On-line Programs on WWW for Some Statistical Analyses

Yuichi Mori*, Masaya Iizuka and Yoshihiro Yamanishi*****

* *Faculty of Informatics, Okayama University of Science*, mori@soci.ous.ac.jp
** *Faculty of Law, Okayama University*, masa@law.okayama-u.ac.jp
*** *Graduate School, Okayama University*, yoshi@stat1.stat.ems.okayama-u.ac.jp

1. Introduction

There are many web pages for statistics on WWW, which are providing statistical resources and computational environment. They are very useful especially for computation in elementary statistics, education of statistical methods and data analysis in some special fields. While the Internet environment for statistics are expanding widely, it is also true from the aspect of actual applications that we meet problems such as how to construct more effective systems for computation on WWW.

Here we discuss various usages of WWW for statistics, and then we consider how we should develop on-line programs by taking the case of our statistical programs, "Sensitivity Analysis on the Web" (Yamanishi, 1999) and "VASPCA (VAriable Selection in Principal component Anaysis)" (Iizuka et al., 1999; Mori and Iizuka, 1999; Mori et al., 1999, 2000) as examples.

2. Using WWW for statistics

We can find web pages for statistics classified into the following categories: on-line programs for statistical computing; on-line textbooks for statistics; archives of data, software and references; on-line library; and links to related topics. Moreover the statistical consulting on web and the on-line symposium on statistical topics are now available.

What we want to focus on here is on-line programs for statistical computing. This category can be classified into further subcategories such as purpose, programming technique to be used, and types of implementation as show in Table 1.

We can observe two typical types of purposes; one is for education and the other data analysis. There exist large number of web sites for statistical education and it is easy to say that teaching or learning statistics by manipulating on-line programs are vary effective even if the programs were not constructed for education. While most of on-line programs are developed for computation in elementary statistics or popular (classical) multivariate methods, some of them are for analyzing data in special fields or providing computational tools for new theories.

Table 1: Aspects to examine on-line programs

Purpose of development	Education	(pages developed for education) (statistical tools available for education)	
	Data analysis	(for trial / for spread of idea and method) (providing full-scale analysis)	
Computational environment and programming technique	Server-side computation		CGI (+script language / +statistical engine)
	Client-side computation		Java applet (for interface) (programs on client machine)
			Java script
			Others (VRML, XML, Original interface, etc.)

On the other hand, we can consider on-line programs from programming aspects. That is, we focus on where the computation is performed, in a server or in a client, and which programming techniques are utilized; CGI, Java applet, Java script, and/or others (VRML, XML, etc.). This aspect strictly depends on the computational environment, how to post the data set and how to provide the results.

Though downloadable packages including add-ins for the spreadsheet software and libraries for some statistical packages can be categorized into client-side computation, we remove them from the subjects of our consideration.

3. Examples: Development of on-line programs

One of our purposes of developing on-line programs is to spread the concept, methods and ideas of some particular statistical methods; that is, we want to provide tools for many analysts to use the statistical methods which have not been widely used yet. The reasons why we made such tools on WWW are to use the methods at any time and from anywhere, not to reduce the client performance, and to update the programs promptly. Based on consideration in the previous section, our programs are categorized in Data analysis for purpose aspect and server-side programming with CGI for programming aspect.

3.1. Sensitivity Analysis on the Web

The purpose of sensitivity analysis is to evaluate the stability or reliability of the results of analysis or to detect so-called influential observations. That is to examine whether there exist any subsets of observations on which the obtained results depend heavily. In our approaches we mainly use the influence function as a mathematical tool to find candidates of influential observations. The empirical influence function (*EIF*) is computed for each parameter of the analysis and its value indicates how large influence a particular observation has. We usually summarize the *EIF* vector into some scalar measures to evaluate the influence of a single observation.

For practical application it is desirable that a means of performing sensitivity analysis is provided. Then we started to develop two programs for sensitivity analysis. One is a statistical package “SAMMIF (Sensitivity Analysis in Multivariate Method based on Influence Functions)” which runs on Windows and provides information to detect not only singly but also jointly influential observations and to confirm whether their influence is really large or not by comparing the results for the sample with and without the specified observations (Mori et al., 1998). The current version (1.00) can be downloaded from, e.g., http://www.f7.ems.okayama-u.ac.jp/sammif/.

