values, the stochastic EM algorithm is employed for the estimation (currently the maximum number of iterations will be executed without attempting to check early converge).

Usage

```r
fit_AR1_t(
  y,
  random_walk = FALSE,
  zero_mean = FALSE,
  fast_and_heuristic = TRUE,
  remove_outliers = FALSE,
  outlier_prob_th = 0.001,
  verbose = TRUE,
  return_iterates = FALSE,
  return_condMean_Gaussian = FALSE,
  tol = 1e-08,
  maxiter = 100,
  n_chain = 10,
  n_thin = 1,
  K = 30
)
```

Arguments

- **y**  
  Time series object coercible to either a numeric vector or numeric matrix (e.g., `zoo` or `xts`) with missing values denoted by `NA`.

- **random_walk**  
  Logical value indicating if the time series is assumed to be a random walk so that $\phi_1 = 1$ (default is `FALSE`).

- **zero_mean**  
  Logical value indicating if the time series is assumed zero-mean so that $\phi_0 = 0$ (default is `FALSE`).

- **fast_and_heuristic**  
  Logical value indicating whether a heuristic but fast method is to be used to estimate the parameters of the Student’s $t$ AR(1) model (default is `TRUE`).

- **remove_outliers**  
  Logical value indicating whether to detect and remove outliers.
outlier_prob_th
Threshold of probability of observation to declare an outlier (default is 1e-3).

verbose
Logical value indicating whether to output messages (default is TRUE).

return_iterates
Logical value indicating if the iterates are to be returned (default is FALSE).

return_condMean_Gaussian
Logical value indicating if the conditional mean and covariance matrix of the
time series (excluding the leading and trailing missing values) given the ob-
served data are to be returned (default is FALSE).

tol
Positive number denoting the relative tolerance used as stopping criterion (de-
default is 1e-8).

maxiter
Positive integer indicating the maximum number of iterations allowed (default
is 100).

n_chain
Positive integer indicating the number of the parallel Markov chains in the
stochastic EM method (default is 10).

n_thin
Positive integer indicating the sampling period of the Gibbs sampling in the
stochastic EM method (default is 1). Every n_thin-th samples is used. This is
aimed to reduce the dependence of the samples.

K
Positive number controlling the values of the step sizes in the stochastic EM
method (default is 30).

Value
If the argument y is a univariate time series (i.e., coercible to a numeric vector), then this function
will return a list with the following elements:

phi0
The estimate for phi0 (real number).

phi1
The estimate for phi1 (real number).

sigma2
The estimate for sigma^2 (positive number).

nu
The estimate for nu (positive number).

phi0_iterates
Numeric vector with the estimates for phi0 at each iteration (returned only when
return_iterates = TRUE).

phi1_iterates
Numeric vector with the estimates for phi1 at each iteration (returned only when
return_iterates = TRUE).

sigma2_iterates
Numeric vector with the estimates for sigma^2 at each iteration (returned only when
return_iterates = TRUE).

nu_iterate
Numeric vector with the estimates for nu at each iteration (returned only when
return_iterates = TRUE).

f_iterates
Numeric vector with the objective values at each iteration (returned only when
return_iterates = TRUE).

cond_mean_y_Gaussian
Numeric vector (of same length as argument y) with the conditional mean of the
time series (excluding the missing values at the head and tail) given the observed
data based on Gaussian AR(1) model (returned only when return_condMean_Gaussian
= TRUE).
index_miss  Indices of missing values imputed.
index_outliers  Indices of outliers detected/corrected.

If the argument \( y \) is a multivariate time series (i.e., with multiple columns and coercible to a numeric matrix), then this function will return a list with each element as in the case of univariate \( y \) corresponding to each of the columns (i.e., one list element per column of \( y \)), with the following additional elements that combine the estimated values in a convenient vector form:

- \( \phi_0 \_vct \): Numeric vector (with length equal to the number of columns of \( y \)) with the estimates for \( \phi_0 \) for each of the univariate time series.
- \( \phi_1 \_vct \): Numeric vector (with length equal to the number of columns of \( y \)) with the estimates for \( \phi_1 \) for each of the univariate time series.
- \( \sigma^2 \_vct \): Numeric vector (with length equal to the number of columns of \( y \)) with the estimates for \( \sigma^2 \) for each of the univariate time series.
- \( \nu \_vct \): Numeric vector (with length equal to the number of columns of \( y \)) with the estimates for \( \nu \) for each of the univariate time series.

