cppEDM Version 1.15.1 October 27, 2023

cppEDM is a C++ implementation of empirical dynamic modeling (EDM) algorithms. It is designed as an application programming interface (API) to functions in the libEDM.a library. A Python interface is provided in PyPI package pyEDM, an R interface in CRAN package rEDM. A Jupyter notebook GUI front-end is provided in jpyEDM.

Table of Contents

University of California at San Diego
Scripps Institution of Oceanography
Sugihara Lab

Joseph Park, Cameron Smith

Introduction

cppEDM is a C++ implementation of empirical dynamic modeling (EDM) algorithms. Primary algorithms are listed in Table 1. The core code is an object oriented implementation supporting application programming interface (API) functions accepting parameters and returning data objects. EDM functions are accessed from a user-compiled library created from C++ source files and a unix-like compiler supporting the C++11 standard. cppEDM has Python and R interfaces implemented in the pyEDM and rEDM packages.

Algorithm	API Interface	Reference
Simplex projection	`Simplex()`	Sugihara and May (1990)
Sequential Locally Weighted Global Linear Maps (S-map)	`SMap()`	Sugihara (1994)
Predictions from multivariate embeddings	`Simplex(), SMap()`	Dixon et. al. (1999)
Convergent cross mapping	`CCM()`	Sugihara et. al. (2012)
Multiview embedding	`Multiview()`	Ye and Sugihara (2016)

Convenience functions to prepare and evaluate data are listed in Table 2.

Function	Purpose	Parameter Range
`Embed()`	Timeseries delay dimensional embedding	User defined
`EmbedDimension()`	Evaluate prediction skill vs. embedding dimension	E = [1, 10]
`PredictInterval()`	Evaluate prediction skill vs. forecast interval	Tp = [1, 10]
`PredictNonlinear()`	Evaluate prediction skill vs. SMap nonlinear localisation	θ = 0.01, 0.1, 0.3, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9
`ComputeError()`	Pearson correlation, MAE, RMSE	

Installation

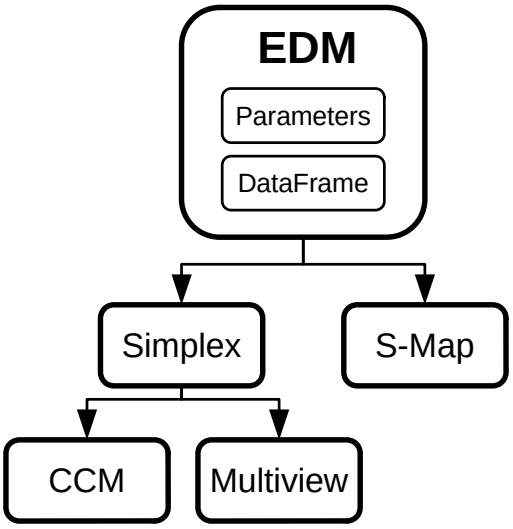
cppEDM is available at github.com/SugiharaLab/cppEDM.

cppEDM requires a C++14 standard compiler, and the LAPACK library. The libEDM.a library can be built by running "`make`" in the cppEDM/src/ directory. This copies libEDM.a into the cppEDM/lib/ directory, where it can be linked to user applications.

Once libEDM.a is built, there are a series of test applications in the cppEDM/tests/ directory. The applications can be built with the "`make`" command, and executed at the command line. API examples can also be found in cppEDM/etc/Test.cc.

Class Objects

cppEDM is a C++ Object Oriented (OO) implementation. The primary data class is `EDM`, from which the algorithm classes are derived. The class hierarchy is:




```
          EDM Methods

void GetTarget();
void EmbedData();
void PrepareEmbedding();
void Distances();
void FindNeighbors();
void Project(); // Simplex, Smap...
void CheckDataRows();
void RemovePartialData();
void FormatOutput();
void FillTimes();

            API
Embed()           Simplex()
CCM()             Multiview()
SMap()            EmbedDimension()
PredictInterval() PredictNonlinear()
```

Two C++ class objects are used for data access and parameter coordination, the `DataFrame` and `Parameters` classes, described below.

DataFrame

The `DataFrame` class is the fundamental data object of cppEDM. It stores data in a contiguous block of memory using the C++ `valarray` type in a row-major format.

A `DataFrame` can be initialised with data from a csv file by calling the `DataFrame` constructor with `path` and `fileName` parameters. Data input files are assumed to be in csv format. The files are required to have a single line header with column names.

It is assumed that the first column of the csv file is a vector of times or time indices for each observation (row). All subsequent columns are expected to be numeric. However, a `DataFrame` can be created with the `noTime = true` parameter to read in numeric data with no time in the first column. Such data frames should not be passed to prediction functions.

The `WriteData(path, file)` class method can be called explicitly to write data to a csv format file. If the `DataFrame` does not have column names, then column names are created as V1, V2...

Primary `DataFrame` access functions are listed in table 3.