The other is an on-line program which is now being constructed (Yamanishi, 1999). You can try sensitivity analysis on http://august.f6.ems.okayama-u.ac.jp/~yoshi/sensitivity/. The present version is just a prototype to perform sensitivity analysis in principal component analysis (PCA), but it provides an easy way for sensitivity analysis on WWW.

(1) Execution

At first you put a data set in the data input form and input the number of rows and columns of the data in the boxes (Figure 1). After selecting a needed output and clicking the [execute] button, the corresponding results based on the influence function will be obtained. Figure 2 is a numerical result and Figure 3 is an index plot of the summarized *EIF* of the first eigenvector.

(2) On-line computation

- All the computation codes are written in Pearl script.
- The Interface is provided by HTML and CGI.

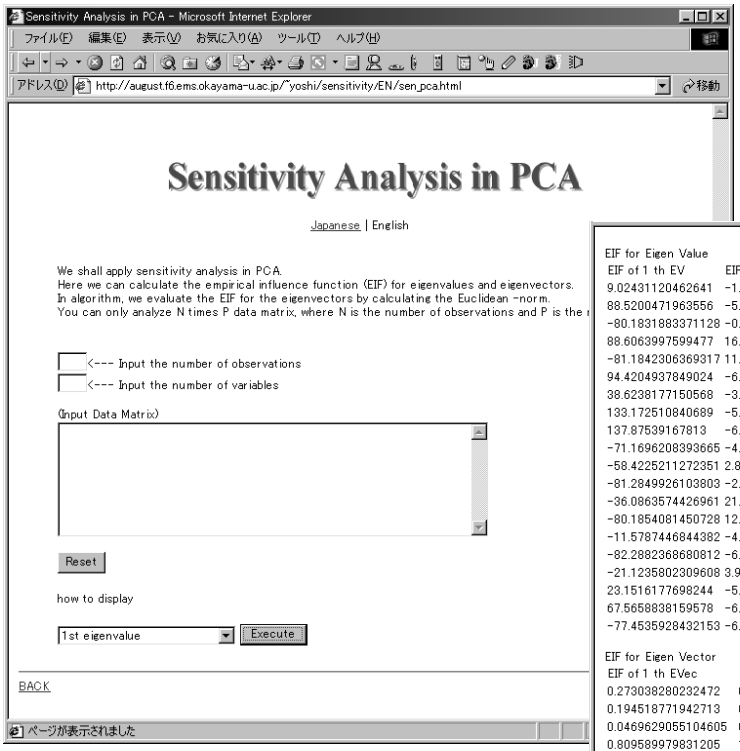

Figure 1: First page of Sensitivity Analysis in PCA

EIF for Eigen Value

EIF of 1 th EV	EIF of 2 th EV	EIF of 3 th EV	EIF of 4 th EV
9.02431120462641	-1.88743180718841	-0.103672283709602	-0.221788404107955
88.5200471963556	-5.76467915294062	0.0524170731375816	-0.244355654008485
-80.1831883371128	-0.307455573433131	-0.396546332113849	-0.137369902311232
88.6063997599477	16.3340144464705	0.031098294212546	0.680025849650158
-81.1842306369317	11.1825666250456	-0.443233260851087	-0.181663557696032
94.4204937849024	-6.72396509882906	-0.443365629818383	-0.100716848137909
38.6238177150568	-3.82275611323945	-0.371604266759303	-0.244965452433672
133.172510840689	-5.7768387044846	-0.394297330344184	0.0791033998931682
137.87539167813	-6.50420354359994	-0.0113618904010113	0.146597800771618
-71.1696208393665	-4.29347756832044	-0.320888879000778	0.250446746133079
-58.4225211272351	2.81175948328325	-0.372390498133034	-0.245373371047157
-81.2849926103803	-2.36931833330779	1.01484587043134	-0.212057024961023
-36.0863574426961	21.1662834922605	0.173537189873163	-0.219114596078719
-80.1854081450728	12.632485400819	2.30210011681914	-0.0726712624532192
-11.5787446844382	-4.82901557588354	-0.200590477405053	-0.236227007811447
-82.2882368680812	-6.49928238014168	-0.0489768246174815	0.0489411592164016
-21.1235802309608	3.96315468092634	0.272953711415751	0.0628593136984513
23.1516177698244	-5.99787755552571	-0.0504901417599315	0.3302242242305
67.5658838159578	-6.70089616185228	-0.436958522987584	0.0974242518302811
-77.4535928432153	-6.61306656005841	-0.252575917988234	0.420680335623194