**Author(s)**

Junyan Liu and Daniel P. Palomar

**References**


**See Also**

`impute_AR1_t, fit_AR1_Gaussian`

**Examples**

```r
library(imputeFin)
data(ts_AR1_t)
y_missing <- ts_AR1_t$y_missing
fitted <- fit_AR1_t(y_missing)
```

---

**impute_AR1_Gaussian**  
**Impute missing values of time series based on a Gaussian AR(1) model**

**Description**

Impute inner missing values (excluding leading and trailing ones) of time series by drawing samples from the conditional distribution of the missing values given the observed data based on a Gaussian AR(1) model as estimated with the function `fit_AR1_Gaussian`. Outliers can be detected and removed.
Usage

```r
impute_AR1_Gaussian(
  y,
  n_samples = 1,
  random_walk = FALSE,
  zero_mean = FALSE,
  remove_outliers = FALSE,
  outlier_prob_th = 0.001,
  verbose = TRUE,
  return_estimates = FALSE,
  tol = 1e-10,
  maxiter = 100
)
```

Arguments

- **y**: Time series object coercible to either a numeric vector or numeric matrix (e.g., `zoo` or `xts`) with missing values denoted by `NA`.
- **n_samples**: Positive integer indicating the number of imputations (default is 1).
- **random_walk**: Logical value indicating if the time series is assumed to be a random walk so that \( \phi_1 = 1 \) (default is `FALSE`).
- **zero_mean**: Logical value indicating if the time series is assumed zero-mean so that \( \phi_0 = 0 \) (default is `FALSE`).
- **remove_outliers**: Logical value indicating whether to detect and remove outliers.
- **outlier_prob_th**: Threshold of probability of observation to declare an outlier (default is `1e-3`).
- **verbose**: Logical value indicating whether to output messages (default is `TRUE`).
- **return_estimates**: Logical value indicating if the estimates of the model parameters are to be returned (default is `FALSE`).
- **tol**: Positive number denoting the relative tolerance used as stopping criterion (default is `1e-8`).
- **maxiter**: Positive integer indicating the maximum number of iterations allowed (default is 100).

Value

By default (i.e., for `n_samples = 1` and `return_estimates = FALSE`), the function will return an imputed time series of the same class and dimensions as the argument `y` with one new attribute recording the locations of missing values (the function `plot_imputed` will make use of such information to indicate the imputed values), as well as locations of outliers removed.

If `n_samples > 1`, the function will return a list consisting of `n_samples` imputed time series with names: `y_imputed.1`, `y_imputed.2`, etc.

If `return_estimates = TRUE`, in addition to the imputed time series `y_imputed`, the function will return the estimated model parameters:
impute_AR1_t

phi0  The estimate for phi0 (numeric scalar or vector depending on the number of time series).

phi1  The estimate for phi1 (numeric scalar or vector depending on the number of time series).

sigma2  The estimate for sigma2 (numeric scalar or vector depending on the number of time series).

Author(s)

Junyan Liu and Daniel P. Palomar

References


See Also

plot_imputed, fit_AR1_Gaussian, impute_AR1_t

Examples

library(imputeFin)
data(ts_AR1_Gaussian)
y_missing <- ts_AR1_Gaussian$y_missing
y_imputed <- impute_AR1_Gaussian(y_missing)
plot_imputed(y_imputed)

| impute_AR1_t | Impute missing values of time series based on a Student’s t AR(1) model |

Description

Impute inner missing values (excluding leading and trailing ones) of time series by drawing samples from the conditional distribution of the missing values given the observed data based on a Student’s t AR(1) model as estimated with the function fit_AR1_t. Outliers can be detected and removed.
Usage

```r
impute_AR1_t(
  y,
  n_samples = 1,
  random_walk = FALSE,
  zero_mean = FALSE,
  fast_and_heuristic = TRUE,
  remove_outliers = FALSE,
  outlier_prob_th = 0.001,
  verbose = TRUE,
  return_estimates = FALSE,
  tol = 1e-08,
  maxiter = 100,
  K = 30,
  n_burn = 100,
  n_thin = 50
)
```

Arguments

- **y**: Time series object coercible to either a numeric vector or numeric matrix (e.g., `zoo` or `xts`) with missing values denoted by `NA`.
- **n_samples**: Positive integer indicating the number of imputations (default is 1).
- **random_walk**: Logical value indicating if the time series is assumed to be a random walk so that \( \phi_1 = 1 \) (default is `FALSE`).
- **zero_mean**: Logical value indicating if the time series is assumed zero-mean so that \( \phi_0 = 0 \) (default is `FALSE`).
- **fast_and_heuristic**: Logical value indicating whether a heuristic but fast method is to be used to estimate the parameters of the Student’s t AR(1) model (default is `TRUE`).
- **remove_outliers**: Logical value indicating whether to detect and remove outliers.
- **outlier_prob_th**: Threshold of probability of observation to declare an outlier (default is `1e-3`).
- **verbose**: Logical value indicating whether to output messages (default is `TRUE`).
- **return_estimates**: Logical value indicating if the estimates of the model parameters are to be returned (default is `FALSE`).
- **tol**: Positive number denoting the relative tolerance used as stopping criterion (default is `1e-8`).
- **maxiter**: Positive integer indicating the maximum number of iterations allowed (default is `100`).
- **K**: Positive number controlling the values of the step sizes in the stochastic EM method (default is `30`).
n_burn  Positive integer controlling the length of the burn-in period of the Gibbs sampling (default is 100). The first \((n\_burn \times n\_thin)\) samples generated will be ignored.

n_thin  Positive integer indicating the sampling period of the Gibbs sampling in the stochastic EM method (default is 1). Every \(n\_thin\)-th samples is used. This is aimed to reduce the dependence of the samples.

