SANDIA REPORT

SAND2008-6212
Unlimited Release
Printed September 2008

Formulas for Robust, One-Pass Parallel Computation of Covariances and Arbitrary-Order Statistical Moments

Philippe Pébay

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.

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

SAND2008-6212
Unlimited Release
Printed September 2008

Formulas for Robust, One-Pass Parallel Computation of Covariances and Arbitrary-Order Statistical Moments

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

Abstract

We present a formula for the pairwise update of arbitrary-order centered statistical moments. This formula is of particular interest to compute such moments in parallel for large-scale, distributed data sets. As a corollary, we indicate a specialization of this formula for incremental updates, of particular interest to streaming implementations. Finally, we provide pairwise and incremental update formulas for the covariance.

Acknowledgments

The author would like to thank Karen McWilliams for proof-reading this report, David Thompson and Jackson Mayo for their constructive comments, as well as Timothy B. Terriberry, for having indicated in July 2008 that – to the best of his knowledge – no generalization of his third- and fourth-order pairwise update formulas to arbitrary order was currently available.

Contents

This page intentionally left blank

1 Introduction

Centered statistical moments are one of the most widely used tools in descriptive statistics. It is therefore essential for statistical analysis packages that robust and efficient algorithms be devised and implemented. However, robustness and speed of execution, in this context as well as in others, tend to be orthogonal. For instance, it is well known[1] that algorithms for calculating centered statistical moments that utilize sum of powers for the sake of execution speed (one-pass algorithms) lead to unacceptable numerical instability. Remedies to this problem can be found by using two-pass algorithms, e.g.:

1. In the first pass, calculate the mean of the data set, and
2. In the second pass, calculate the required powers of the deviations to the previously calculated mean.

This approach is, in general, much more stable than the naïve one-pass approach. However, execution speed is severely impaired as each data point must be accessed twice; for large data sets in particular, one cannot rely on the expectation that all data could be retained in cache memory. Additional refinements of the two-pass approach have been proposed in order to further improve numerical stability. However, the issue of execution speed remains.

In addition, having the capability to analyze large-scale, distributed data sets is one of the stated goals of some of the most recent data analysis packages that are currently being developed [WBS08, WTP+08]. In this context, two-pass algorithms become entirely impractical because costs of distributed memory access massively dominate computation costs, and one must thus devise a one-pass algorithm that allows direct updates (i.e., where no updates of the mean are necessary) while being as numerically stable as possible. Moreover, such one-pass (or on-line) algorithms are directly amenable to streaming processing and thus to implementation for Graphics Processing Units (GPU). However, on-line algorithms necessitate recurrence formulas for updating the desired centered moments each time a new data point is added to the system.

In the case of the variance, on-line algorithms have long been known; see for instance [Wel62]. The gist of technique consists in the following recurrence formula for the mean:

$$\mu = \mu_1 + \frac{y - \mu_1}{n}, \tag{1.1}$$

where $\mathscr{S}_1$ is a data set with finite and non-negative cardinality $n-1$, μ_1 is its mean, and μ is the mean of the data set $\mathscr{S} = \mathscr{S}_1 \cup \{y\}$, y being an additional data point. Using (1.1), the following recurrence formula for the sum $M_{2,\mathscr{S}} = \sum_{x\in\mathscr{S}}(x-\mu)^2$ can then be used directly:

$$M_{2,\mathscr{S}} = M_{2,\mathscr{S}_1} + (y-\mu_1)(y-\mu). \tag{1.2}$$

[1] This can be easily verified by the reader by computing the variance of $\{1-\varepsilon, 1+\varepsilon\}$, where ε is smaller than than the square root of the machine's EPSILON.

from whence, for instance, the unbiased estimator for the variance of $\mathscr{S}$ is readily obtained as $\sigma^2_{n-1,\mathscr{S}} = \frac{1}{n-1} M_{2,\mathscr{S}}$.

With large-scale, distributed data sets, one would like, ideally, to conduct the calculations in an embarrassingly parallel fashion, except for final aggregation of the results obtained on each part of the distributed data set, with these final updates having a negligible cost. For example, if the data set of interest is partitioned into $\mathscr{S}_1$ and $\mathscr{S}_2$, one can use a one-pass algorithm to calculate separately, e.g, in different processes, the desired centered moments of $\mathscr{S}_1$ and $\mathscr{S}_2$. Upon completion of these partial computations, one then needs a global update formula for calculating the desired moments of $\mathscr{S} = \mathscr{S}_1 \cup \mathscr{S}_2$. For instance, in the case of the mean and the variance, [CGL79] derived the following formulas:

