SANDIA REPORT

SAND2009-1687
Unlimited Release
Printed March 2009

Scalable Multi-Correlative Statistics and Principal Component Analysis with Titan

Janine Bennett, Philippe Pébay, Diana Roe, David Thompson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2009-1687
Unlimited Release
Printed March 2009

Scalable Multi-Correlative Statistics and Principal Component Analysis with Titan

Janine Bennett
Sandia National Laboratories
M.S. 9159, P.O. Box 969
Livermore, CA 94551, U.S.A.
jcbenne@sandia.gov

Philippe Pébay
Sandia National Laboratories
M.S. 9159, P.O. Box 969
Livermore, CA 94551, U.S.A.
pppebay@sandia.gov

Diana Roe
Sandia National Laboratories
M.S. , P.O. Box 969
Livermore, CA 94551, U.S.A.
dcroe@sandia.gov

David Thompson
Sandia National Laboratories
M.S. 9159, P.O. Box 969
Livermore, CA 94551, U.S.A.
dcthomp@sandia.gov

Abstract

This report summarizes existing statistical engines in VTK/Titan and presents the recently parallelized multi-correlative and principal component analysis engines. It is a sequel to [PT08] which studied the parallel descriptive and correlative engines. The ease of use of these parallel engines is illustrated by the means of C++ code snippets. Furthermore, this report justifies the design of these engines with parallel scalability in mind; then, this theoretical property is verified with test runs that demonstrate optimal parallel speed-up with up to 200 processors.

Acknowledgments

The authors would like to thank:

- Brian Wylie, for his comments on the integration of scalable statistical tools in VTK/Titan,
- Jackson Mayo, for his precursor work on a serial version of the multi-correlative statistics algorithm.

Contents

1 Introduction 7
1.1 The Titan Informatics Toolkit 7
1.2 Statistics Functionality in Titan 8
2 Parallel Statistics Classes 12
2.1 Implementation Details 12
2.2 Usage 13
3 Results 16
3.1 Algorithm Scalability 16
3.2 Algorithm Correctness 21
References 24

This page intentionally left blank

1 Introduction

This report is a sequel to [PT08], which focused on the parallel descriptive and bivariate statistics engines; please refer to this reference for a detailed presentation of these engines as well as an assessment of their scalability and speed-up properties.

1.1 The Titan Informatics Toolkit

The Titan Informatics Toolkit is a collaborative effort between Sandia National Laboratories and Kitware Inc. It represents a significant expansion of the Visualization ToolKit (VTK) to support the ingestion, processing, and display of informatics data. By leveraging the VTK engine, Titan provides a flexible, component based, pipeline architecture for the integration and deployment of algorithms in the fields of intelligence, semantic graph and information analysis.

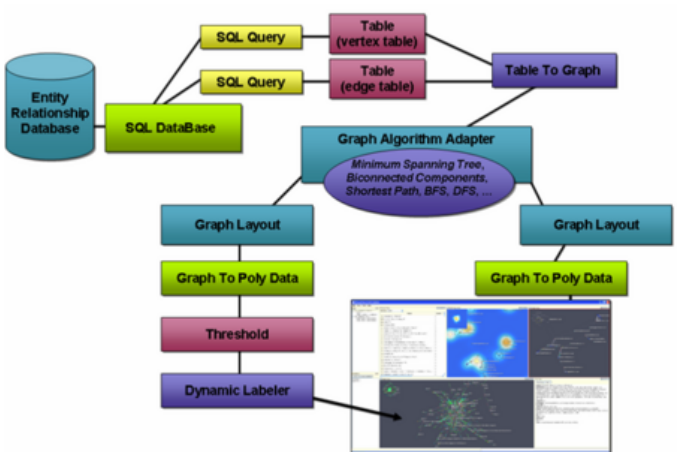

Figure 1. A theoretical application built with Titan.

A theoretical application built from Titan/VTK components is schematized in Figure 1. The flexibility of the pipeline architecture allows effective utilization of the Titan components for different problem domains. An actual implementation is OverView, a generalization of the ParaView scientific visualization application to support the ingestion, processing, and display of informatics data. The ParaView client-server architecture provides a mature framework for performing scalable analysis on distributed memory platforms, and OverView will use these capabilities to analyze informatics problems that are too large for individual workstations.

The Titan project represents one of the first software development efforts to address the merging of scientific visualization and information visualization on a substantive level. The VTK parallel client-server layer will provide an excellent framework for doing scalable analysis on distributed memory platforms. The benefits of combining the two fields are already reaping rewards in the form of functionality such as the cell lineage application below.

1.2 Statistics Functionality in Titan

A number of univariate, bivariate, and multivariate statistical tools have been implemented in Titan. Each tool acts upon data stored in one or more tables; the first table serves as observations and further tables serves as model data. Each row of the first table is an observation, while the form of further tables depends on the type of statistical analysis. Each column of the first table is a variable.