DataFrame Method	Parameters	Type	Purpose
(row, column)	size_t row size_t column	double or int	Access data element
DataFrame(path, file)	string path string fileName	DataFrame<double>	Create DataFrame from csv file
WriteData(path, file)	string outputFilePath string outputFileName		Write DataFrame to file
Elements()		valarray	Access data valarray
NColumns()		size_t	Get number of columns
NRows()		size_t	Get number of rows
size()		size_t	Get number of elements
ColumnNames()		vector< string >	Access column names
ColumnNameToIndex()		map<string, size_t>	Access column name to index map
MaxRowPrint()		size_t	Access maximum number of rows to ostream
Column(col)	size_t col	valarray	Get data vector at column
Row(row)	size_t row	valarray	Get data vector at row
VectorColumnName(column)	string column	valarray	Get data vector at column with name
ColumnMajorData()		valarray	Elements() in column major format.
DataFrameFromColumnIndex (columns)	vector<size_t> columns	DataFrame<double>	Get DataFrame subset from column indices
DataFrameFromColumnNames (columns)	vector<string> columns	DataFrame<double>	Get DataFrame subset from column names
DataFrameFromRowIndex (rows)	vector<size_t> rows	DataFrame<double>	Get DataFrame subset from row indices
WriteRow(row, array)	size_t row std::valarray<T> array		Write valarray to row
WriteColumn(col, array)	size_t col valarray<T> array		Write valarray to column

Parameters

The `Parameters` class is used to store and access API function parameters in a unified object. Generally this is an internal object that does not need to be instantiated, accessed or dynamically modified. API parameter names and purpose are listed in table 4.

Parameter	Type	Default	Purpose
`pathIn`	`string`	`"./"`	Input data file path
`dataFile`	`string`	`""`	Data file name
`pathOut`	`string`	`"./"`	Output file path
`predictFile`	`string`	`""`	Prediction output file
`smapCoefFile`	`string`	`""`	SMap coefficient output file
`smapSVFile`	`string`	`""`	SMap singular values output file
`lib`	`string`	`""`	library start : stop row indices
`pred`	`string`	`""`	prediction start : stop row indices
`E`	`int`	`0`	Embedding dimension
`Tp`	`int`	`0 or 1`	Prediction interval (rows)
`knn`	`int`	`0`	Number nearest neighbors
`tau`	`int`	`-1`	Embedding offset (time series rows)
`theta`	`double`	`0`	SMap localisation
`exclusionRadius`	`int`	`0`	Prediction vector exclusion row radius
`columns`	`string`	`""`	Column names or indices for prediction
`target`	`string`	`""`	Target library column name or index
`embedded`	`bool`	`false`	Is data an embedding?
`const_pred`	`bool`	`false`	Include non-projected forecast data
`ignoreNan`	`bool`	`true`	SMap detect and remove nan from lib
`verbose`	`bool`	`false`	Echo messages
`validLib`	`vector<bool>`	`[]`	Conditional Embedding (CE)
`generateSteps`	`int`	`0`	Simplex, SMap feedback prediction
`generateLibrary`	`bool`	`false`	Increment EDM library with feedback
`parameterList`	`bool`	`false`	Return Parameters map
`multiview`	`int`	`0`	Number of ensembles, 0 = sqrt(N)
`D`	`int`	`0`	Multiview dimension
`trainLib`	`bool`	`true`	Multiview use lib as training library
`excludeTarget`	`bool`	`false`	Multiview exlcude target from combos
`libSizes`	`string`	`""`	CCM library sizes
`sample`	`int`	`0`	CCM number of random samples
`random`	`bool`	`true`	CCM use random samples?
`replacement`	`bool`	`false`	CCM sample with replacement?
`includeData`	`bool`	`false`	CCM include all projections in return
`seed`	`unsigned`	`0`	CCM RNG seed, 0 = random seed

Application Programming Interface (API)

Embed

Create a data block of Takens (1981) time-delay embedding from each of the columns in the csv file or `DataFrame`. The `columns` parameter can be a list of column names, or a list of column indices. If `columns` is a list of indices, then column names are created as V1, V2...

Note: The returned `DataFrame` will have `|tau|*(E-1)` fewer rows than the input data from the removal of partial vectors as a result of the embedding.

Note: The returned `DataFrame` does not have the time column.

```
//------------------------------------------------------------
// Overload 1: Explicit data file path/name
//------------------------------------------------------------
DataFrame< double > Embed ( std::string path     = "",
                            std::string dataFile = "",
                            int         E        = 0,
                            int         tau      = -1,
                            std::string columns  = "",
                            bool        verbose  = false );



//------------------------------------------------------------
// Overload 2: DataFrame provided
//------------------------------------------------------------
DataFrame< double > Embed ( DataFrame< double > dataFrame,
                            int                 E       = 0,
                            int                 tau     = -1,
                            std::string         columns = "",
                            bool                verbose = false );



//------------------------------------------------------------
// Called from Embed to create the time-delay embedding
//------------------------------------------------------------
DataFrame< double > MakeBlock ( DataFrame< double >      dataFrame,
                                int                      E,
                                int                      tau,
                                std::vector<std::string> columnNames,
                                bool                     deletePartial = false );
```

Simplex

Simplex projection of the input data file or DataFrame. See the Parameters table for parameter definitions.

