Testing for Superior Predictive Ability using Ox
A Manual for SPA for Ox*

Peter Reinhard Hansen† Jungho Kim‡ Asger Lunde§

February 12, 2003

Abstract

This manual describes how to use the Ox implementation of the test for Superior Predictive Ability (SPA) introduced by Hansen (2001). The manual describes: what is needed to run the program, how to use the program, and how to interpret the outputs. Further, a brief explanation of the theory behind the SPA test is given.

Contents

*Download SPA for Ox from: http://www.econ.brown.edu/fac/Peter_Hansen

†Brown University, Department of Economics, Box B, Providence RI 02912, email: Peter_Hansen@brown.edu.
‡Brown University, Department of Economics, Box B, Providence RI 02912, email: Jungho_Kim@brown.edu.
§The Aarhus School of Business, Fuglesangs Allé 4, DK-8210 Aarhus V, email: alunde@asb.dk.

1. A Quick Guide to SPA for Ox

SPA for Ox can be used to compare two or more forecasting models. A model is chosen as a benchmark model, and the SPA-test can analyze whether any of the competing models are significantly better than the benchmark model, in terms of predictive accuracy.

1.1. What You Need

- Prior to using SPA for Ox, you need to acquire and install Ox on a computer. Ox can be purchased from the following web site.

  `http://www.timberlake-consultancy.com/`

- For academic use, a free version of Ox is available from the web site.

  `http://www.nuff.ox.ac.uk/Users/Doornik/`

- Download the files `SPA.ox,spa_src.h` and `spa_src.oxo` and the data set `mydata.mat` (to be replaced by your own data set) from the web site.

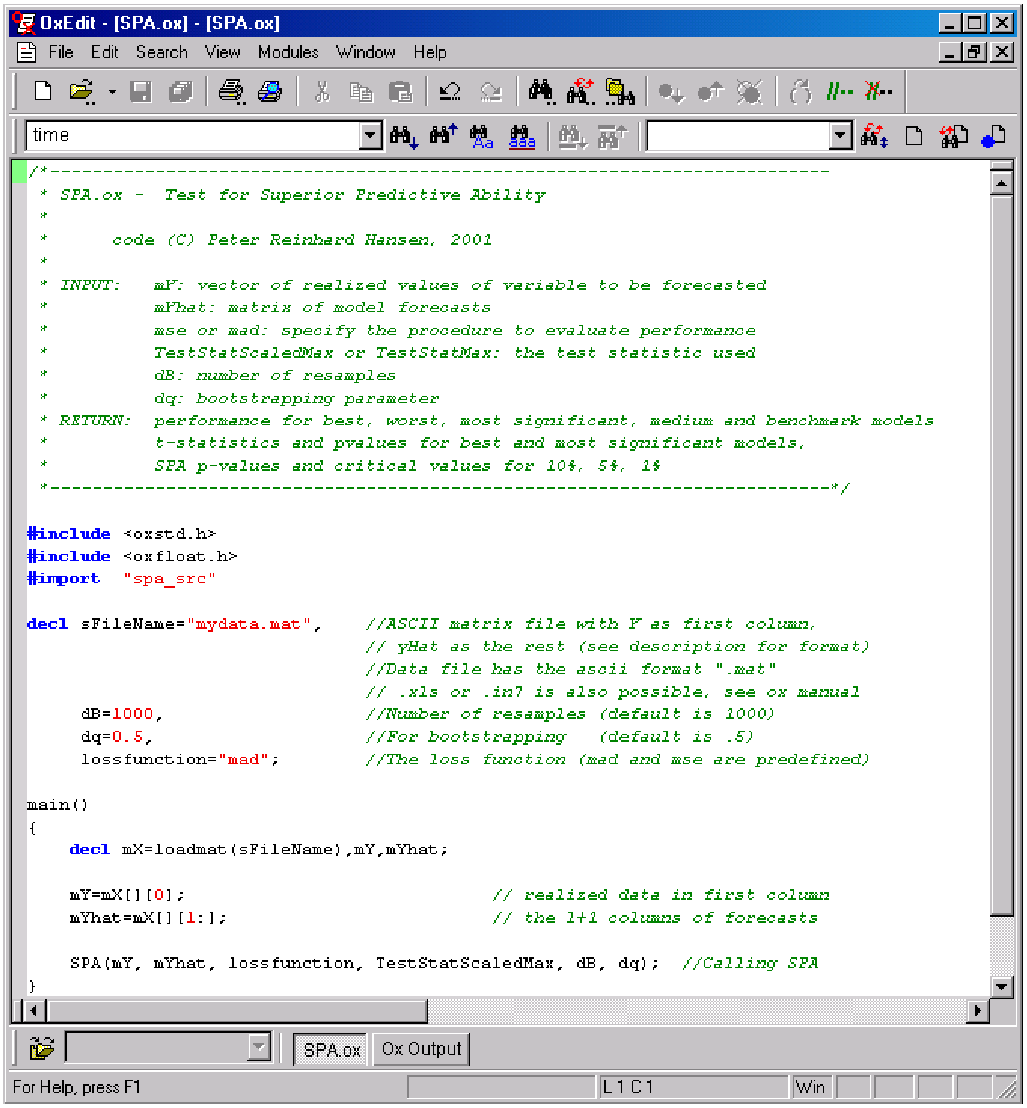
  `http://www.econ.brown.edu/fac/Peter_Hansen/`