1.2.1 Variables

A univariate statistics algorithm only uses information from a single column and, similarly, a bivariate from 2 columns. Because an input table may have many more columns than an algorithm can make use of, Titan must provide a way for users to denote columns of interest. Because it may be more efficient to perform multiple analyses of the same type on different sets of columns at once as opposed to one after another, Titan provides a way for users to make multiple analysis requests of a single filter.

Table 1. A table of observations that might serve as input to a statistics algorithm.

row	A	B	C	D	E
1	0	1	0	1	1.03315
2	1	2	2	2	0.76363
3	0	3	4	6	0.49411
4	1	5	6	24	0.04492
5	0	7	8	120	0.58395
6	1	11	10	720	1.66202

As an example, consider Table 1. It has 6 observations of 5 variables. If the correlations between A, B, and C, and also between B, C and D are desired, two requests, R_1 and R_2 must be made. The first request R_1 would have columns of interest $\{A,B,C\}$ while R_2 would have columns of interest $\{B,C,D\}$. Calculating covariances for R_1 and R_2 in one pass is more efficient than computing each separately since $\text{cov}(B,B)$, $\text{cov}(C,C)$, and $\text{cov}(B,C)$ are required for both requests but need only be computed once.

1.2.2 Phases

Each statistics algorithm performs its computations in a sequence of common phases, regardless of the particular analysis being performed. These phases can be described as:

Learn: Calculate a "raw" statistical model from an input data set. By "raw", we mean the minimal representation of the desired model, that contains only primary statistics. For example, in

the case of descriptive statistics: sample size, minimum, maximum, mean, and centered M_2, M_3 and M_4 aggregates (*cf.* [Pó8]). For Table 1 with a request $R_1 = \{B\}$, these values are 6, 1, 11, $4.8\bar{3}$, $68.8\bar{3}$, $159.\bar{4}$, and $1759.819\bar{4}$, respectively.

Derive: Calculate a "full" statistical model from a raw model. By "full", we mean the complete representation of the desired model, that contains both primary and derived statistics. For example, in the case of descriptive statistics, the following derived statistics are calculated from the raw model: unbiased variance estimator, standard deviation, and two estimators (g and G) for both skewness and kurtosis. For Table 1 with a request $R_1 = \{B\}$, these additional values are $13.7\bar{6}$, 3.7103, 0.520253, 0.936456, -1.4524, and -1.73616 respectively.

Assess: Given a statistical model – from the same or another data set – mark each datum of a given data set. For example, in the case of descriptive statistics, each datum is marked with its relative deviation with respect to the model mean and standard deviation (this amounts to the one-dimensional Mahalanobis distance). Table 1 shows this distance for $R_1 = \{B\}$ in column E.

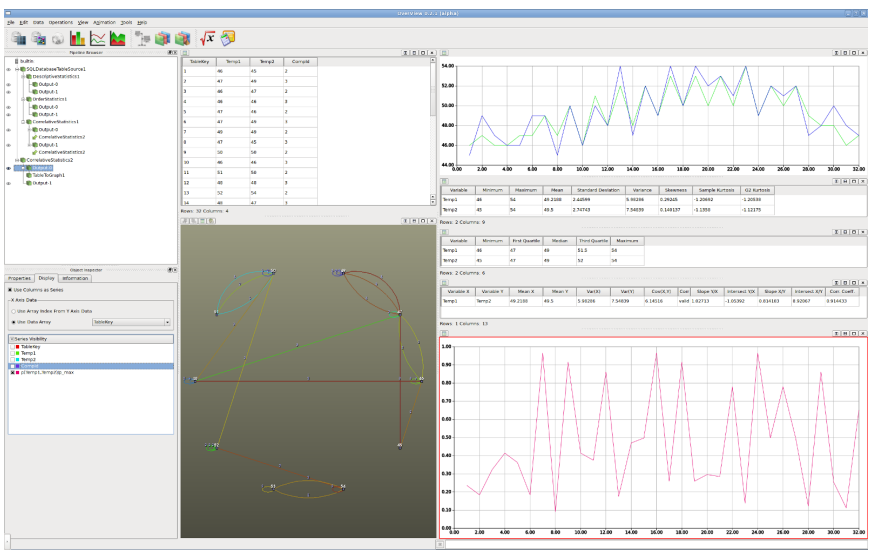

Figure 2. An example utilization of Titan's statistics algorithms in OverView.

An example of the utilization of Titan's statistical tools in OverView is illustrated in Figure 2; specifically, the descriptive, correlative, and order statistics classes are used in conjunction with various table views and plots. With the exception of contingency statistics which can be performed on any type (nominal, cardinal, or ordinal) of variables, all currently implemented algorithms require cardinal or ordinal variables as inputs.