$$\mu = \mu_1 + n_2 \frac{\delta_{2,1}}{n}, \tag{1.3}$$

and

$$M_{2,\mathscr{S}} = M_{2,\mathscr{S}_1} + M_{2,\mathscr{S}_2} + n_1 n_2 \frac{\delta^2_{2,1}}{n}, \tag{1.4}$$

where n, n_1 and n_2 are the respective cardinalities of $\mathscr{S}$, $\mathscr{S}_1$ and $\mathscr{S}_2$, and $\delta_{2,1} = \mu_2 - \mu_1$. Evidently, (1.1) and (1.2) are particular cases of, respectively, (1.3) and (1.4), occurring when $\mathscr{S}_2$ is reduced to a singleton $\{y\}$.

For third- and fourth-order moments, which are needed to calculate skewness and kurtosis of the data set, formulas have been derived by [Ter08], in the form of pairwise update formulas for $M_{3,\mathscr{S}} = \sum_{x\in\mathscr{S}} (x-\mu)^3$ and $M_{4,\mathscr{S}} = \sum_{x\in\mathscr{S}} (x-\mu)^4$, as follows:

$$M_{3,\mathscr{S}} = M_{3,\mathscr{S}_1} + M_{3,\mathscr{S}_2} + n_1 n_2 (n_1 - n_2) \frac{\delta^3_{2,1}}{n^2} + 3(n_1 M_{2,\mathscr{S}_2} - n_2 M_{2,\mathscr{S}_1}) \frac{\delta_{2,1}}{n}, \tag{1.5}$$

and

$$\begin{aligned} M_{4,\mathscr{S}} = M_{4,\mathscr{S}_1} + M_{4,\mathscr{S}_2} + n_1 n_2 (n_1^2 - n_1 n_2 + n_2^2) \frac{\delta^4_{2,1}}{n^3} + 6(n_1^2 M_{2,\mathscr{S}_2} + n_2^2 M_{2,\mathscr{S}_1}) \frac{\delta^2_{2,1}}{n^2} \\ + 4(n_1 M_{3,\mathscr{S}_2} - n_2 M_{3,\mathscr{S}_1}) \frac{\delta_{2,1}}{n}. \end{aligned} \tag{1.6}$$

Evidently, these pairwise update formulas can be readily specialized to the particular case (incremental updates) where one of the two subsets is reduced to a singleton, in a similar fashion to the specialization of (1.4) to (1.2). The incremental update formulas can, in particular, be used to calculate M_3 and M_4 on each subset of the partitioned data set. Although [Ter08] does not provide them, their derivation from (1.5) and (1.6) is trivial and thus not detailed in this report.

The results recalled thus far provide the necessary formulas for efficient, one-pass, robust parallel calculations of statistical moments up to the fourth order, thus addressing the issue that arises when the considered application requires that mean, variance, skewness, and kurtosis be calculated. Although this probably covers the needs of the vast majority of traditional applications of descriptive statistics, more recent applications require that higher-order statistics be used: for example, in the context of signal processing, cf. [KHSS05] for a method using up to the sixth-order centered statistical moments, but which in theory can be used with arbitrarily higher-order moments.

To our knowledge, there are currently no published formulas for parallel updates of higher-order moments. Therefore, the goal of this report is to address this issue; more precisely, rather than providing a number of additional formulas for given orders, this report presents a general result that is valid for all orders. Finally, because covariance estimators are also of broad interest (e.g., for correlation analysis), and no parallel update formulas seem to have been proposed thus far, this report also presents formulas for the pairwise update of the covariance.

2 Arbitrary-Order Update Formulas

Proposition 2.1. *Let p be a natural number greater than 1. Then, using the same notations as in Section 1, the pairwise update formula for $M_{p,\mathscr{S}} = \sum_{x\in\mathscr{S}}(x-\mu)^p$ is:*

$$M_{p,\mathscr{S}} = M_{p,\mathscr{S}_1} + M_{p,\mathscr{S}_2} + \sum_{k=1}^{p-2}\binom{k}{p}\left[\left(-\frac{n_2}{n}\right)^k M_{p-k,\mathscr{S}_1} + \left(\frac{n_1}{n}\right)^k M_{p-k,\mathscr{S}_2}\right]\delta_{2,1}^k + \left(\frac{n_1 n_2}{n}\delta_{2,1}\right)^p\left[\frac{1}{n_2^{p-1}} - \left(\frac{-1}{n_1}\right)^{p-1}\right]. \tag{2.1}$$