Value

By default (i.e., for \(n\_samples = 1\) and \(return\_estimates = FALSE\)), the function will return an imputed time series of the same class and dimensions as the argument \(y\) with one new attribute recording the locations of missing values (the function \texttt{plot_imputed} will make use of such information to indicate the imputed values), as well as locations of outliers removed.

If \(n\_samples > 1\), the function will return a list consisting of \(n\_sample\) imputed time series with names: \(y\_imputed.1, y\_imputed.2, \ldots\).

If \(return\_estimates = TRUE\), in addition to the imputed time series \(y\_imputed\), the function will return the estimated model parameters:

- \(\phi_0\): The estimate for \(\phi_0\) (numeric scalar or vector depending on the number of time series).
- \(\phi_1\): The estimate for \(\phi_1\) (numeric scalar or vector depending on the number of time series).
- \(\sigma^2\): The estimate for \(\sigma^2\) (numeric scalar or vector depending on the number of time series).
- \(\nu\): The estimate for \(\nu\) (numeric scalar or vector depending on the number of time series).

Author(s)

Junyan Liu and Daniel P. Palomar

References


See Also

\texttt{plot_imputed, fit_AR1_t, impute_AR1_Gaussian}

Examples

```r
library(imputeFin)
data(ts_AR1_t)
y_missing <- ts_AR1_t$y_missing
y_imputed <- impute_AR1_t(y_missing)
plot_imputed(y_imputed)
```
plot_imputed

Description
Plot single imputed time series (as returned by functions `impute_AR1_Gaussian` and `impute_AR1_t`), highlighting the imputed values in a different color.

Usage
```r
plot_imputed(
  y_imputed,
  column = 1,
  title = "Imputed time series",
  color_imputed = "red",
  type = c("ggplot2", "simple")
)
```

Arguments
- **y_imputed**: Imputed time series (can be any object coercible to a numeric vector or a numeric matrix). If it has the attribute "index_miss" (as returned by any of the imputation functions `impute_AR1_Gaussian` and `impute_AR1_t`), then it will highlight the imputed values in a different color.
- **column**: Positive integer indicating the column index to be plotted (only valid if the argument `y_imputed` is coercible to a matrix with more than one column). Default is 1.
- **title**: Title of the plot (default is "Imputed time series").
- **color_imputed**: Color for the imputed values (default is "red").
- **type**: Type of plot. Valid options: "ggplot2" and "simple". Default is "ggplot2" (the package ggplot2 must be installed).

Author(s)
Daniel P. Palomar

Examples
```r
library(imputeFin)
data(ts_AR1_t)
y_missing <- ts_AR1_t$y_missing
y_imputed <- impute_AR1_t(y_missing)
plot_imputed(y_missing, title = "Original time series with missing values")
plot_imputed(y_imputed)
```
ts_AR1_Gaussian  \textit{Synthetic AR(1) Gaussian time series with missing values}

\textbf{Description}

Synthetic AR(1) Gaussian time series with missing values for estimation and imputation testing purposes.

\textbf{Usage}

data(ts_AR1_Gaussian)

\textbf{Format}

List with the following elements:

- \textbf{y\_missing} 300 x 3 zoo object with three AR(1) Gaussian time series along the columns: the first column contains a time series with 10\% consecutive missing values; the second column contains a time series with 10\% missing values randomly distributed; and the third column contains the union of the previous missing values.
- \textbf{phi0} Value of phι0 used to generate the time series.
- \textbf{phi1} Value of phι1 used to generate the time series.
- \textbf{sigma2} Value of sigma2 used to generate the time series.

\begin{center}
\begin{tabular}{ll}
\hline
\textbf{ts\_AR1\_t} & \textit{Synthetic AR(1) Student's t time series with missing values} \\
\hline
\end{tabular}
\end{center}

\textbf{Description}

Synthetic AR(1) Student's t time series with missing values for estimation and imputation testing purposes.

\textbf{Usage}

data(ts\_AR1\_t)

\textbf{Format}

List with the following elements:

- \textbf{y\_missing} 300 x 3 zoo object with three AR(1) Student's t time series along the columns: the first column contains a time series with 10\% consecutive missing values; the second column contains a time series with 10\% missing values randomly distributed; and the third column contains the union of the previous missing values.
- \textbf{phi0} Value of phι0 used to generate the time series.
\( \text{ts}\_\text{ARI}\_t \)

- phi1: Value of phi1 used to generate the time series.
- sigma2: Value of sigma2 used to generate the time series.
- nu: Value of nu used to generate the time series.
Index

* dataset
  ts_AR1_Gaussian, 14
  ts_AR1_t, 14

fit_AR1_Gaussian, 2, 3, 8, 10
fit_AR1_t, 2, 5, 6, 10, 12

impute_AR1_Gaussian, 2, 5, 8, 12, 13
impute_AR1_t, 2, 8, 10, 10, 13
imputeFin-package, 2

plot_imputed, 2, 9, 10, 12, 13

ts_AR1_Gaussian, 2, 14
ts_AR1_t, 2, 14