At the time of writing, the following algorithms are available in Titan:

1. Univariate statistics:

(a) Descriptive statistics:

Learn: calculate minimum, maximum, mean, and centered M_2, M_3 and M_4 aggregates;

Derive: calculate unbiased variance estimator, standard deviation, skewness (1_2 and G_1 estimators), kurtosis (g_2 and G_2 estimators);

Assess: mark with relative deviations (one-dimensional Mahlanobis distance).

(b) Order statistics:

Learn: calculate histogram;

Derive: calculate arbitrary quartiles, such as "5-point" statistics (quartiles) for box plots, deciles, percentiles, etc.;

Assess: mark with quartile index.

2. Bivariate statistics:

(a) Correlative statistics:

Learn: calculate minima, maxima, means, and centered M_2 aggregates;

Derive: calculate unbiased variance and covariance estimators, Pearson correlation coefficient, and linear regressions (both ways);

Assess: mark with squared two-dimensional Mahlanobis distance.

(b) Contingency statistics:

Learn: calculate contingency table;

Derive: calculate joint, conditional, and marginal probabilities, as well as information entropies;

Assess: mark with joint and conditional PDF values.

3. Multivariate statistics:

These filters all accept multiple requests R_i, each of which is a set of n_i variables upon which simultaneous statistics should be computed.

(a) Multi-Correlative statistics:

Learn: calculate means and pairwise centered M_2 aggregates;

Derive: calculate the upper triangular portion of the symmetric $n_i \times n_i$ covariance matrix and its (lower) Cholesky decomposition;

Assess: mark with squared multi-dimensional Mahlanobis distance.

(b) PCA statistics:

Learn: identical to the multi-correlative filter;

Derive: everything the multi-correlative filter provides, plus the n_i eigenvalues and eigenvectors of the covariance matrix;

Assess: perform a change of basis to the principal components (eigenvectors), optionally projecting to the first m_i components, where $m_i \leq n_i$ is either some user-specified value or is determined by the fraction of maximal eigenvalues whose sum is above a user-specified threshold. This results in m_i additional columns of data for each request R_i.

In the following sections, we present implementation details on the parallel versions of the descriptive, correlative, multi-correlative, and PCA statistics algorithms, provide a basic user manual of these, and examine their correctness as well as their parallel speed-up properties.

2 Parallel Statistics Classes

2.1 Implementation Details

The purpose of building a full statistical model in two phases is parallel computational efficiency. In our approach, inter-processor communication and updates are performed only for primary statistics. The calculations to obtain derived statistics from primary statistics are typically fast and simple and need only be calculated once, without communication, upon completion of all parallel updates of primary variables. Data to be assessed is assumed to be distributed in parallel across all processes participating in the computation, thus no communication is required as each process assesses its own resident data.

Therefore, in the parallel versions of the statistical engines, inter-processor communication is required only for the Learn phase, while both Derive and Assess are executed in an embarrassingly parallel fashion due to data parallelism. This design is consistent with the data parallelism methodology used to enable parallelism within VTK, most notably in ParaView. Because the focus of this report is on the parallel speed-up properties of statistics engines, we will not report on the Derive or Assess phases, as these are executed independently from each other, on a separate process for each part of the data partition. However, because the Derive phase provides the derived quantities to which one is naturally accustomed (e.g., variance as opposed to M_2 aggregate), the numerical results reported here are those that are yielded by the consecutive application of the Learn and then Derive phases.

At this point (March 2009) of the development of scalable statistics algorithms in Titan, the following 4 parallel classes are implemented:

1. `vtkPDescriptiveStatistics`;
2. `vtkPCorrelativeStatistics`;
3. `vtkPMultiCorrelativeStatistics`;
4. `vtkPPCAStatistics`.

Each of these parallel algorithms is implemented as a subclass of the respective serial version of the algorithm and contains a `vtkMultiProcessController` to handle inter-processor communication. Within each of the parallel statistics classes, the Learn phase is the only phase whose behavior is changed (by reimplementing its virtual method) due to the data parallelism inherent in the Derive and Assess phases. The Learn phase of the parallel algorithms performs two primary tasks:

1. Calculate correlative statistics on local data by executing the Learn code of the superclass.
2. If parallel updates are needed (i.e. the number of processes is greater than 1), perform necessary data gathering and aggregation of local statistics into global statistics.

The descriptive, correlative and multi-correlative statistics algorithms perform the aggregation necessary for the statistics which they are computing using the arbitrary-order update and covariance update formulas presented in [PÓ8]. Because the PCA statistics class derives from the multi-correlative statistics algorithm and inherits its Learn phase, we define a static method within the parallel multi-correlative statistics algorithm to gather all necessary statistics. This function is called from within both the parallel multi-correlative and PCA filter Learn phases to maximize code reuse.