- Save the files in the same directory

1.2. Running SPA for Ox the First Time

To run SPA for Ox using the free version of Ox, you use the command: “`oxl spa.ox`”. If you have purchased Ox, you can run the program from the editor OxEdit or GiveWin.

If you open SPA.ox using OxEdit version 1.61, the window looks like this:

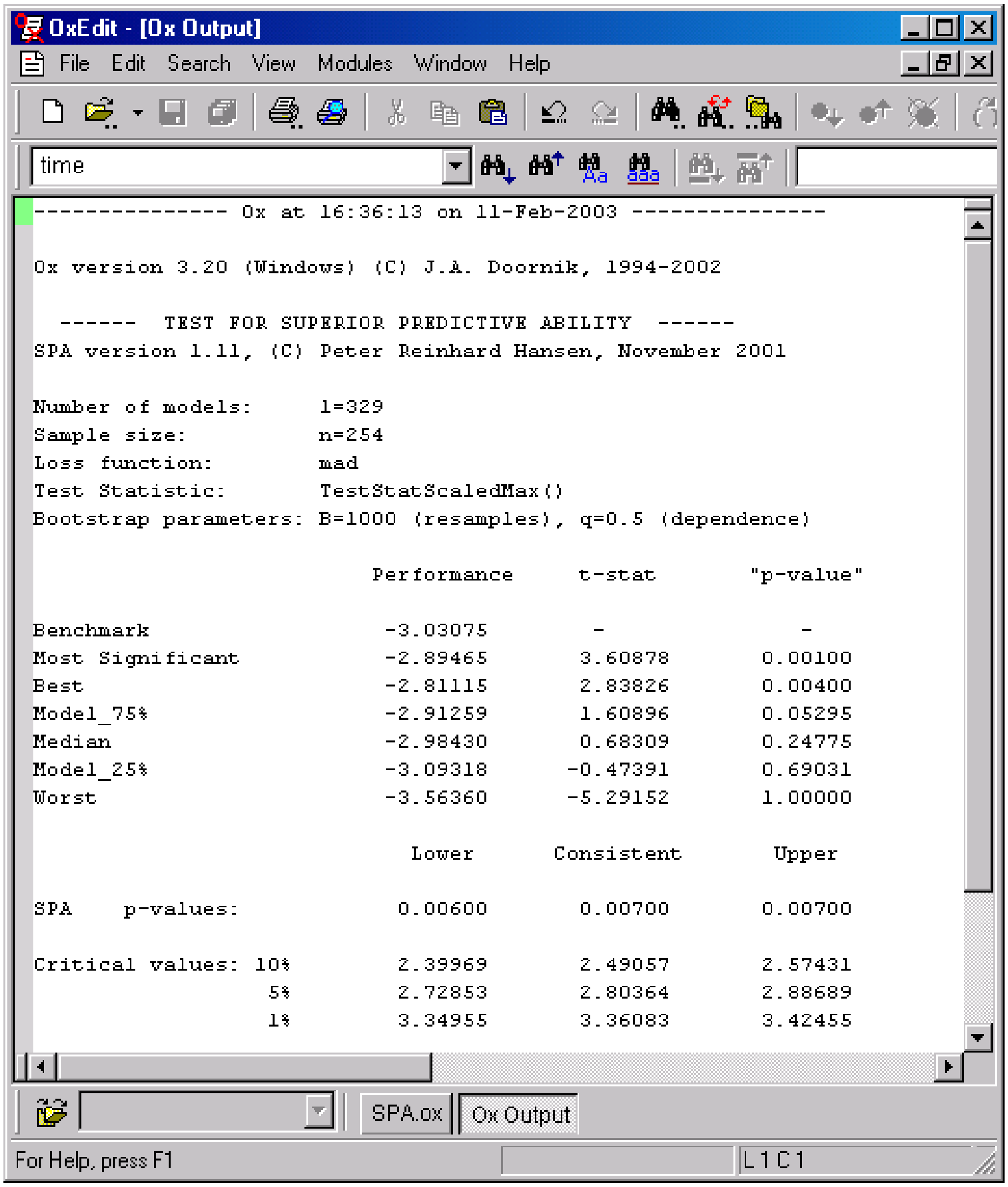



```
/*------------------------------------------------------------------------------
 * SPA.ox -  Test for Superior Predictive Ability
 *
 *      code (C) Peter Reinhard Hansen, 2001
 *
 * INPUT:   mY: vector of realized values of variable to be forecasted
 *          mYhat: matrix of model forecasts
 *          mse or mad: specify the procedure to evaluate performance
 *          TestStatScaledMax or TestStatMax: the test statistic used
 *          dB: number of resamples
 *          dq: bootstrapping parameter
 * RETURN:  performance for best, worst, most significant, medium and benchmark models
 *          t-statistics and pvalues for best and most significant models,
 *          SPA p-values and critical values for 10%, 5%, 1%
 *------------------------------------------------------------------------------*/

#include <oxstd.h>
#include <oxfloat.h>
#import  "spa_src"

decl sFileName="mydata.mat",     //ASCII matrix file with Y as first column,
                                 // yHat as the rest (see description for format)
                                 //Data file has the ascii format ".mat"
                                 // .xls or .in7 is also possible, see ox manual
     dB=1000,                    //Number of resamples (default is 1000)
     dq=0.5,                     //For bootstrapping   (default is .5)
     lossfunction="mad";         //The loss function (mad and mse are predefined)

main()
{
    decl mX=loadmat(sFileName),mY,mYhat;

    mY=mX[][0];                             // realized data in first column
    mYhat=mX[][1:];                         // the 1+1 columns of forecasts

    SPA(mY, mYhat, lossfunction, TestStatScaledMax, dB, dq);  //Calling SPA
}
```

SPA for Ox will produce an output to `SPA.output` (the free version will show the result in the DOS window), which can be redirected a file, see the Ox-manual for details.