`Simplex()` returns a `SimplexValues` structure:

```
struct SimplexValues {
    DataFrame< double >                  predictions;
    std::map< std::string, std::string > parameterMap;
};
```

The `predictions` `DataFrame` has 3 columns "Time", "Observations", "Predictions". nan values are inserted where there is no observation or prediction. The `parameterMap` returns a map of parameters used in the predictions if `parameterList = true.`

Parameters

`lib` and `pred` specify [start stop] row indices of the input data for the library and predictions.

If `embedded` is `false` the data columns are embedded to dimension `E` with time offset `tau`. If `embedded` is `true` the data columns are assumed to be a multivariable data block.

If `knn` is not specified, and `embedded` is `false`, it is set equal to `E+1`. If `embedded` is `true, knn` is set equal to the number of `columns + 1`.

`exclusionRadius` defines the number of library rows excluded from the state-space library with respect to a temporal "radius" from the prediction state. If `exclusionRadius = 1,` library state-space points from observation time series that are within ±1 sequential observation row of the prediction state are not included in the library. Note that units of the radius are time series rows, not time values.

`validLib` implements conditional embedding (CE). It is a boolean vector the same length as the number of time series rows. A `false` entry means that the state-space vector derived from the corresponding time series row will not be included in the state-space library.

If `generateSteps` > `0`, then `Simplex` and `SMap` operate in feedback generative mode. The values of `pred` are over-riden to start at the end of the data. At each step one prediction is made, added to the `columns` data, a new time-delay embedded is created, and the cycle repeated for `generateSteps`. Feedback generation only operates on a univariate time series that is time-delay embedded. The `columns` and `target` variables must be the same. If `generateLibrary` is `false` the state-space library is not expanded as predictions are generated, it is static. If `generateLibrary` is `true` the state-space library has the generated prediction added to the library at each step.

If `parameterList = true`, then `parameterMap` is populated.

```
//----------------------------------------------------------------
// Overload 1: Explicit data file path/name
//----------------------------------------------------------------
SimplexValues Simplex( std::string pathIn          = "./data/",
                       std::string dataFile        = "",
                       std::string pathOut         = "./",
                       std::string predictFile     = "",
                       std::string lib             = "",
                       std::string pred            = "",
                       int         E               = 0,
                       int         Tp              = 1,
                       int         knn             = 0,
                       int         tau             = -1,
                       int         exclusionRadius = 0,
                       std::string columns         = "",
                       std::string target          = "",
                       bool        embedded        = false,
                       bool        const_pred    =   = false,
                       bool        verbose         true,
                       std::vector<bool> validLib  = std::vector<bool>(),
                       int         generateSteps   = 0,
                       bool        generateLibrary = false,
                       bool        parameterList   = false );


//----------------------------------------------------------------
// Overload 2: DataFrame reference provided
//----------------------------------------------------------------
SimplexValues Simplex( DataFrame< double > & dataFrameIn,
                       std::string pathOut         = "./",
                       std::string predictFile     = "",
                       std::string lib             = "",
                       std::string pred            = "",
                       int         E               = 0,
                       int         Tp              = 1,
                       int         knn             = 0,
                       int         tau             = -1,
                       int         exclusionRadius = 0,
                       std::string columns         = "",
                       std::string target          = "",
                       bool        embedded        = false,
                       bool        const_pred      = false,
                       bool        verbose         = true,
                       std::vector<bool> validLib  = std::vector<bool>(),
                       int         generateSteps   = 0,
                       bool        generateLibrary = false,
                       bool        parameterList   = false );
```

SMap

SMap projection of the input data file or DataFrame. See the Parameters table for parameter definitions.

`SMap()` returns a `SMapValues` structure:

```
struct SMapValues {
    DataFrame< double > predictions;
    DataFrame< double > coefficients;
    DataFrame< double > singularValues;
    std::map< std::string, std::string > parameterMap;
};
```

The `predictions` `DataFrame` has 3 columns "Time", "Observations", "Predictions". nan values are inserted where there is no observation or prediction. If `predictFile` is provided the `predictions` will be written to it in csv format.

The `coefficients` `DataFrame` will have `E+2` columns. The first column is the "Time" vector, the remaining `E+1` columns are the SMap SVD fit coefficients. The first column "C0" is the bias term, following coefficients are ∂ columns[i] / ∂ target.

The `singularValues` `DataFrame` will have `E+2` columns. The first column is the "Time" vector, the remaining `E+1` columns are the SMap SVD singular values.

If `nan` are detected in the `columns` or `target`, and argument `ignoreNan` is true (default), the library is automatically redefined to exclude nan observations. If argument `ignoreNan` is `false`, no library adjusment is performed, the user may define the library (`lib`) to exclude observation rows with `nan`.