2.2 Usage

It is fairly easy to use the serial statistics classes of Titan; it is not much harder to use their parallel versions. All that is required is a parallel build of Titan and a version of MPI installed on your system.

For example, Listing 1 demonstrates how to calculate descriptive statistics, in parallel, on each column of an input set `inputData` of type `vtkTable*`, with no subsequent data assessment. Note that if, instead, the Assess phase were turned on with `pds->SetAssess( true )` then, by default, unsigned deviations (Mahalanobis distance) would be calculated. To obtained signed deviations, `pds->SignedDeviationsOn()` should be called. The option to elect signed deviations as opposed to the Mahlanobis distance for the assessment of data is available only for descriptive statistics as the concepts of left and right do not extend to dimensions higher than one.

In Listing 1, requests for each column of interest are specified by calling `AddColumn()`, as is done for all univariate algorithms. Bivariate algorithms need only call `AddColumn()` an even number of times to unambiguously specify a set of requests. However, multivariate filters have a slightly different usage pattern. In order to queue a request for multivariate statistics algorithms, `SetColumnStatus()` should be called to turn on columns of interest (and to turn off any previously-selected columns that are no longer of interest). Once the desired set of columns has been specified, a call to `RequestSelectedColumns()` should be made. Consider the example from §1.2.1 and Table 1 where 2 requests are mentioned: $\{A,B,C\}$ and $\{B,C,D\}$. The code snippet in Listing 2 shows how to queue these requests for a `vtkPPCAStatistcs` object.

The examples thus far assume that you have already prepared an MPI communicator, loaded a dataset into the `inputData` object, and are running in a parallel environment. It is outside the scope of this report to discuss I/O issues, and in particular how a `vtkTable` can be created and filled with the values of the variables of interest. See VTK's online documentation for details [vtk].

In the code example from Listing 1, the `vtkMultiProcessController` object passed to `Foo()` is used to determine the set of processes (which may be a subset of a larger job) among which input data is distributed. VTK uses subroutines of this form to execute code across many processes. In Listing 3 we demonstrate that, to prepare a parallel controller to execute `Foo()` in parallel using MPI, one must first (e.g. in the main routine) create a `vtkMPIController` and pass it the address of `Foo()`. Note that, when using MPI, the number of processes is determined by the external program which launches the application.

```
void Foo( vtkMultiProcessController* controller, void* arg )
{
  // Use the specified controller on all parallel filters by default:
  vtkMultiProcessController::SetGlobalController( controller );

  // Assume the input dataset is passed to us:
  vtkTable* inputData = static_cast<vtkTable*>( arg );

  // Create parallel descriptive statistics class
  vtkPDescriptiveStatistics* pds = vtkPDescriptiveStatistics::New();

  // Set input data port
  pds->SetInput( 0, inputData );

  // Select all columns in inputData
  for ( int c = 0; c < inputData->GetNumberOfColumns(); ++ c )
    {
    pds->AddColumn( inputData->GetColumnName[c] );
    }

  // Calculate statistics with Learn and Derive phases only
  pds->SetLearn( true );
  pds->SetDerive( true );
  pds->SetAssess( false );
  pds->Update();
}
```

Listing 1: A subroutine – that should be run in parallel – for calculating descriptive statistics.

```
  vtkPPCAStatistics* pps = vtkPPCAStatistics::New();

  // Turn on columns of interest
  pps->SetColumnStatus( "A", 1 );
  pps->SetColumnStatus( "B", 1 );
  pps->SetColumnStatus( "C", 1 );
  pps->RequestSelectedColumns();

  // Columns A, B, and C are still selected, so first we turn off
  // column A so it will not appear in the next request.
  pps->SetColumnStatus( "A", 0 );
  pps->SetColumnStatus( "D", 1 );
  pps->RequestSelectedColumns();
```

Listing 2: An example of requesting multiple multi-variate analyses.

```
vtkTable* inputData;
vtkMPIController* controller = vtkMPIController::New();
controller->Initialize( &argc, &argv );

// Execute the function named Foo on all processes
controller->SetSingleMethod( Foo, &inputData );
controller->SingleMethodExecute();

// Clean up
controller->Finalize();
controller->Delete();
```

Listing 3: A snippet of code to show how to execute a subroutine (`Foo()`) in parallel. In reality, `inputData` would be prepared in parallel by `Foo()` but is assumed to be pre-populated here to simplify the example.

3 Results

The parallel runs have been executed on Sandia National Laboratories' `catalyst` computational cluster, which comprises 120 dual 3.06GHz Pentium Xeon compute nodes with 2GB of memory each. This cluster has a Gigabit Ethernet user network for job launch, I/O to storage, and user interaction with jobs, and a 4X Infiniband fabric high-speed network using a Voltaire 9288 InfiniBand switch. Its operating system has a Linux 2.6.17.11 kernel, and its batch scheduling system is the TORQUE resource manager [tor].