1.3. Interpreting the Results

The output produced by Ox using the standard configuration is the following:




```
--------------- Ox at 16:36:13 on 11-Feb-2003 ---------------

Ox version 3.20 (Windows) (C) J.A. Doornik, 1994-2002

  ------ TEST FOR SUPERIOR PREDICTIVE ABILITY ------
SPA version 1.11, (C) Peter Reinhard Hansen, November 2001

Number of models:      l=329
Sample size:           n=254
Loss function:         mad
Test Statistic:        TestStatScaledMax()
Bootstrap parameters: B=1000 (resamples), q=0.5 (dependence)

                          Performance     t-stat       "p-value"

Benchmark                  -3.03075         -              -
Most Significant           -2.89465       3.60878       0.00100
Best                       -2.81115       2.83826       0.00400
Model_75%                  -2.91259       1.60896       0.05295
Median                     -2.98430       0.68309       0.24775
Model_25%                  -3.09318      -0.47391       0.69031
Worst                      -3.56360      -5.29152       1.00000

                             Lower       Consistent      Upper

SPA    p-values:            0.00600       0.00700       0.00700

Critical values: 10%        2.39969       2.49057       2.57431
                  5%        2.72853       2.80364       2.88689
                  1%        3.34955       3.36083       3.42455
```

The output in the picture above can be divided into three parts.[1] The first part contains descriptive statistics, such as the number of competing models, sample size, bootstrap parameters, etc. The second part informs about model performance, and six pair-wise comparisons. The six models being compared to the benchmark model are: (1) The ”Most Significant” model is the model that had the most “significant” performance relative to the benchmark model in the sample being analyzed, and (2-6) those models with a performance that corresponded to the 75%, 50% (median), 25%, and 0% (worst) quantile of model performances. This provides some information about the

[1] The data that were analyzed to produce this output are taken from **?**.

population of model performances in the sample being analyzed. The `Performance` is measured in terms of the loss function that has been specified by the user, in this case the mean absolute deviation (`mad`) has been used, another predefined loss-function is the mean squared error (`mse`). The `t-stat` indicates the significance of a model's performance relative to the benchmark model, although it is important to note that this statistic is not t-distributed, and the "`p-value`" reported next to it cannot be interpreted as a p-value, rather it is a number that is calculated like a p-value, but it ignores the search over models that preceded the selection of the model being compared to the benchmark.

The third part contains the relevant information for testing the hypothesis that the benchmark model is the best forecasting model. The SPA p-value takes the space of models into account and the SPA-Consistent p-value instructs you if there is evidence against the hypothesis. A low p-value (less that .05–.10 say) informs you that the benchmark model is inferior to one or more of the competing models. A high p-value tells you that there the sample being analyzed does not yield strong evidence that the benchmark is outperformed. The number the left (right) of the SPA-Consistent p-value, is a lower (upper) bound for the true p-value. Critical values at the 10%, 5%, and 1% significance levels are given below the SPA p-values.

1.4. Modifying the Configuration to Your Own Empirical Analysis

1. `sFileName`: ASCII matrix file with Y in first column and $\widehat{Y}$ as the rest where Y is a vector of realized values of variable to be forecasted and $\widehat{Y}$ is the matrix containing model-forecasts. The first column of $\hat{Y}$ contains the forecast of the benchmark model. Files with the extension "`.mat`" is recognized by Ox as ASCII files, other formats such as "`.xls`" (Excel) or "`.in7`" (PcGive) are also recognized by Ox. If the data file is not in the same directory as `SPA.ox`, the path must be specified, for example "`C:\Data\mydata.mat`". You can refer to the Ox manual for more information.

2. `dB`: number of bootstrap resamples that are used to the p-values of the test statistic. The default value is 1000.

3. `dq`: bootstrapping parameter. The default value is 0.5.

4. `mse` *or* `mad`: the criterion to evaluate forecasting performance. `mse` refers to Mean Squared Error (MSE), and `mad` stands for Mean Absolute Deviation (MAD). If you are familiar with Ox, it is straight forward to define other loss functions.

2. What is the Test for SPA?

The Superior Predictive Ability test is a test that can be used for comparing the performances of two or more forecasting models. The forecasts are evaluated using a prespecified loss function, and the "best" forecast model is the model that produces the smallest expected loss. Two loss functions are predefine in the SPA code the two are: the Mean Squared Error (`mse`) and Mean Absolute Deviation (`mad`).

