[R] Library **envlp**: User's Guide

January 25, 2016

Contents **1**

1

Quick start

1.1. Overview

The envelope model is a new area in multivariate analysis. It uses dimension reduction techniques to achieve efficient estimation of parameters, for example, the regression coefficients in multivariate linear regression (MLR).

This package provides a general routine, **envMU**, which allows estimation of the **M** envelope of span(**U**) given estimators of **M** and **U**. The routine **envMU** does not presume a model. In the context of multivariate linear regression (MLR), this package implements response envelopes (**env** function), partial response envelopes (**penv** function) and envelopes in the predictor space (**envMU** function). For each of these model-based routines the package provides inference tools including bootstrap, cross validation, estimation and prediction, hypothesis testing on coefficients are included. Tools for selection of dimension include AIC, BIC and likelihood ratio testing. Background is available at http://arxiv.org/abs/1509.03767 . Optimization is based on a clockwise coordinate descent algorithm.

envMU Implements estimation of the **M** envelope of span(**U**) given estimators of **M** and **U**. The routine **envMU** does not presume a model.

env Implements the response envelope model in multivariate linear model. The response envelope model is a general tool for efficient estimation in the context of MLR, and it has the potential to achieve substantial efficiency gains when part of the response variables or their linear combination is invariant to the changes of the predictors [Cook et al., 2010].

penv Implements the partial envelope model. The partial envelope model can be applied when part of the predictors that are of main interest. It often gives more efficiency gains than the envelope model in estimating the coefficients of the main predictors [Su and Cook, 2011].

xenv Implements the envelope model in the predictor space. The envelope model in the predictor space is used when some of the predictors or their linear combinations do not contribute to the change of the responses. It can potentially bring a better prediction performance than the standard model, or even partial least squares [Cook et al., 2013].

Module	Dimension Selection	Inference Tools	Section
envMU		Estimation	2.1
env	AIC BIC LRT m-fold CV	Estimation and Prediction Bootstrap for Estimating Standard Errors Hypothesis Test on Coefficients	2.2.1
penv	AIC BIC LRT m-fold CV	Estimation and Prediction Bootstrap for Estimating Standard Errors Hypothesis Test on Coefficients	2.2.2
xenv	AIC BIC LRT m-fold CV	Estimation and Prediction Bootstrap for Estimating Standard Errors Hypothesis Test on Coefficients	2.2.3
predict2env	AIC BIC LRT	Prediction for a new $\mathbf{X}_{\text{new}}$	2.2.4

Table 1.1: Applicability of the library envlp

predict2.env A way to do prediction given a new $\mathbf{X}_{\text{new}}$ using partial envelopes to envelope $\beta\mathbf{X}_{\text{new}}$, which leads to a new method of prediction that has the potential to yield predicted and fitted values with smaller variation than the standard predictions based on the classical linear model (2.1) or the full envelope predictions **env**.

The complete applicability of this toolbox is described in Table 1.1.

1.2. Technical Support

We provide a support website, on which the users can download the toolbox and check recent updates http://www.stat.ufl.edu/~zhihuasu/Renvlp/. For further help, the users can contact the authors of the toolbox: Dennis Cook (dennis@stat.umn.edu), Liliana Forzani (liliana.forzani@gmail.com) and Zhihua Su (zhihuasu@stat.ufl.edu).

1.3. Installation

The usual in [R].
If a previous version of the toolbox is present, it should be removed before installing the new version. To remove, type

```
remove.packages("envlp")
```

1.4. A Guided Tour

This "envlp" library have the following capacity:

1. Finding estimators of envelopes
2. Multivariate Linear regression
    - Dimension selection: Select the dimension of the envelope subspace.
    - Model Fitting with Selected Dimension: Fit the model.
    - Post processing: Inference based on the model fitting.
    - Direct prediction using partial envelopes to envelope $\boldsymbol{\beta}\mathbf{X}_{\text{new}}$.

We will present a tour of our library through all functions.

2

Envelope Models

2.1. Estimation of the M envelope of $\text{span}(\mathbf{U})$

The $\mathbf{M}$ envelope subspace of $\text{span}(\mathbf{U})$ is denote by $\mathcal{E}_{\mathbf{M}}(\text{span}(\mathbf{U}))$ and it is the smallest subspace that contains $\text{span}(\mathbf{U})$ and reduces $\mathbf{M}$.

Given $\sqrt{n}$ consistent estimators of $\mathbf{M}$ and $\mathbf{U}$ respectively. We can find an estimator of $\mathcal{E}_{\mathbf{M}}(\text{span}(\mathbf{U}))$ for a known dimension u by

```
ModelOutput = envMU(M, U, u)
```

2.2. Envelope models for multivariate linear regression

The envelope models are based on the multivariate linear regression model,

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the multivariate response, $\mathbf{X} \in \mathbb{R}^p$ is non-stochastic predictor, and $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ follows a distribution with mean 0, and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ are the unknown intercept and coefficients. By applying sufficient dimension reduction techniques, the envelope models have the potential to obtain efficiency gains in estimating $\boldsymbol{\beta}$, compared to the OLS estimators.