3.1 Algorithm Scalability

In order to assess speed-up independently of the load-balancing scheme, a series of (pseudo-) randomly-generated samples is used. Specifically, input tables are created at run time by generating 4 separate samples of independent pseudo-random variables, the two first (resp. last) variables having a standard normal (resp. standard uniform) distribution. Since our objective is to assess the scalability of the parallel statistics engines only, equally-sized slabs of data are created by each process in order to work with perfectly load-balanced cases. For the same reason, the amount of time needed to create the input data table is excluded from the analysis. In this test, `vtkPDescriptiveStatistics`, with Learn, Derive, and Assess modes on, is executed for each of the 4 columns, and the corresponding wall clock time is reported. Subsequently, `vtkPCorrelativeStatistics`, with Learn, Derive, and Assess modes turned on is executed on a single pair of columns (standard normal ones), and the corresponding wall clock time is also reported.

With these synthetic examples, we assess:

1. relative speed-up (at constant total work), and
2. scalability of the rate of computation (at constant work per processor).

3.1.1 Relative Speed-Up

Given a problem of size N (as measured in our case by sample size), the wall clock time measured to complete the work with p processors is denoted $T_N(p)$. Then, relative speed-up with p processors is

$$S_N(p) = \frac{T_N(1)}{T_N(p)}.$$

Evidently, optimal (linear) speedup is attained with p processors when $S_N(p) = p$ and, therefore, relative speed-up results for S_N may be visually inspected by plotting S_N *versus* the number of processors: optimal speed-up is revealed by a line, the angle bisector of the first quadrant.

Table 2. Relative speed-up (at constant total work), with a total sample size of $N = 25,600,000$.

N/p	p	Descriptive (sec. / $S_N(p)$)	Correlative (sec. / $S_N(p)$)	MC (8) (sec. / $S_N(p)$)	PCA (8) (sec. / $S_N(p)$)
25,600,000	1	69 / 1.00	55 / 1.00	53.4 / 1.00	65.9 / 1.00
12,800,000	2	35 / 1.99	28 / 2.00	26.9 / 1.98	33.5 / 1.97
6,400,000	4	18 / 3.93	15 / 3.80	14.4 / 3.70	16.8 / 3.93
3,200,000	8	9.0 / 7.65	7.0 / 7.92	6.9 / 7.73	8.4 / 7.81
1,600,000	16	4.6 / 14.91	3.7 / 15.01	3.7 / 14.60	4.7 / 13.90
800,000	32	2.3 / 29.95	1.8 / 30.51	1.8 / 29.85	2.2 / 29.95
400,000	64	1.3 / 54.43	0.9 / 58.13	1.0 / 54.52	1.1 / 59.89

In the first series of test runs, in order to assess relative speed-up, the a sample is subdivided into 4 columns of size $25,600,000$. Thus, the input data of the entire test case contains a total of $N = 102,400,000$ values. The values of p were chosen to be increasing powers of 2, for convenience only: making use of other values did not modify speed-up results. The results obtained on `catalyst` are provided in Tables 2 and plotted in Figure 3.

As expected based on the embarrassingly parallel nature of the algorithms, the measured relative speed-up is optimal (within $\pm 10\%$ fluctuations attributable to OS jitter and such), until total wall time measurements become too small to remain accurate (less than 1 sec.), and the decreasing amount of work per processor ultimately results in a situation where overheads, even small in absolute terms, become dominant as compared to the amount of actual computational work. In this current example, it appears that with 32 processors, minimal reliably measurable wall clock time has been or is almost reached. Note that this corresponds to a per processor load of $N/p = 3,200,000$ points per processor.

3.1.2 Rate of Computation Scalability

The rate of computation is defined as

$$r(p) = \frac{N(p)}{T_{N(p)}(p)},$$

where $N(p)$, the sample size, now varies with the number of processors p. We then measure its scalability by normalizing it with respect to the rate of computation obtained with a single processor, as follows:

