Paper AS10

SAS/STAT High-performance procedures contrasted with their legacy counterparts

Roman Ganzha, GCE Solutions, Kyiv, Ukraine

ABSTRACT

Given the persistent trend of growth of clinical trial data and in computing capabilities, the SAS/STAT high-performance procedures (HPPs) were developed to provide extensive support in predictive modeling including regression, logistic regression, generalized linear models, linear mixed models, nonlinear models, and decision trees. A certain amount of information dedicated to functionality and commonly used options have already been published. However, there could not be awareness on such procedures per se and on the difference between HPPs and their SAS/STAT legacy counterparts.

INTRODUCTION

This paper serves as an overview of HPPs contrasted with legacy procedures. The paper makes awareness on such procedures (SAS/STAT 12.3 User's Guide: High-Performance Procedures; SAS/STAT 14.3 User's Guide: High-Performance Procedures) and on the differences between key HPPs and their legacy counterparts in terms of capabilities and functionality (SAS/STAT 12.3 User's Guide: High-Performance Procedures; SAS/STAT 14.3 User's Guide: High-Performance Procedures).

HPCANDISC COMPARED WITH CANDISC

The HPCANDISC procedure is a high-performance procedure that performs canonical discriminant analysis. Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. The methodology that is used in deriving the canonical coefficients parallels that of a one-way multivariate analysis of variance (MANOVA). The HPCANDISC procedure performs a canonical discriminant analysis, computes squared Mahalanobis distances between class means, and performs both univariate and one-way multivariate analyses of variance. Two output data sets can be produced: one that contains the canonical coefficients and another that contains, among other things, scored canonical variables. It is customary to standardize the canonical coefficients so that the canonical variables have means that are equal to 0 and pooled within-class variances that are equal to 1. PROC HPCANDISC displays both standardized and unstandardized canonical coefficients. Correlations between the canonical variables and the original variables in addition to the class means for the canonical variables are also displayed; these correlations, sometimes known as loadings, are called canonical structures (SAS/STAT 14.3 User's Guide: High-Performance Procedures).

Some differences between HPCANDISC procedure and the CANDISC procedure in SAS/STAT are displayed below

What	HPCANDISC	CANDISC
Input	Only the ordinary SAS data set	Ordinary SAS data set and other types of special SAS data sets
ID statement	Available	Not available

HPFMM CONTRASTED WITH FMM

The HPFMM procedure is designed to fit finite mixtures of regression models or finite mixtures of generalized linear models in which the covariates and regression structure can be the same across components or can be different. You can fit finite mixture models by maximum likelihood or Bayesian methods.

One can fit the same kinds of models and get the same kinds of tabular, graphical, and data set results from PROC HPFMM as from PROC FMM (SAS/STAT 14.3 User's Guide: High-Performance Procedures). PROC HPFMM and PROC FMM use the same statements. However, PROC HPFMM includes more options compared to the legacy procedure in the CLASS and PERFORMACE statements. More detailed comparisons are presented on the SAS/STAT 14.3 User's Guide: High-Performance Procedures.

GAMPL CONTRASTED WITH GAM

The GAMPL procedure is a high-performance procedure that fits generalized additive models that are based on low-rank regression splines providing powerful tools for nonparametric regression and smoothing.

Generalized additive models are extensions of generalized linear models. They relax the linearity assumption in generalized linear models by allowing spline terms in order to characterize nonlinear dependency structures. Each spline term is constructed by the thin-plate regression spline technique. A roughness penalty is applied to each spline term by a smoothing parameter that controls the balance between goodness of fit and the roughness of the spline curve.

The difference between GAMPL and GAM procedures include using slightly different approaches for constructing spline basis expansions, fitting generalized additive models and testing smoothing components (SAS/STAT 14.3 User's Guide: High-Performance Procedures). In addition, PROC GAMPL presents more options on Distribution Families and Link Functions compared to legacy procedure (e.g. the negative binomial family). More detailed comparisons are presented on the SAS/STAT 14.3 User's Guide: High-Performance Procedures.

HPGENSELECT CONTRASTED WITH GENMOD

The HPGENSELECT procedure is a high-performance procedure that provides model fitting and model building for generalized linear models. Some differences between HPGENSELECT procedure and the legacy procedure in SAS/STAT are displayed below:

What	HPGENSELECT	GENMOD
CLASS STATEMENT	2 parameterizations: GLM and reference	13 parameterizations
Optimization Techniques	8 options	Newton-Raphson optimization with ridging
Post fitting analyses	wide variety	limited
BAYES statement	Not available	Available
EXACT statement	Not available	Available
PARTITION statement	Available (creation of training and testing datasets)	Not available

Of note, the EXACT statements and the BAYES statement are not available on the HPP. It is not ruled out these may be the more computational demanding statements.

HPNLMOD CONTRASTED WITH THE NLIN AND NLMIXED

The HPNLMOD procedure is a high-performance procedure that uses either nonlinear least squares or maximum likelihood to fit nonlinear regression models. PROC HPNLMOD enables you to specify the model by using SAS programming statements, which give you greater flexibility in modeling the relationship between the response variable and independent (regressor) variables than do SAS procedures that use a more structured MODEL statement (SAS/STAT 14.3 User's Guide: High-Performance Procedures).