The `parameterMap` returns a map of parameters used in the predictions if `parameterList = true`.

Parameters

`lib` and `pred` specify [start, stop] row indices of the input data for the library and predictions.

If `embedded` is `false` the data columns are embedded to dimension `E` with offset `tau`. If `embedded` is true the data columns are assumed to be a multivariable data block. If `smapFile` is provided the `coefficients` will be written to it in csv format.

If a multivariate data set is used (number of `columns > 1`) it must use `embedded = true` with `E` equal to the number of `columns`. This prevents the function from internally time-delay embedding the multiple columns to dimension E. If the internal time-delay embedding is performed, then state-space columns will not correspond to the intended dimensions in the matrix inversion, coefficient assignment, and prediction. In the multivariate case, the user should first prepare the embedding (using `Embed()` for time-delay embedding if desired), then pass this embedding to `SMap` with appropriately specified `columns`, `E`, and `embedded = true`.

If knn is not specified, it is set equal to the library size. If knn is specified, it must be greater than E.

exclusionRadius defines the number of library rows excluded from the state-space library with respect to a temporal "radius" from the prediction state. If exclusionRadius = 1, library state-space points from observation time series that are within ±1 sequential observation row of the prediction state are not included in the library. Note that units of the radius are time series rows, not time values.

ignoreNan automatically redefines the library to avoid nan observations and associated state vectors. If ignoreNan is false the library is not changed. The user can manually specify library row segements to ignore nan values.

validLib implements conditional embedding (CE). It is a boolean vector the same length as the number of time series rows. A false entry means that the state-space vector derived from the corresponding time series row will not be included in the state-space library.

The default solver for the SMap coefficient matrix is the LAPACK SVD function dgelss(). If the default solver is used, SMap singular values are returned.

If generateSteps > 0, then Simplex and SMap operate in feedback generative mode. The values of pred are over-riden to start at the end of the data. At each step one prediction is made, added to the columns data, a new time-delay embedded is created, and the cycle repeated for generateSteps. Feedback generation only operates on a univariate time series that is time-delay embedded. The columns and target variables must be the same. If generateLibrary is false the state-space library is not expanded as predictions are generated, it is static. If generateLibrary is true the state-space library has the generated prediction added to the library at each step.

If parameterList = true, parameterMap is populated.

SMap

```
//----------------------------------------------------------------
// Overload 1: Explicit data file path/name
//----------------------------------------------------------------
SMapValues SMap( std::string pathIn          = "./data/",
                 std::string dataFile        = "",
                 std::string pathOut         = "./",
                 std::string predictFile     = "",
                 std::string lib             = "",
                 std::string pred            = "",
                 int         E               = 0,
                 int         Tp              = 1,
                 int         knn             = 0,
                 int         tau             = -1,
                 double      theta           = 0,
                 int         exclusionRadius = 0,
                 std::string columns         = "",
                 std::string target          = "",
                 std::string smapCoefFile    = "",
                 std::string smapSVFile      = "",
                 bool        embedded        = false,
                 bool        const_pred      = false,
                 bool        verbose         = true,
                 std::vector<bool> validLib  = std::vector<bool>(),
                 bool        ignoreNan       = true,
                 int         generateSteps   = 0,
                 bool        generateLibrary = false,
                 bool        parameterList   = false );


//----------------------------------------------------------------
// Overload 2: DataFrame reference provided
//----------------------------------------------------------------
SMapValues SMap( DataFrame< double > &dataFrameIn,
                 std::string pathOut         = "./",
                 std::string predictFile     = "",
                 std::string lib             = "",
                 std::string pred            = "",
                 int         E               = 0,
                 int         Tp              = 1,
                 int         knn             = 0,
                 int         tau             = -1,
                 double      theta           = 0,
                 int         exclusionRadius = 0,
                 std::string columns         = "",
                 std::string target          = "",
                 std::string smapCoefFile    = "",
                 std::string smapSVFile      = "",
                 bool        embedded        = false,
                 bool        const_pred      = false,
                 bool        verbose         = true,
                 std::vector<bool> validLib  = std::vector<bool>(),
                 bool        ignoreNan       = true,
                 int         generateSteps   = 0,
                 bool        generateLibrary = false,
                 bool        parameterList   = false );
```

```
//------------------------------------------------------------------------------
// Overload 3: Data path/file with external solver object, init to default SVD
//------------------------------------------------------------------------------
SMapValues SMap( std::string pathIn          = "./data/",
                 std::string dataFile        = "",
                 std::string pathOut         = "./",
                 std::string predictFile     = "",
                 std::string lib             = "",
                 std::string pred            = "",
                 int         E               = 0,
                 int         Tp              = 1,
                 int         knn             = 0,
                 int         tau             = -1,
                 double      theta           = 0,
                 int         exclusionRadius = 0,
                 std::string columns         = "",
                 std::string target          = "",
                 std::string smapCoefFile    = "",
                 std::string smapSVFile      = "",
                 SVDValues  (*solver)(DataFrame < double >,
                                      std::valarray < double >) = &SVD,
                 bool        embedded        = false,
                 bool        const_predict   = false,
                 bool        verbose         = true,
                 std::vector<bool> validLib  = std::vector<bool>(),
                 bool        ignoreNan       = true,
                 int         generateSteps   = 0,
                 bool        generateLibrary = false,
                 bool        parameterList   = false );
```