Proof. By definition of $M_{p,\mathscr{S}}$, and because $\{\mathscr{S}_1,\mathscr{S}_2\}$ is a partition of $\mathscr{S}$, one has

$$\begin{aligned}
M_{p,\mathscr{S}} &= \sum_{x\in\mathscr{S}}(x-\mu)^p && (2.2)\\
&= \sum_{x\in\mathscr{S}_1}(x-\mu)^p + \sum_{x\in\mathscr{S}_2}(x-\mu)^p && (2.3)\\
&= \sum_{x\in\mathscr{S}_1}\left(x-\frac{n_1\mu_1+n_2\mu_2}{n}\right)^p + \sum_{x\in\mathscr{S}_2}\left(x-\frac{n_1\mu_1+n_2\mu_2}{n}\right)^p && (2.4)\\
&= \sum_{x\in\mathscr{S}_1}\left(x-\mu_1-\frac{n_2}{n}\delta_{2,1}\right)^p + \sum_{x\in\mathscr{S}_2}\left(x-\mu_2+\frac{n_1}{n}\delta_{2,1}\right)^p && (2.5)\\
&= \sum_{k=0}^{p}\binom{k}{p}\left[M_{p-k,\mathscr{S}_1}\left(-\frac{n_2}{n}\delta_{2,1}\right)^k + M_{p-k,\mathscr{S}_2}\left(\frac{n_1}{n}\delta_{2,1}\right)^k\right], && (2.6)
\end{aligned}$$

thanks to the commutativity of summations over finite sets, which allows us here to permute $\sum_{k=0}^{p}$ with $\sum_{x\in\mathscr{S}_1}$ and $\sum_{x\in\mathscr{S}_2}$. Now, a few simplifications are in order. First, the $k=0$ term of the above summation is simply $M_{p,\mathscr{S}_1}+M_{p,\mathscr{S}_2}$. Second, by definition, both $M_{1,\mathscr{S}_1}$ and $M_{1,\mathscr{S}_2}$ are zero, thus eliminating the $k=p-1$ term from the summation. Last, $M_{0,\mathscr{S}_1}=n_1$, $M_{0,\mathscr{S}_2}=n_2$, and

$$\begin{aligned}
n_1\left(-\frac{n_2}{n}\delta_{2,1}\right)^p + n_2\left(\frac{n_1}{n}\delta_{2,1}\right)^p &= \left(\frac{n_1 n_2}{n}\delta_{2,1}\right)^p\left[\frac{(-1)^p}{n_1^{p-1}} + \frac{1}{n_2^{p-1}}\right] && (2.7)\\
&= \left(\frac{n_1 n_2}{n}\delta_{2,1}\right)^p\left[\frac{1}{n_2^{p-1}} - \left(\frac{-1}{n_1}\right)^{p-1}\right]. && (2.8)
\end{aligned}$$

Therefore, by substituting (2.8) for the $k=p$ term in (2.6), one finally obtains (2.1). □

One can readily verify that the pairwise update formulas for M_3 and M_4 indicated in [Ter08], respectively recalled in (1.5) and (1.6), are particular cases of Proposition 2.1 when $p=3$ and $p=4$, respectively.

Corollary 2.2. *In the case where $\mathscr{S}_2$ is reduced to a singleton $\{y\}$, and denoting $\delta = y-\mu_1$, Proposition 2.1 reduces to the incremental update formula for $\mathscr{S} = \mathscr{S}_1\cup\{y\}$ as follows:*

$$M_{p,\mathscr{S}} = M_{p,\mathscr{S}_1} + \sum_{k=1}^{p-2}\binom{k}{p}M_{p-k,\mathscr{S}_1}\left(\frac{-\delta}{n}\right)^k + \left(\frac{(n-1)}{n}\delta\right)^p\left[1-\left(\frac{-1}{n-1}\right)^{p-1}\right] \tag{2.9}$$

Proof. Corollary 2.2 is an immediate specialization of Proposition 2.1, obtained when $n_1 = n - 1$ and $n_2 = 1$. In this case, $\delta_{2,1} = \delta$ and each $M_{p,\mathscr{S}_2}$ vanishes since $\mu_2 = y$, and thus (2.1) is immediately simplified into (2.9). □

Remark 2.1. By noticing that (1.1) is equivalent to

$$y - \mu = \frac{n-1}{n}(y - \mu_1) \tag{2.10}$$

one directly retrieves (1.2) from Corollary 2.2 with $p = 2$.