$$R(p) = \frac{r(p)}{r(1)} = \frac{N(p)T_{N(1)}(1)}{N(1)T_{N(p)}(p)},$$

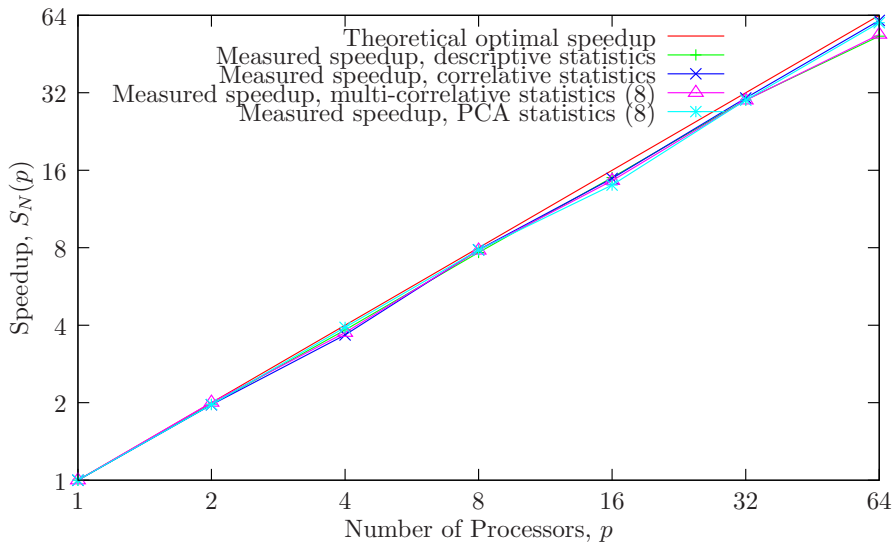

Figure 3. Relative speed-up at constant total work with a total data size of $N = 102,400,000$.

In particular, if the sample size is made to vary in proportion to the number of processors, i.e., if $N(p) = pN(1)$, then

$$R(p) = \frac{pT_{N(1)}(1)}{T_{pN(1)}(p)} = \frac{pT_{N(1)}(1)}{pT_{N(1)}(p)} = \frac{T_{N(1)}(1)}{T_{N(1)}(p)},$$

and thus, optimal (linear) scalability is also attained with p processors when $R(p) = p$. Note that without linear dependency between N and p, the latter equality no longer implies optimal scalability. Hence, under the above assumptions, scalability can also be visually inspected, with a plot of R *versus* the number of processors, where optimal scalability is also indicated by the angle bisector of the first quadrant. In order to assess rate of computation scalability (at constant work per processor), increasingly large samples are created, containing np quadruples, where $n = 10^6$ and $p \in \{1,2,4,8,16,32,64\}$ respectively denote the number of sample points per processor, and the number of processors, thus resulting in a total sample size of $N(p) = 4np$. Note that whether one or two cores per node are occupied by the np processes in each case is left for the scheduler to decide; forcing all cluster nodes to utilize either exactly one, or exactly two of their cores dit did not result in a measurable difference.

In each case, a table of size $n \times 4$ is created by each process. Corresponding wall clock times measured on `catalyst` are given in Table 3 (for multi-correlative statistics) and Table 4 (for PCA statistics), and plotted in Figure 4 ; these clearly exhibit optimal scalability (again within $\pm 10\%$ fluctuations attributable to OS jitter and such), thus experimentally verifying the embarrassingly parallel nature of these algorithms. It is also worth noting that using 1 or 2 cores per node did not result in any measurable difference.

Table 3. Rate of computation scalability (at constant load per processor), `vtkPMultiCorrelativeStatistics`.

$N(p)$ (million)	p	Correlative (sec. / R)	MC (4) (sec. / R)	MC (6) (sec. / R)	MC (8) (sec. / R)
4	1	9 / 1.00	9 / 1.00	13 / 1.00	17 / 1.00
8	2	9 / 2.01	9 / 1.98	13 / 2.00	17 / 2.02
16	4	9 / 4.11	10 / 3.89	13 / 3.93	17 / 3.89
32	8	9 / 8.09	10 / 7.68	13 / 7.81	18 / 7.48
64	16	10 / 13.7	10 / 15.5	14 / 14.21	18 / 14.76
128	32	9 / 31.0	10 / 30.0	14 / 28.37	20 / 27.05
256	64	9 / 61.2	11 / 52.6	14 / 57.26	20 / 53.21
512	128	10 / 117	11 / 112	14 / 115.7	19 / 112.60

Table 4. Rate of computation scalability (at constant load per processor), `vtkPPCAStatistics`.

$N(p)$ (million)	p	Correlative (sec. / R)	PCA (4) (sec. / R)	PCA (6) (sec. / R)	PCA (8) (sec. / R)
4	1	9 / 1.00	8 / 1.00	14 / 1.00	21 / 1.00
8	2	9 / 2.01	8 / 2.01	13 / 2.02	21 / 2.02
16	4	9 / 4.11	8 / 3.98	14 / 3.98	22 / 3.78
32	8	9 / 8.09	8 / 7.77	14 / 7.92	21 / 7.95
64	16	10 / 13.7	8 / 15.4	14 / 15.19	21 / 15.86
128	32	9 / 31.0	9 / 29.2	15 / 28.45	24 / 28.36
256	64	9 / 61.2	10 / 52.7	15 / 59.08	24 / 54.94
512	128	10 / 117	9 / 117	15 / 116.2	24 / 111.77