SMap

```
//------------------------------------------------------------------------
// Overload 4: DataFrame with external solver object, init to default SVD
//------------------------------------------------------------------------
SMapValues SMap( DataFrame< double > &dataFrameIn,
                 std::string pathOut         = "./",
                 std::string predictFile     = "",
                 std::string lib             = "",
                 std::string pred            = "",
                 int         E               = 0,
                 int         Tp              = 1,
                 int         knn             = 0,
                 int         tau             = -1,
                 double      theta           = 0,
                 int         exclusionRadius = 0,
                 std::string columns         = "",
                 std::string target          = "",
                 std::string smapCoefFile    = "",
                 std::string smapSVFile      = "",
                 SVDValues  (*solver)(DataFrame < double >,
                                      std::valarray < double >) = &SVD,
                 bool        embedded        = false,
                 bool        const_predict   = false,
                 bool        verbose         = true,
                 std::vector<bool> validLib  = std::vector<bool>(),
                 bool        ignoreNan       = true,
                 int         generateSteps   = 0,
                 bool        generateLibrary = false,
                 bool        parameterList   = false );
```

CCM

Convergent cross mapping of `columns` against `target` via Simplex. Normally, one `column` and one `target` are specified. The `column` time series is time-delay embedded to dimension E, cross mapped with the `target` time series. The target time series is then embedded to E and cross mapped against the column as the "target" time series, not an embedding.

If there are multiple `columns` and `embedded` is false, each column is time-delay embedded to dimension `E` creating an N-columns * E dimensional "mixed" embedding. If `embedded` is true, no time-delay embedding is done, creating a multivariate embedding of the speficied columns. The same logic applies if multiple `target` are specified for the "reverse" mapping. If `embedded` is false, each target is time-delay embedded to dimension `E` creating an N-target * E dimensional "mixed" embedding cross mapped to the first `column` as the cross map target. If `embedded` is true, no time-delay embedding is done, creating a multivariate embedding of the speficied target(s).

Cross mappings are performed between `column : target`, and, `target : column` in separate threads. Threading can be disabled in the makefile by removing `-DCCM_THREADED`.

Note: The entire prediction vector is used in the Simplex prediction at each library subset size.
See the Parameters table for parameter definitions.

`CCM()` returns a CCM`Values` structure:

```
struct CCMValues {
    DataFrame< double > AllLibStats;  // unified mean libsize, rho, RMSE, MAE
    CrossMapValues      CrossMap1;
    CrossMapValues      CrossMap2;
    std::map< std::string, std::string > parameterMap;
};
struct CrossMapValues {
    DataFrame< double > LibStats;     // mean libsize, rho, RMSE, MAE
    DataFrame< double > PredictStats; // each predict libsize, rho, RMSE, MAE
    std::forward_list< DataFrame< double > > Predictions;
};
```

`CCMValues::LibStats DataFrame` has 3 columns. The first column is "LibSize", the second and third columns are Pearson correlation coefficients for "column : target" and "target : column" cross mapping.

If the `includeData` parameter is true, then the `CrossMapValues` structures `CrossMap1` and `CrossMap2` in `CCMValues` are populated with the DataFrame `PredictStats` and list of DataFrames `Predictions`. `LibStats` is used for internal storage, and is empty upon return.

Parameters

`The libSizes` parameter is a string of whitespace or comma separated library sizes. If the string has 3 values, and, if the third value is less than the second value, then the three values are interpreted as a sequence generator specifying "start stop increment" row values, i.e. "10 80 10" will evaluate library sizes from 10 to 80 in increments of 10.

If `random` is `true`, `sample` observations are radomly selected from the subset of each library size. If `replacement` is `true`, observations are selected with replacement. If `seed=0` a random seed is generated for the random number generator. Otherwise, `seed` is used to initialise the random number generator.

If `random` is false, `sample` is ignored and contiguous library rows up to the current library size are used. Note this is not Convergent Cross Mapping (CCM).

```
//----------------------------------------------------------------
// Overload 1: Explicit data file path/name
//----------------------------------------------------------------
CCMValues CCM( std::string pathIn          = "./data/",
               std::string dataFile        = "",
               std::string pathOut         = "./",
               std::string predictFile     = "",
               int         E               = 0,
               int         Tp              = 0,
               int         knn             = 0,
               int         tau             = -1,
               int         exclusionRadius = 0,
               std::string columns         = "",
               std::string target          = "",
               std::string libSizes        = "",
               int         sample          = 0,
               bool        random          = true,
               bool        replacement     = false,
               unsigned    seed            = 0,     // seed=0: use RNG
               bool        embedded        = false,
               bool        includeData     = false,
               bool        parameterList   = false,
               bool        verbose         = true );


//----------------------------------------------------------------
// Overload 2: DataFrame reference provided
//----------------------------------------------------------------
CCMValues CCM( DataFrame< double > &dataFrameIn,
               std::string pathOut         = "./",
               std::string predictFile     = "",
               int         E               = 0,
               int         Tp              = 0,
               int         knn             = 0,
               int         tau             = -1,
               int         exclusionRadius = 0,
               std::string columns         = "",
               std::string target          = "",
               std::string libSizes        = "",
               int         sample          = 0,
               bool        random          = true,
               bool        replacement     = false,
               unsigned    seed            = 0,     // seed=0: use RNG
               bool        embedded        = false,
               bool        includeData     = false,
               bool        parameterList   = false,
               bool        verbose         = true );
```