EIF for Eigen Vector

EIF of 1 th EVec	EIF of 2 th EVec	EIF of 3 th EVec	EIF of 4 th EVec
0.273038280232472	0.332510084815246	0.471519525758018	0.427526680504701
0.194518771942713	0.193489882838448	0.188189861088108	0.11333349382823
0.0469629055104605	0.156748904362574	0.360285097776422	0.370386843114563
0.809589979831205	1.15424661615877	3.16808831954633	3.20188164620339
0.0565153356243764	0.1719240299594	0.0913859031100107	0.175993914342265
0.0628743497199958	0.0215683803577588	0.119914898303705	0.133219885214337
0.238709726040163	0.24664687065332	0.08505123897678	0.0310965468826655

Figure 2: Numerical result

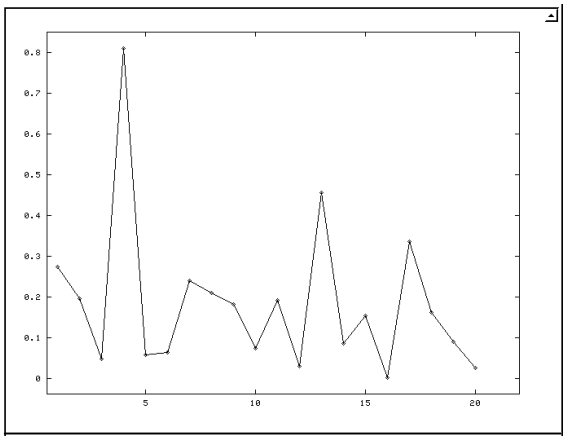

Figure 3: Index plot (1st Eigenvector)

3.2. VASPCA (VAriable Selection in PCA)

Consider a situation where we wish to select subsets of variables in the context in PCA. There are various methods and criteria to select q variables among p original ones as shown in Table 2 ($1 \le q \le p$). These existing methods and criteria often provide the different results (selected subsets of variables) from each other. In practical application of variable selection, it is necessary to apply a method suitable for the purpose of selection and/or to try some methods and select one by comparing the results.

So far, however, we had no device to perform any method easily. Thus we are developing statistical software “VASPCA (VAriable Selection in PCA)” which contains a variety of selection methods and criteria to select a subset of variables in PCA. It has two versions, an on-line program VASPCA/Web and off-line program VASPCA/Win (Windows package).

Table 2: Classification of methods and criteria of variable selection in PCA

Method	Criterion		
	Variation	Closeness of configurations	Others
Jolliffe's B2	-	-	loadings (S to L)
Jolliffe's B4	-	-	loadings (L to S)
McCabe	partial covariance	-	-
Falguerolles & Jmel	Gaussian model	-	-
Krzanowski	-	Procrastes analysis	-
Robert & Escoufier	-	*RV*-coefficient	-
Bonifas	-	*RV*-coefficient	-
Tanaka & Mori	proportion	*RV*-coefficient	-
Senisitivity Analysis	-	-	influence of variable
PRESS	-	-	*PRESS*
M.Reg	multiple correlation	-	-
Cluster analysis	-	-	cluster (subjectively)

As in Figure 4, the web page of VASPCA at http://face.f7.ems.okayama-u.ac.jp/~masa/vaspca/indexE.html or http://mo161.soci.ous.ac.jp/vaspca/indexE.html consists of three parts: general information on variable selection in PCA, the programs supplies (VASPCA/Web and VASPCA/Win) and the related information. At the VASPCA/Web execution pages in the programs supplies you can try and do any selection method supplied there.

(1) Execution

Before selection you should create a data file in your local disk and determine the number of principal components (PCs) r by applying PCA to your data ($1 \leq r \leq q$). In current version of VASPCA/Web you can choose one selection method among (A) selection using criteria in Modified PCA; (B) selection using the ideas of Principal Variables; (C) selection using Procrustes analysis; (D) selection using *RV*-coefficient; (E) selection using Influence analysis of variables; (F) selection using the idea of prediction error; or (G) selection focusing loadings. After choice of method you proceed to the execution pages consisting of three pages. In Page 1 you enter your data with ID name from your local disk (Figure 5). In Page 2 you specify the following selection parameters, the number of PCs r (except (G)), selection criterion corresponding to the chosen method and selection procedure among Backward, Forward, Backward-forward or Forward-backward (Figure 6). The result of selection will be displayed after a short time in Page 3 (Figure 7). You can also obtain a graph of change of criterion values and a plane text of result on your demand.