As another and more interesting illustration of Corollary 2.2, one can readily evince the incremental update formulas for M_3 and M_4:

Example 2.1. If $p = 3$, then (2.9) becomes:

$$M_{3,\mathscr{S}} = M_{3,\mathscr{S}_1} - 3M_{2,\mathscr{S}_1}\frac{\delta}{n} + \left(\frac{(n-1)}{n}\delta\right)^3\left[1 - \left(\frac{-1}{n-1}\right)^2\right] \tag{2.11}$$

$$= M_{3,\mathscr{S}_1} - 3M_{2,\mathscr{S}_1}\frac{\delta}{n} + \frac{(n-1)^3\delta^3}{n^3} \times \frac{n^2 - 2n}{(n-1)^2} \tag{2.12}$$

$$= M_{3,\mathscr{S}_1} - 3M_{2,\mathscr{S}_1}\frac{\delta}{n} + (n-1)(n-2)\frac{\delta^3}{n^2}. \tag{2.13}$$

Example 2.2. If $p = 4$, then (2.9) becomes:

$$M_{4,\mathscr{S}} = M_{4,\mathscr{S}_1} - 4M_{3,\mathscr{S}_1}\frac{\delta}{n} + 6M_{2,\mathscr{S}_1}\left(\frac{\delta}{n}\right)^2 + \left(\frac{(n-1)}{n}\delta\right)^4\left[1 - \left(\frac{-1}{n-1}\right)^3\right] \tag{2.14}$$

$$= M_{4,\mathscr{S}_1} - 4M_{3,\mathscr{S}_1}\frac{\delta}{n} + 6M_{2,\mathscr{S}_1}\left(\frac{\delta}{n}\right)^2 + \frac{(n-1)^4\delta^4}{n^4} \times \frac{n^3 - 3n^2 + 3n}{(n-1)^3} \tag{2.15}$$

$$= M_{4,\mathscr{S}_1} - 4M_{3,\mathscr{S}_1}\frac{\delta}{n} + 6M_{2,\mathscr{S}_1}\left(\frac{\delta}{n}\right)^2 + (n-1)(n^2 - 3n + 3)\frac{\delta^4}{n^3}. \tag{2.16}$$

Note that we are providing implementations of these in the open-source Visualization Tool Kit (VTK), more precisely as part of the `vtkDescriptiveStatistics` class.

3 Covariance Update Formulas

In this section, we provide formulas for both incremental and pairwise update of the covariance. These are of interest, in particular, for Pearson correlation analysis, which we wish to conduct on large-scale, distributed data sets.

In this section, $\mathscr{S}$ denotes a set of doubles $x = (u, v)$. Existing notations from Section 1 are retained; in addition, we will use $\mu_{u,1}$, $\mu_{v,1}$, $\mu_{u,2}$, and $\mu_{v,2}$ to denote the means of u and v on $\mathscr{S}_1$ and $\mathscr{S}_2$, respectively. Also, we define $\delta_{u,2,1} = \mu_{u,2} - \mu_{u,1}$, and $\delta_{v,2,1} = \mu_{v,2} - \mu_{v,1}$.

Proposition 3.1. *The pairwise update formula for $C_{2,\mathscr{S}} = \sum_{(u,v)\in\mathscr{S}}(u-\mu_u)(v-\mu_v)$ is:*

$$C_{2,\mathscr{S}} = C_{2,\mathscr{S}_1} + C_{2,\mathscr{S}_2} + \frac{n_1 n_2}{n}\delta_{u,2,1}\delta_{v,2,1}. \tag{3.1}$$

Proof. By definition of $C_{2,\mathscr{S}}$, and because $\{\mathscr{S}_1, \mathscr{S}_2\}$ is a partition of $\mathscr{S}$, one has

$$\begin{aligned} C_{2,\mathscr{S}} &= \sum_{(u,v)\in\mathscr{S}} (u-\mu_u)(v-\mu_v) && (3.2)\\ &= \sum_{(u,v)\in\mathscr{S}_1} (u-\mu_u)(v-\mu_v) + \sum_{(u,v)\in\mathscr{S}_2} (u-\mu_u)(v-\mu_v). && (3.3) \end{aligned}$$

As in (2.4), we expand the means over $\mathscr{S}$ into expressions that relate them to the means on $\mathscr{S}_1$ and $\mathscr{S}_2$:

$$\begin{aligned} \sum_{(u,v)\in\mathscr{S}_1} (u-\mu_u)(v-\mu_v) &= \sum_{(u,v)\in\mathscr{S}_1} \left(u - \frac{n_1\mu_{u,1} + n_2\mu_{u,2}}{n}\right)\left(v - \frac{n_1\mu_{v,1} + n_2\mu_{v,2}}{n}\right) && (3.4)\\ &= \sum_{(u,v)\in\mathscr{S}_1} \left(u - \mu_{u,1} - \frac{n_2}{n}\delta_{u,2,1}\right)\left(v - \mu_{v,1} - \frac{n_2}{n}\delta_{v,2,1}\right) && (3.5)\\ &= \sum_{(u,v)\in\mathscr{S}_1} \left[\begin{array}{c} (u-\mu_{u,1})(v-\mu_{v,1}) - \frac{n_2}{n}\delta_{u,2,1}(u-\mu_{u,1}) \\ -\frac{n_2}{n}\delta_{v,2,1}(v-\mu_{v,1}) + \frac{n_2^2}{n^2}\delta_{u,2,1}\delta_{v,2,1} \end{array}\right]. && (3.6) \end{aligned}$$