Multiview

Multiview embedding and forecasting of the input data file or DataFrame. See the Parameters table for parameter definitions.

`Multiview()` returns a `MultiviewValues` structure:

```
struct MultiviewValues {
    DataFrame< double > ComboRho;
    DataFrame< double > Predictions;
    std::map< std::string, std::vector< std::string >  > ColumnNames;
    std::map< std::string, std::string > parameterMap;
};
```

The `ComboRho` `DataFrame` has `D+3` columns. The first `D` columns are the the column indices in the input `DataFrame` embedding (not the input `DataFrame`) and are applied to multivariate Simplex predictions. The last three columns are "rho", "MAE", "RMSE" corresponding to the prediction Pearson correlation, maximum absolute error and root mean square error.

`ColumnNames` is a map of string vectors listing the multiview embedding column names of the input `DataFrame` embedding.

The `Predictions` `DataFrame` has 3 columns "Time", "Observations", "Predictions". `nan` values are inserted where there is no observation or prediction. If `predictFile` is provided `Predictions` will be written to it in csv format.

Parameters

D represents the number of variables to combine for each assessment, if not specified, it is the number of columns.

E is the embedding dimension of each variable. If E = 1, no time delay embedding is done.

`multiview` is the number of top-ranked D-dimensional predictions to "average" for the final prediction. Corresponds to parameter k in Ye & Sugihara with default k = sqrt(C) where C is the number of combinations C(n,D) available from the embedding columns taken D at-a-time.

`trainLib` specifies whether projections used to rank the column combinations are done in-sample (pred = lib, the default), or, using the lib and pred specified as input options (`trainLib false`).

`lib` and `pred` specify [start, stop] row indices of the input data for the library and predictions.

If `knn` is not specified, it is set equal to `D+1`.

```
//------------------------------------------------------------------
// Overload 1: Explicit data file path/name
//------------------------------------------------------------------
MultiviewValues Multiview( std::string pathIn          = "./",
                           std::string dataFile        = "",
                           std::string pathOut         = "./",
                           std::string predictFile     = "",
                           std::string lib             = "",
                           std::string pred            = "",
                           int         D               = 0,
                           int         E               = 1,
                           int         Tp              = 1,
                           int         knn             = 0,
                           int         tau             = -1,
                           std::string columns         = "",
                           std::string target          = "",
                           int         multiview       = 0,
                           int         exclusionRadius = 0,
                           bool        trainLib        = true,
                           bool        excludeTarget   = false,
                           bool        parameterList   = false,
                           bool        verbose         = false,
                           unsigned    nThreads        = 4 );



//------------------------------------------------------------------
// Overload 2: DataFrame provided
//------------------------------------------------------------------
MultiviewValues Multiview( DataFrame< double >,
                           std::string pathOut         = "./",
                           std::string predictFile     = "",
                           std::string lib             = "",
                           std::string pred            = "",
                           int         D               = 0,
                           int         E               = 1,
                           int         Tp              = 1,
                           int         knn             = 0,
                           int         tau             = -1,
                           std::string columns         = "",
                           std::string target          = "",
                           int         multiview       = 0,
                           int         exclusionRadius = 0,
                           bool        trainLib        = true,
                           bool        excludeTarget   = false,
                           bool        parameterList   = false,
                           bool        verbose         = false,
                           unsigned    nThreads        = 4 );
```

EmbedDimension

Evaluate Simplex prediction skill for embedding dimensions from 1 to `maxE` (default 10). The returned `DataFrame` has columns "E" and "rho". See the Parameters table for parameter definitions.