(2) On-line computation

- "R" is used as a statistical engine. VASPCA/Web calls R functions with the data using CGI. The reasons why we use R are that R has many statistical functions to make programming easy, that R is a free package, and that R works on multi-platform.
- The Interface is provided by HTML and CGI.
- The system requires some restrictions to the data such as data format and data size.
- The result is stored in our server as an HTML file. So the user can move to the other pages before computation is done and see at any time after the computation.

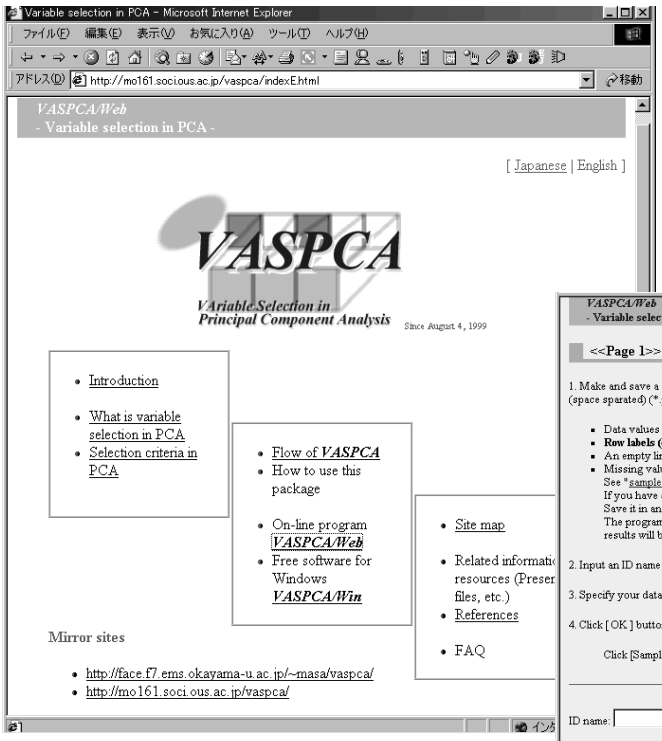

Figure 4: Top page of VASPCA on WWW

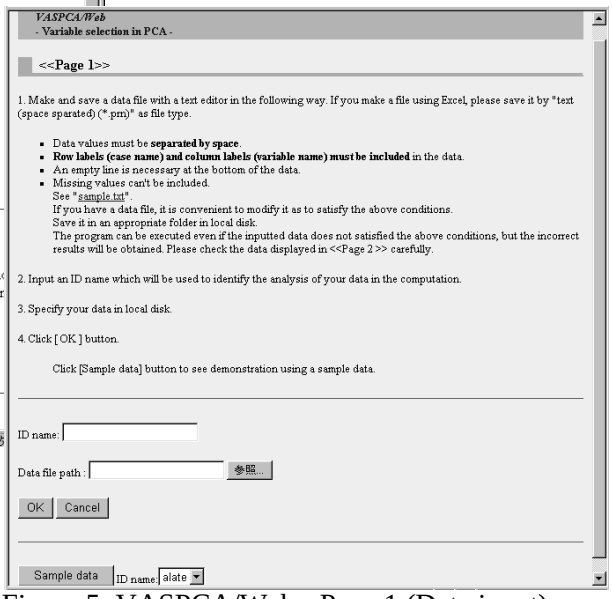

Figure 5: VASPCA/Web - Page 1 (Data input)