Let $L(Y_t, \hat{Y}_t)$ denote the loss if one had made the prediction, $\hat{Y}_t,$ when the realized value turned out to be Y_t. The performance of model $k,$ relative to the benchmark model (at time t), can be defined as:

$$X_k(t) = L(Y_t, \widehat{Y}_{0t}) - L(Y_t, \widehat{Y}_{kt}), \qquad k = 1, \dots, l, \qquad t = 1, \dots, n.$$

The question of interest is whether any of the models $k = 1, \dots, l$ are better than the benchmark model. To analyze this question we formulate the testable hypothesis that the benchmark model is the best forecasting model. This hypothesis can be expressed parametrically as

$$\mu_k = E\left[X_k(t)\right] \leq 0, \qquad k = 1, \dots, l.$$

For notational convenience, we define an l-dimensional vector $\boldsymbol{\mu}$ by

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_l \end{pmatrix} = E \begin{pmatrix} X_1(t) \\ \vdots \\ X_l(t) \end{pmatrix}.$$

Since a positive value of μ_k corresponds to model k being better than the benchmark, we want to test the hypothesis $H_0 : \mu_k \leq 0,$ for $k = 1, ..., l.$ The equivalent vector formulation is

$$H_0 : \boldsymbol{\mu} \leq \mathbf{0}.$$

One way to test this hypothesis is to consider the test statistic

$$T_n^{sm} = \max_k \frac{n^{1/2} \bar{X}_k}{\hat{\sigma}_k},$$

where

$$\bar{X}_k = \frac{1}{n} \sum_{t=1}^{n} X_k(t),$$

and

$$\hat{\sigma}_k^2 = \widehat{\mathrm{var}}(n^{1/2} \bar{X}_k).$$

The latter is estimated by using the bootstrap method. The superscript "*sm*" refers to *standardized maximum*.

Under the regularity condition it holds that

$$T_n^{sm} = \max_k \frac{\bar{X}_k}{\hat{\sigma}_k} \xrightarrow{p} \max_k \frac{\mu_k}{\sigma_k},$$

which is greater than zero if and only if $\mu_k > 0$ for some k. So we can test H_0 using the test statistic T_n^{sm}. The only remaining problem is to derive the distribution of T_n^{sm}, under the assumption of a true null hypothesis. Testing multiple inequalities is more complicated than testing equalities (or a single inequality) because the distribution is not unique under the null hypothesis. Nevertheless, a consistent estimate of the *p*-value can be obtained, as well as an upper and a lower bound, and those are the *p*-values produced by the SPA program.

2.1. SPA *p*-values

The upper bound is the *p*-value of a conservative test which tacitly assumes that all the competing models ($k = 1, \ldots, l$) are as precisely as good as the benchmark in terms of expected loss ($\mu_1 = \cdots = \mu_l$). The lower bound is the *p*-value of a liberal test whose null hypothesis assumes that the models with worse performance than the benchmark are poor models in the limit. Therefore, these can be viewed as asymptotic upper and lower bounds for the actual *p*-value, respectively. The consistent *p*-value is produced by the test for SPA of **?**. This test will asymptotically determine which models are worse than the benchmark and asymptotically prevent them from affecting the distribution of the test statistic, as should be the case. While the conservative test is sensitive to including poor and irrelevant models in the comparison, the consistent and liberal tests are not. The technical reason that the consistent test is unbiased (and more powerful) is the fact that it is asymptotically similar on the boundary of H_0, see **?** for details.

2.2. Critical values for 10%, 5%, and 1%

Since *p*-values are designed to control for the size of a test, they are not informative of the power of a test, which is the probability of rejecting H_0 when H_1 is true. When you fail to reject the null hypothesis, the critical value can be informative about the power of a test. By definition, the critical values show how well a model would need to perform (in standardized performance $\bar{X}_k/\hat{\sigma}_k$) in order to reject the null hypothesis and conclude that a competing model is better than the benchmark.

If the critical values are large, it indicates that the data being analyzed are not very informative about the hypothesis of interest, and the SPA test may lack power. Lack of power makes it difficult to discover a superior model, even if one existed, so it is useful to consider the critical values, if one gets a large *p*-value in the test for Superior Predictive Ability.