2.2.1. Response envelope model

In the multivariate linear regression context, a coordinate form of the response envelope model [Cook et al., 2010] is

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T,$$

where the regression coefficients $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\mathcal{B} = \text{span}(\boldsymbol{\beta})$, $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ spans $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ – the envelope subspace, $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ spans the orthogonal complement of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$, $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$, $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are coordinates, and u is the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$.

We can fit the response envelope model for a known u by

```
ModelOutput = env(X, Y, u, asy = T)
```

The output ModelOutput is a list, which contains the MLEs of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\alpha}$, and also statistics computed from the model, including the maximized log likelihood, the asymptotic covariance matrix of $\text{vec}(\widehat{\boldsymbol{\beta}})$, the asymptotic standard errors of elements in $\widehat{\boldsymbol{\beta}}$, the ratios of the asymptotic standard errors of the standard model versus the envelope model for elements in $\boldsymbol{\beta}$ and the number of observations in the data. If asy = F, then only the MLEs will be calculated. After fitting the data and get ModelOutput, we can perform post processing inference as computing bootstrap standard errors of $\widehat{\boldsymbol{\beta}}$ by

```
bootse = boot.env(X, Y, u, B)
```

or computing the fitted value or predicted value given an $\mathbf{X}$ by

```
PredictOutput = predict.env(ModelOutput, Xnew)
```

or testing if some linear combination of $\boldsymbol{\beta}$ is equal to a particular matrix, i.e. given $\mathbf{L}$, $\mathbf{R}$, and $\mathbf{A}$, testing if $\mathbf{L}\boldsymbol{\beta}\mathbf{R} = \mathbf{A}$,

```
TestOutput = testcoef.env(ModelOutput, L, R, A)
```

The inputs and outputs of these post processing functions are discussed in the [R] help functions.

The computation of the prediction error for the response envelope estimator using cross validation can be done via

```
CVOutput = cv.env(X, Y, u, m = number of folders, nperm)
```

To select the dimension of the response envelope subspace u, via Akaike information criterion (AIC), Bayesian information criterion (BIC) and likelihood ratio testing with specified significance level for the response envelope model, we can use the following function

```
u = u.env(X, Y, alpha = 0.01) # Users can specify other significance level
```

The possible values of u can be any integer from 0 to r. When $u = r$, the envelope model is equivalent to the standard multivariate linear model. And when $u = 0$, it means that $\boldsymbol{\beta} = 0$, then the changes in $\mathbf{Y}$ do not depend on $\mathbf{X}$. The outputs **u.aic**, **u.bic** and **u.lrt** are the chosen values for u for AIC, BIC and LRT respectivly.

2.2.2. Partial Envelope Model

The partial envelope model [Su and Cook, 2011] can be applied when part of the predictors are of main interest. This is particular useful when the number of the predictors p is large, but only a small number of predictors are of main interest. Suppose that $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T)$, where $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ are predictors of main interest, and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$ are covariates, $p_1 + p_2 = p$. Then the standard model is formulated as $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\varepsilon}$, and the coordinate form of the partial envelope model is

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T,$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\Gamma}\boldsymbol{\eta} \in \mathbb{R}^{r\times p_1}$, $\mathcal{B}_1 = \text{span}(\boldsymbol{\beta}_1)$, $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ spans the partial envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_1)$, $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times u}$ spans the orthogonal complement of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_1)$, $\boldsymbol{\eta} \in \mathbb{R}^{u\times p_1}$, $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$, $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are coordinates, $\boldsymbol{\beta}_2 \in \mathbb{R}^{r\times p_2}$ contains the coefficients for $\mathbf{X}_2$ and u is the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}_1)$.

We can fit the partial envelope model for a known u by

```
ModelOutput = penv(X1, X2, Y, u, asy = T)
```

The output ModelOutput is a list containing the MLEs of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\alpha}$, as well as statistics computed from the model, including the maximized log likelihood, the asymptotic covariance matrix of $(\text{vec}(\widehat{\boldsymbol{\beta}}_2)^T, \text{vec}(\widehat{\boldsymbol{\beta}}_1)^T)^T$, the asymptotic standard errors of elements in $\widehat{\boldsymbol{\beta}}_1$, the ratios of the asymptotic standard errors of the standard model versus the partial envelope model for elements in $\boldsymbol{\beta}_1$ and the number of observations in the data. If asy = F, then only the MLEs will be calculated. The functions for post processing inference are

```
bootse = boot.penv(X1, X2, Y, u, B)
PredictOutput = predict.penv(ModelOutput, X1new, X2new)
TestOutput = testcoef.penv(ModelOutput, L, R, A)
CVOutput = cv.penv(X1, X2, Y, u, m = number of folders, nperm)
```

These functions are used similarly as those for the envelope model in Section 2.2.1.
The following function can be applied to select u,

```
u = u.penv(X, Y, alpha = 0.01)
```

The possible values of u are any integer from 0 to r, when $u = r$, the partial envelope model reduces the standard model, and when $u = 0$, the changes in $\mathbf{Y}$ do not depend on the changes in $\mathbf{X}_1$ given $\mathbf{X}_2$.