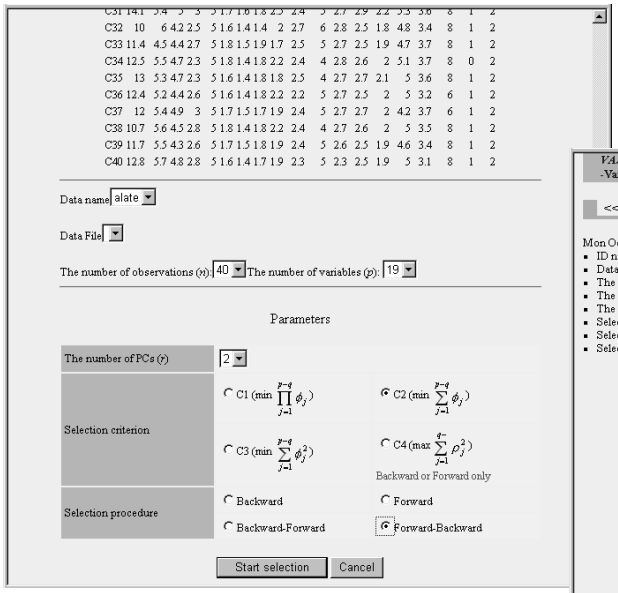

Figure 6: VASPCA/Web - Page 2
(Specification of parameters for selection; C2 in (B) selection using the ideas of Principal Variables and Forward-Backward)

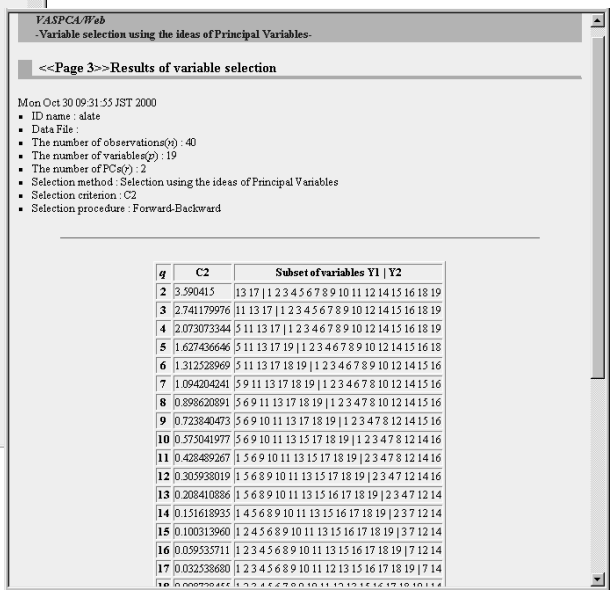

Figure 7: VASPCA/Web - Page 3
(Result - subsets of variables and criterion values)

4. Concluding remarks

We can provide useful statistical tools for sensitivity analysis and variable selection in PCA. In our trials, we tried to make full-scale systems on WWW to spead the new ideas as well as to provide the tools to perform the analyses. We chose server-side programming not to reduce the performance of the user's machine and made systems with CGI because of its easiness to modify programs and construct HTML pages.

In our experiences we still have the problems to be solved in developing an on-line program: which method is taken, developing the entire parts of program by ourselves or utilizing some statistical engine; how to protect the data and the results from unconcerned persons; how large the size of data to be handled is; how to provide more effective interface including a new statistical environment through WWW.

References:

Iizuka, M., Mori, Y., Tarumi, T. and Tanaka, Y. (1999). Variable selection program "VASPCA/Web". *Proceedings of the 13th symposium of Japanese Society of Computational Statistics*, 60-63.

Mori, Y and Iizuka, M. (1999). VASPCA (Variable Selection in Principal Component Analysis), http://face.f7.ems.okayama-u.ac.jp/~masa/vaspca/.html, http://mo161.soci.ous.ac.jp/vaspca/

Mori, Y., Iizuka, M. Tarumi, T. and Tanaka, Y. (1999). Variable selection in "Principal Component Analysis Based on a Subset of Variables", *Bulletin of the International Statistical Institute* (*52nd Session Contributed Papers Book2*), 333-334.

Mori, Y., Iizuka, M. Tarumi, T. and Tanaka, Y. (2000). Statistical Software "VASPCA" for Variable Selection in Principal Component Analysis, In *COMPSTAT2000 Proceedings in Computational Statistics* (*Short Communications*), 73-74.

Mori, Y., Watadani, S., Tarumi, T. & Tanaka, Y. (1998). Development of Statistical Software SAMMIF for Sensitivity Analysis in Multivariate Methods, In *COMPSTAT98 Proceedings in Computational Statistics* (ed. R.Payne & P.Green), 395-400. Heidelberg: Physica-Verlag.

Yamanishi, Y. (1999). Sensitivity Analysis on the Web, http://august.f6.ems.okayama-u.ac.jp/~yoshi/sensitivity/