Note: `nThreads` defines the number of worker threads for the 10 embeddings. The maximum number of threads is `maxE`.

```
//----------------------------------------------------------------
// Overload 1: Explicit data file path/name
//----------------------------------------------------------------
DataFrame<double> EmbedDimension( std::string pathIn          = "./data/",
                                  std::string dataFile        = "",
                                  std::string pathOut         = "./",
                                  std::string predictFile     = "",
                                  std::string lib             = "",
                                  std::string pred            = "",
                                  int         maxE            = 10,
                                  int         Tp              = 1,
                                  int         tau             = -1,
                                  int         exclusionRadius = 0,
                                  std::string columns         = "",
                                  std::string target          = "",
                                  bool        embedded        = false,
                                  bool        verbose         = true,
                                  std::vector<bool> validLib  =
                                                    std::vector<bool>(),
                                  unsigned    nThreads        = 4 );


//----------------------------------------------------------------
// Overload 2: DataFrame reference provided
//----------------------------------------------------------------
DataFrame<double> EmbedDimension( DataFrame< double > &dataFrameIn,
                                  std::string pathOut         = "./",
                                  std::string predictFile     = "",
                                  std::string lib             = "",
                                  std::string pred            = "",
                                  int         maxE            = 10,
                                  int         Tp              = 1,
                                  int         tau             = -1,
                                  int         exclusionRadius = 0,
                                  std::string columns         = "",
                                  std::string target          = "",
                                  bool        embedded        = false,
                                  bool        verbose         = true,
                                  ustd::vector<bool> validLib =
                                                     std::vector<bool>(),
                                  nsigned    nThreads         = 4 );
```

PredictInterval

Evaluate Simplex prediction skill for forecast intervals from 1 to `maxTp` (default 10). The returned `DataFrame` has columns "Tp" and "rho". See the Parameters table for parameter definitions.

Note: `nThreads` defines the number of worker threads for the 10 prediction interval forecasts. The maximum number of threads is `maxTp`.

```
//----------------------------------------------------------------
// Overload 1: Explicit data file path/name
//----------------------------------------------------------------
DataFrame<double> PredictInterval( std::string pathIn          = "./data/",
                                   std::string dataFile        = "",
                                   std::string pathOut         = "./",
                                   std::string predictFile     = "",
                                   std::string lib             = "",
                                   std::string pred            = "",
                                   int         maxTp           = 10,
                                   int         E               = 0,
                                   int         tau             = -1,
                                   int         exclusionRadius = 0,
                                   std::string columns         = "",
                                   std::string target          = "",
                                   bool        embedded        = false,
                                   bool        verbose         = true,
                                   std::vector<bool> validLib  =
                                                     std::vector<bool>(),
                                   unsigned    nThreads        = 4 );


//----------------------------------------------------------------
// Overload 2: DataFrame reference provided
//----------------------------------------------------------------
DataFrame<double> PredictInterval( DataFrame< double > &dataFrameIn,
                                   std::string pathOut         = "./",
                                   std::string predictFile     = "",
                                   std::string lib             = "",
                                   std::string pred            = "",
                                   int         maxTp           = 10,
                                   int         E               = 0,
                                   int         tau             = -1,
                                   int         exclusionRadius = 0,
                                   std::string columns         = "",
                                   std::string target          = "",
                                   bool        embedded        = false,
                                   bool        verbose         = true,
                                   std::vector<bool> validLib  =
                                                     std::vector<bool>(),
                                   unsigned    nThreads        = 4 );
```

PredictNonlinear

Evaluate SMap prediction skill for localisation parameters θ specifed in the string `theta`. Default values of theta are `"0.01 0.1 0.3 0.5 0.75 1 1.5 2 3 4 5 6 7 8 9"`. Default `knn` (0) sets to size of the library. Smaller `knn` values can be specified.

The returned `DataFrame` has columns "theta" and "rho". See the Parameters table for parameter definitions.

Note: `nThreads` defines the number of worker threads for the θ value forecasts.

```
//----------------------------------------------------------------
// Overload 1: Explicit data file path/name
//----------------------------------------------------------------
DataFrame<double> PredictNonlinear( std::string pathIn          = "./data/",
                                    std::string dataFile        = "",
                                    std::string pathOut         = "./",
                                    std::string predictFile     = "",
                                    std::string lib             = "",
                                    std::string pred            = "",
                                    std::string theta           = "",
                                    int         E               = 0,
                                    int         Tp              = 1,
                                    int         knn             = 0,
                                    int         tau             = -1,
                                    int         exclusionRadius = 0,
                                    std::string columns         = "",
                                    std::string target          = "",
                                    bool        embedded        = false,
                                    bool        verbose         = true,
                                    std::vector<bool> validLib  =
                                                      std::vector<bool>(),
                                    bool        ignoreNan       = true,
                                    unsigned    nThreads        = 4 );
```

```
//------------------------------------------------------------------
// Overload 2: DataFrame reference provided
//------------------------------------------------------------------
DataFrame<double> PredictNonlinear( DataFrame< double > &dataFrameIn,
                                    std::string pathOut         = "./",
                                    std::string predictFile     = "",
                                    std::string lib             = "",
                                    std::string pred            = "",
                                    std::string theta           = "",
                                    int         E               = 0,
                                    int         Tp              = 1,
                                    int         knn             = 0,
                                    int         tau             = -1,
                                    int         exclusionRadius = 0,
                                    std::string columns         = "",
                                    std::string target          = "",
                                    bool        embedded        = false,
                                    bool        verbose         = true,
                                    std::vector<bool> validLib  =
                                                        std::vector<bool>(),
                                    bool        ignoreNan       = true,

                                    unsigned    nThreads        = 4 );
```

ComputeError

Compute Pearson correlation coefficient, maximum absolute error (MAE) and root mean square error (RMSE) between two vectors.