2.2.3. Envelope Model in the Predictor Space

The envelope model in the predictor space has the potential to have better prediction performance compared to the standard model, or even the partial least squares. In fact, in the population version, the envelope estimator in the predictor space is equivalent to the partial least squares estimator, but in the sample version, its performance is normally superior to the partial least squares estimator.

We slightly change the formulation of the standard model to $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$, to be consistent with the notations in Cook et al. [2013]. Then the coordinate form of the envelope model in the predictor space is as follows:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{X} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T,$$

where $\boldsymbol{\mu} \in \mathbb{R}^r$ is the intercept, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} \in \mathbb{R}^{p\times r}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ spans the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{B})$, and $\mathcal{B} = \text{span}(\boldsymbol{\beta}^T)$, $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ spans the orthogonal complement of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{B})$, $\boldsymbol{\eta} \in \mathbb{R}^{u\times r}$, $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$, and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$ carry coordinates, and u is the dimension of the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{B})$.

We can fit the envelope model in the predictor space for a known u by

```
ModelOutput = xenv(X, Y, u, asy = T)
```

The output ModelOutput is a list, which contains the MLEs of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_\mathbf{X}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\mu}$, and also statistics computed from the model, including the maximized log likelihood, the asymptotic covariance matrix of $\text{vec}(\widehat{\boldsymbol{\beta}})$, the asymptotic standard errors of elements in $\widehat{\boldsymbol{\beta}}$, the ratios of the asymptotic standard errors of the standard model versus the envelope model for elements in $\boldsymbol{\beta}$ and the number of observations in the data. If asy = F, then only the MLEs will be calculated. After model fitting, the following post processing inference can be performed:

```
bootse = boot.xenv(X, Y, u, B)
PredictOutput = predict.xenv(ModelOutput, Xnew)
TestOutput = testcoef.xenvt(ModelOutput, L, R, A)
CVOutput=cv.xenv(X, Y, u, m=number of folders, nperm)
```

These functions are used similarly as those for the envelope model in Section 2.2.1.
To select the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{B})$, we apply the following function

```
u = u.xenv(X, Y, alpha = 0.01)
```

The possible values for u can be any integer from 0 to p, when $u = p$, the envelope model reduces to the standard model, and when $u = 0$, the changes in $\mathbf{Y}$ do not depend on $\mathbf{X}$.

2.2.4. Direct prediction using partial envelopes

There is a way to do prediction given a new $\mathbf{X}_{\text{new}}$ using partial envelopes to envelope $\boldsymbol{\beta}\mathbf{X}_{\text{new}}$, which leads to a new method of prediction that has the potential to yield predicted and fitted values with smaller variation than the standard predictions based on the classical linear model (2.1) or the full envelope predictions **env**. Select an $\mathbf{A}_0 \in \mathbb{R}^{p\times(p-1)}$ so that $\mathbf{A} = (\mathbf{X}_{\text{new}}, \mathbf{A}_0) \in \mathbb{R}^{p\times p}$ has full rank. It may be helpful, although not necessary, to choose the columns of $\mathbf{A}_0$ to be orthogonal to $\mathbf{X}_{\text{new}}$. Let $\boldsymbol{\phi}_1 = \boldsymbol{\beta}\mathbf{X}_{\text{new}}$, $\boldsymbol{\phi}_2 = \boldsymbol{\beta}\mathbf{A}_0$, $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ and $\mathbf{Z} = \mathbf{A}^{-1}\mathbf{X} = (Z_1, \mathbf{Z}_2^T)^T$, where $\mathbf{Z}_2 \in \mathbb{R}^{p-1}$. Then we can parameterize model (2.1) as

$$\begin{aligned}\mathbf{Y} &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{A}\mathbf{A}^{-1}\mathbf{X} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\alpha} + (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)\mathbf{Z} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\alpha} + \boldsymbol{\phi}_1 Z_1 + \boldsymbol{\phi}_2 \mathbf{Z}_2 + \boldsymbol{\varepsilon}\end{aligned}$$

We can now parameterize the model in terms of a basis $\boldsymbol{\Gamma}$ of the $\boldsymbol{\Sigma}$-envelope of span$(\boldsymbol{\phi}_1)$, leading to a partial envelope representation for prediction at $\mathbf{X}_{\text{new}}$:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta} Z_1 + \boldsymbol{\phi}_2 \mathbf{Z}_2 + \boldsymbol{\varepsilon},\ \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.$$

This method to predict $\mathbf{Y}$ for a new X_{new} given the dimension u can be done by

```
ModelOutput = predict2env(X,Y, u, Xnew)
```

and to choose the dimension we use

```
u.predict2 = u.predict2.env(X, Y, Xnew)
```

Bibliography

R. D. Cook, I. Helland, and Z. Su. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society B*, 75:851–877, 2013.

R.D. Cook, B. Li, and F. Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statist. Sinica*, 20:927–1010, 2010.

Z. Su and R.D. Cook. Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, 98(1):133–146, 2011. ISSN 0006-3444.