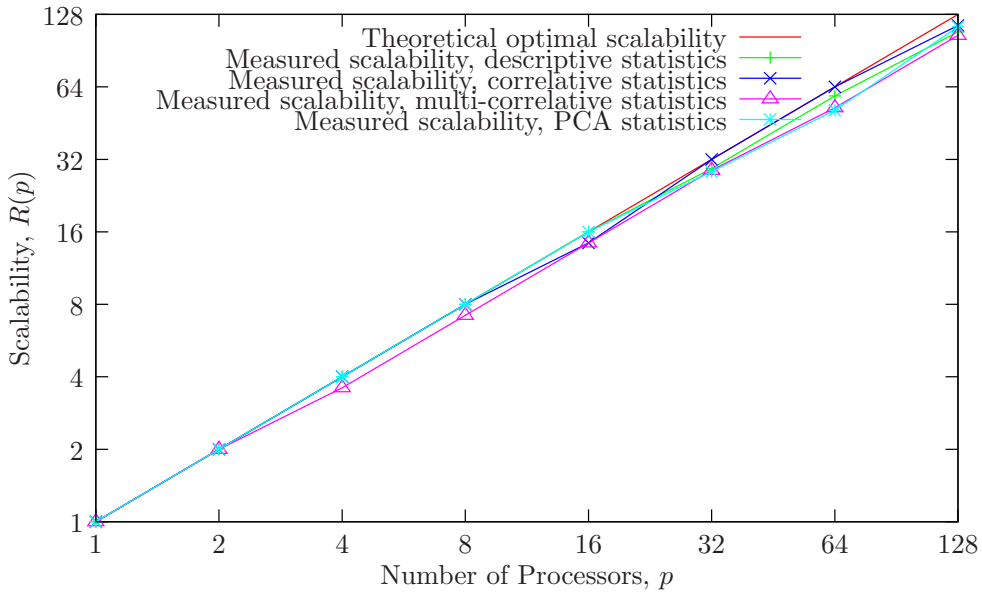

Figure 4. Rate of computation scalability at constant work per processor of $N(p)/p = 4,000,000$.

3.2 Algorithm Correctness

In order to assess algorithm correctness, we make use of the same test cases as § 3.1, for which we inspect the numerical results obtained by both the `vtkPMultiCorrelativeStatistics` and `vtkPPCA`classes. More precisely, we examine the statistical models obtained when both Learn and Derive options are turned on. Since the statistical properties of the test cases are known, we can immediately compare them to the calculated results. Note that this is the same metholody as the one we followed in [PT08] to verify the correctness of `vtkPDescriptiveStatistics` and `vtkPCorrelativeStatistics`.

Relatively large input sets are used ($n = 10^6$), in order to mitigate the risk of statistical bias due to insufficient sampling. In addition, the test case is run 100 times for each random variable, and we examine the statistical dispersion of the results of the ensemble of these runs. We compare the results obtained with the Learn and Derive option of the statistical engines to the theoretical values of the random variables which serve as models for the pseudo-random inputs, namely, $\mathscr{N}(0,1)$ and $\mathscr{U}(0,1)$. This comparison is done by simple visual inspection of the numerical results, by:

1. comparing the sample mean of the quantity of interest (e.g., mean) across the a number n_r of runs to the corresponding theoretical quantity (e.g, expectation), and
2. examining the variability of the results by checking the standard deviation of the quantity of interest across the n_r runs.

Table 5. Means of 8 pseudo-random independent samples of 4 standard normal and 4 standard uniform distributions, averaged across 100 runs, *versus* theoretical values. The last column indicates the standard deviation of the means across the 100 runs.

Sample	Expectation (theoretical mean)	Average of sample means	Standard deviation of sample means
$\mathscr{N}(0,1)_0$	0	0.0000328	0.000194
$\mathscr{N}(0,1)_1$	0	0.0000181	0.000159
$\mathscr{N}(0,1)_2$	0	−0.0000278	0.000173
$\mathscr{N}(0,1)_3$	0	0.0000278	0.000178
$\mathscr{U}(0,1)_4$	0.5	0.4999987	$5.448978 \cdot 10^{-5}$
$\mathscr{U}(0,1)_5$	0.5	0.4999961	$6.079136 \cdot 10^{-5}$
$\mathscr{U}(0,1)_6$	0.5	0.5000044	$5.586889 \cdot 10^{-5}$
$\mathscr{U}(0,1)_7$	0.5	0.5000030	$5.306256 \cdot 10^{-5}$

Using this methodology with either $n_r = 100$ runs over 32 processors, the results provided in

Table 5 show that pseudo-random input samples possess the desired first-order statistics, with egligible variation across the 100 runs.

Table 6. Cholesky decomposition of covarance matrix for 4 independent standard normal (top) and uniform (bottom) samples, averaged across 100 runs. Theoretical values are those of the identity matrix $\mathbb{I}_4$.