Some differences between HPNLMOD procedure and the legacy procedure in SAS/STAT are displayed below:

What	HPNLMOD	NLIN/NLMIXED

Optimization techniques	Different methods compared to legacy procedure	
BOOTSTRAP Statement	Not available	Available (NLIN)
CMPTMODEL Statement*	Not available	Available (NLMIXED)
MODEL Statement	8 models (RESIDUAL or LS as new model)	7 models (NLMIXED)
RANDOM Statement	Not available	Available (NLMIXED)

* The CMPTMODEL statement computes predicted concentrations from a specified one-, two-, or three-compartment model

Of note, the BOOTSTRAP statement is not available on the HPP. It is not ruled out this may be the more computational demanding statement.

HPPLS CONTRASTED WITH PLS

The HPPLS procedure is a high-performance version of the PLS procedure in SAS/STAT software, which fits models by using any one of a number of linear predictive methods, including partial least squares (PLS). For additional details on PLS procedure see the SAS/STAT 14.3 User's Guide.
Some differences between HPPLS procedure and the legacy procedure in SAS/STAT are displayed below:

What	HPPLS	PLS
Algorithm techniques	3 techniques (RLGW not available)	4 techniques
EFFECT Statement*	Not available	Available

* The EFFECT statement enables you to construct special collections of columns for design matrices (e.g. splines/polynomials).

Additional details on the comparison between the two procedures may be found in the SAS/STAT 14.3 User's Guide: High-Performance Procedures

HPPRINCOMP CONTRASTED WITH PRINCOMP

The HPPRINCOMP procedure is a high-performance procedure that performs principal component analysis.
It is a high-performance version of the PRINCOMP procedure in SAS/STAT software, but it provides additional iterative methods to calculate the principal components.

Principal component analysis is a multivariate technique for examining relationships among several quantitative variables, providing an optimal way of reducing dimensionality by projecting the data onto a lower-dimensional orthogonal subspace that explains as much variation in those variables as possible.
Some differences between HPPRINCOMP procedure and the legacy procedure in SAS/STAT are displayed below:

What	HPPRINCOMP	PRINCOMP
PLOTS option	Not available	Available
Method Techniques	EIG | ITERGS<(iter-options)> | NIPALS<(iter-options)>	eigenvalue decomposition
CODE Statement*	Available	Not available
OUTPUT Statement**	Available	Not available

* The CODE statement enables you to write SAS DATA step code for computing the principal component scores either to a file or to a catalog entry.

** The OUTPUT statement creates a data set that contains observationwise statistics

HPQUANTSELECT CONTRASTED WITH QUANTSELECT

The HPQUANTSELECT procedure is a high-performance procedure that fits and performs effect selection for quantile regression analysis.

Some differences between HPQUANTSELECT procedure and the legacy procedure in SAS/STAT are displayed below:

What	HPQUANTSELECT	QUANTSELECT
Algorithm option	Predictor-corrector interior point algorithm, which was originally designed to solve support vector machine classifiers for large data sets	SIMPLEX | SMOOTH
EFFECT Statement*	Not available	Available
Confidence Intervals	Produced	Not directly produced

* The EFFECT statement enables you to construct special collections of columns for design matrices (e.g. splines/polynomials).

Additional details on the comparison between the HPP procedure and some other legacy procedures may be found in the SAS/STAT 14.3 User's Guide: High-Performance Procedures

HPREG CONTRASTED WITH GLMSELECT, GLM AND REG

The HPREG procedure is a high-performance procedure that fits and performs model selection for ordinary linear least squares models.

Some differences between HPQUANTSELECT procedure and the legacy procedure in SAS/STAT are displayed below:

What	HPREG	REG/GLM

Ridge option	Not available	Available (REG)
MANOVA Statement	Not available	Available (GLM)
MTEST Statement	Not available	Available (REG)
RANDOM Statement	Not available	Available (GLM)
REWEIGHT/REFIT statement	Not available	Available (REG)

Of note, the Multivariate analysis statements are not available on the HPP. It is not ruled out these may be the more computational demanding statement.

CONCLUSION

Where computational power is required, HPPs may provide an alternative solution to SAS/STAT base procedure.

However, some considerations should be in place on whether to use the HPPs procedure compared to legacy procedures

What?	**Considerations**
License	If the company does not have a licence for HPPs, evaluate whether it is worth to have it
Size/complexity of datasets	Think about the maximum size of datasets you are dealing with and which type of procedures. Think about the transfer speed of data (the transfer of data may be the bottle neck rather than the execution time)
How frequently I use computational intensive procedures	This could be one additional driver
Some computational intensive statements may not be available in HPPS	This could be one additional driver

Future work may include the evaluation of the performance (e.g. time to complete the procedure) in datasets with increasing complexity (variables, observations) and/or when increasing the number of repetitions of a task (e.g. simulations) and/or when using different estimators.

ACKNOWLEDGMENT

Oscar Bortolami as reviewer of this paper.

REFERENCES

SAS/STAT 12.3 User's Guide: High-Performance Procedures
http://support.sas.com/documentation/cdl/en/stathpug/66410/HTML/default/viewer.htm#titlepage.htm

SAS/STAT 14.3 User's Guide: High-Performance Procedures
https://documentation.sas.com/?docsetId=stathpug&docsetTarget=titlepage.htm&docsetVersion=14.3&locale=en

SAS/STAT 14.3 User's Guide
https://documentation.sas.com/?docsetId=statug&docsetTarget=titlepage.htm&docsetVersion=14.3&locale=en

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Roman Ganzha
GCE Solutions
25-B, P. Sagaydachnogo street
Kyiv, Ukraine, 04070
ganzha.roman.alex@gmail.com
www.gcesolutions.com