`ComputeError()` returns a `VectorError` struct:

```
struct VectorError {
    double rho;
    double RMSE;
    double MAE;
};

//---------------------------------------------------------------
//---------------------------------------------------------------
VectorError ComputeError( std::valarray< double > obsIn,
                          std::valarray< double > predIn );
```

Application Notes

All data input files are assumed to be in csv format. The files are required to have a single line header with column names.

It is assumed that the first column of the csv file is a vector of time values, time strings, or time indices for each observation (row). All subsequent columns are expected to be numeric. However, a `DataFrame` can be created with the `noTime = true` parameter to read in numeric data with no time in the first column. Such data frames should not be passed to prediction functions.

The pyEDM and rEDM wrappers independently implement a `noTime` parameter in their respective API's. There, the wrapper creates a row index vector inserting it as the first column of the DataFrame if `noTime = true` is specified. Subsequent calls to the cppEDM API always provide a DataFrame with a first column of time values.

`SMap()` should be called with `DataFrame` that have columns explicity corresponding to dimensions E. Subsequently, if a multivariate data set is used, it should Not be called with an embedding from `Embed()` since `Embed()` will add lagged coordinates for each variable. These extra columns will not correspond to the intended dimensions in the matrix inversion and prediction reconstruction. In this case, use the `embedded` parameter set to `true` so that the columns selected correspond to the proper dimension.

See `etc/Notes` for details regarding the building an application (`etc/Test.cc`) on Windows.

See `etc/Notes` for an explanation of differences between `Simplex()` and `CCM()` default embeddings, and the effect this has on predictions.

Example Application

This application is assumed to be located in the etc/ directory. Otherwise, adust the -I and -L compiler flags and the `pathIn` argument accordingly. The file etc/Test.cc shows sample invocations for several API functions.

```
// g++ TestApp.cc -o TestApp -g -I../src -L../lib -lstdc++ -lEDM -llapack

#include "API.h"

int main( int argc, char *argv[] ) {

    try {
        //----------------------------------------------------------
        // embedded = false : Simplex embeds all columns to E = 3
        //----------------------------------------------------------
        DataFrame<double> dataFrame =
            Simplex( "../data/",         // pathIn
                     "block_3sp.csv",    // dataFile
                     "./",               // pathOut
                     "Block3sp_E3.csv",  // predictFile
                     "1 100",            // lib
                     "101 195",          // pred
                      3,                 // E
                      1,                 // Tp
                      0,                 // knn
                      -1,                // tau
                      0,                 // exclusionRadius
                     "x_t y_t z_t",      // columns
                     "x_t",              // target
                     false,              // embedded
                     false,              // const_predict
                     true );             // verbose

        dataFrame.MaxRowPrint() = 12; // Set number of rows to print
        std::cout << dataFrame;

        VectorError ve = ComputeError(
            dataFrame.VectorColumnName( "Observations" ),
            dataFrame.VectorColumnName( "Predictions"  ) );

        std::cout << "rho " << ve.rho << "  RMSE " << ve.RMSE
                  << "  MAE " << ve.MAE << std::endl << std::endl;

    }

    catch ( const std::exception & e ) {
       std::cout << "Exception caught in main:\n";
      std::cout << e.what() << std::endl;
       return -1;
    }
    catch (...) {
       std::cout << "Unknown exception caught in main.\n";
       return -1;
    }

    std::cout << "Normal termination.\n";

    return 0;
}
```

Code Notes

1) The OSX XCode compiler/linker seems to be incompatible with the C++ standard implementation allowing template classes to be distributed into declarations (.h) and implementation (.cc). To support OSX, DataFrame.h contains both declarations and implementations. See: etc/libstdc++_Notes.txt.

2) The code relies heavily on class and data containers without explicit heap allocation. This facilitates garbage collection. If the code encounters massive data objects/large problems, this may warrant investigation.

3) SMap uses the LAPACK (http://www.netlib.org/lapack/explore-html/index.html) routine `dgelss()` to solve the local linear maps by singular value decomposition.

References

Dixon, P. A., M. Milicich, and G. Sugihara, 1999. Episodic fluctuations in larval supply. Science 283:1528–1530.

Sugihara G. and May R. 1990. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature, 344:734–741.

Sugihara G. 1994. Nonlinear forecasting for the classification of natural time series. Philosophical Transactions: Physical Sciences and Engineering, 348 (1688) : 477–495.

Sugihara G., May R., Ye H., Hsieh C., Deyle E., Fogarty M., Munch S., 2012. Detecting Causality in Complex Ecosystems. Science 338:496-500.

Takens, F. Detecting strange attractors in turbulence. Lect. Notes Math. 898, 366–381 (1981).

Ye H., and G. Sugihara, 2016. Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. Science 353:922–925.