Standard Normal			
1.000005			
0.000009	0.999989		
0.000015	-0.000007	1.000006	
-0.000005	0.000028	0.000001	1.000013

Standard Uniform			
0.288670			
0.000003	0.288675		
-0.000008	-0.000003	0.288678	
-0.000003	-0.000011	-0.000005	0.288674

Table 6 shows the Cholesky decompositions of the covariance matrices that were calculated by `vtkPMultiCorrelativeStatistics` with, respectively, the standard uniform and standard normal pseudo-random inputs. We observe that the numerical results are in statistical agreement with their theoretical counterparts, up to at least 5 significant digits, in a manner consistent with the fluctuations of the pseudo-random inputs. In particular, and as expected due to the mutual independence of the input samples, only diagonal terms are non-zero–up to the aforementioned precision–in the Cholesky decompositions of the covariance matrices.

It is not necessary to conduct a verification of the calculation of the covariance matrix, and subsequent Cholesky decomposition, with `vtkPPCAStatistics` for this segment the parallel principal component analysis is simply inherited from `vtkPMultiCorrelativeStatistics`. It is, however, necessary that we verify that the PCA decomposition itself is correct, which we are doing by using a slightly modified input, comprising 4 pseudo-random samples of the following distributions:

1. 2 independent standard normal variables, denoted $X_0 = \mathcal{N}(0,1)_0$ and $X_1 = \mathcal{N}(0,1)_1$;
2. 2 variables obtained as the following linear combinations of the above: $X_2 = X_0 + X_1$, and $X_3 = 2X_0 - 3X_1$.

The statistical correlations of the input variables should be reflected in the output of the principal component analysis; in fact, it is trivial, based on elementary properties of Gaussian random

variables, to demonstrate that the random vector $X = (X_1, X_2, X_3, X_4)$ is Gaussian, centered, with covariance matrix:

$$\mathrm{cov}(X) = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & -3 \\ 1 & 1 & 2 & -1 \\ 2 & -3 & -1 & 13 \end{pmatrix}$$

for, if X_i and X_j are two random variable with respective distributions $\mathcal{N}(\mu_i, \sigma_i^2)$ and $\mathcal{N}(\mu_j, \sigma_j^2)$, then, for any arbitrary pair of reals (a, b), the random variable $aX_i + bX_j$ is Gaussian with distribution $\mathcal{N}(a\mu_i + b\mu_j, a^2\sigma_i^2 + b^2\sigma_j^2)$. The off-diagonal terms of the covariance matrix are obtained by using the linearity of the expectation:

$$\mathbb{E}(X_i \times (aX_j + bX_k)) = a\mathbb{E}(X_iX_j) + b\mathbb{E}(X_iX_k).$$

As predicated by the definition of the components of X, its covariante matrix is singular, with rank 2. Accordingly, it has only two non-zero eigenvalues, which can be approximated as $\lambda_1 \approx 14.1$ and $\lambda_2 \approx 2.91$. We are not discussing the eigenvectors here, as evincing them once the eigenvalues are known is trivial, and left to the reader as an exercise. The two zero eigenvalues are indeed retrieved with `vtkPPCAStatistics`, and the values for the two non-zero ones, averaged over $n_r = 100$ runs are shown in Table 7. We observe again agreement with the theoretical values, up to the fifth significant digit; this is in fact a better agreement that what we expected, considering the fact that the successive truncation errors of the covariance, Cholesky, and eigenvalue calculations are coumpounded with the fact that the pseudo-random samples do not exhibit perfect statistical properties, as illustrated by their means in Table 5 and covariances in Table 6. Jitter, as measured by standard deviation across all 100 runs, is essentially insignificant.

Table 7. Means of the non-zeroeigenvalues of 4 pseudo-random independent samples of 2 standard normal distributions, and 2 linear correlations thereof, averaged across 100 runs, *versus* approximated theoretical values. The last column indicates the standard deviation of the means across the 100 runs.

Eigenvalue (approx. theoretical)	Average of eigenvalues	Standard deviation of eigenvalues
14.0902	14.0906	0.011465
2.90983	2.90958	0.0023212

References

[PÓ8] P. Pébay. Formulas for robust, one-pass parallel computation of covariances and arbitrary-order statistical moments. Sandia Report SAND2008-6212, Sandia National Laboratories, September 2008.

[PT08] P. Pébay and D. Thompson. Scalable descriptive and correlative statistics with Titan. Sandia Report SAND2008-8260, Sandia National Laboratories, December 2008.

[tor] TORQUE Resource Manager. http://www.clusterresources.com/pages/products/torque-resource-manager.php.

[vtk] VTK Doxygen documentation. http://www.vtk.org/doc/nightly/html.

DISTRIBUTION:

2	MS 9159	Philippe P. Pébay, 8963
1	MS 9159	David Thompson, 8963
2	MS 9018	Central Technical Files, 8944
1	MS 0899	Technical Library, 9536

Sandia National Laboratories