Again using the commutatitivty of summations over finite sets and noticing that, by definition of $\mu_{u,1}$ and $\mu_{v,1}$,

$$\sum_{(u,v)\in\mathscr{S}_1} \frac{n_2}{n}\delta_{u,2,1}(u-\mu_{u,1}) = \frac{n_2}{n}\delta_{u,2,1}\sum_{u\in\mathscr{S}_1}(u-\mu_{u,1}) = 0 \tag{3.7}$$

and

$$\sum_{(u,v)\in\mathscr{S}_1} \frac{n_2}{n}\delta_{v,2,1}(v-\mu_{v,1}) = \frac{n_2}{n}\delta_{v,2,1}\sum_{v\in\mathscr{S}_1}(v-\mu_{v,1}) = 0, \tag{3.8}$$

we simplify (3.6) into

$$\sum_{(u,v)\in\mathscr{S}_1} (u-\mu_u)(v-\mu_v) = C_{2,\mathscr{S}_1} + \frac{n_1 n_2^2}{n^2}\delta_{u,2,1}\delta_{v,2,1}. \tag{3.9}$$

Similarly, by interchanging the roles of $\mathscr{S}_1$ and $\mathscr{S}_2$, we obtain

$$\sum_{(u,v)\in\mathscr{S}_2}(u-\mu_u)(v-\mu_v) = C_{2,\mathscr{S}_2} + \frac{n_2 n_1^2}{n^2}\delta_{u,2,1}\delta_{v,2,1} \tag{3.10}$$

because $(\mu_{u,2}-\mu_{u,1})(\mu_{v,2}-\mu_{v,1}) = (\mu_{u,1}-\mu_{u,2})(\mu_{v,1}-\mu_{v,2})$. Therefore, (3.3) becomes

$$C_{2,\mathscr{S}} = C_{2,\mathscr{S}_1} + C_{2,\mathscr{S}_2} + \frac{n_1 n_2^2 + n_2 n_1^2}{n^2}\delta_{u,2,1}\delta_{v,2,1}, \tag{3.11}$$

from whence the result arises since $n_1 + n_2 = n$. □

Pairwise update formulas for the estimators of the variance are now obtained immediately; for instance, the unbiased estimator of the covariance of $\mathscr{S}$ is $\frac{1}{n-1}C_{2,\mathscr{S}}$.

Remark 3.1. In passing, we note that in the case where $\mathscr{S}_2$ is reduced to a singleton $\{(s,t)\}$, Proposition 3.1 reduces to the following incremental update formula for $\mathscr{S} = \mathscr{S}_1 \cup \{(s,t)\}$:

$$C_{2,\mathscr{S}} = C_{2,\mathscr{S}_1} + \frac{n-1}{n}(s-\mu_{u,1})(t-\mu_{v,1}). \tag{3.12}$$

References

[CGL79] T. F. Chan, G. H. Golub, and R. J. LeVeque. Updating formulae and a pairwise algorithm for computing sample variances. Technical Report STAN-CS-79-773, Stanford University, Department of Computer Science, 1979.

[KHSS05] N. Kikuchi, S. Hayase, K. Sekine, and S. Sasaki. Performance of chromatic dispersion monitoring using statistical moments of asynchronously sampled waveform histograms. *Photonics Technology Letters*, 17:1103–1105, May 2005.

[Ter08] T. B. Terriberry. Computing higher-order moments online, 2008. `http://people.xiph.org/~tterribe/notes/homs.html`.

[WBS08] B. Wylie, J. Baumes, and T. Shead. Titan informatics toolkit. In *IEEE Visualization Tutorial*, Columbus, OH, October 2008.

[Wel62] B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419420, 1962.

[WTP+08] M. H. Wong, D. C. Thompson, P. Pébay, J. R. Mayo, A. C. Gentile, B. J. Debusschere, and J. M. Brandt. OVIS-2: A robust distributed architecture for scalable RAS. In *Proc. 22nd IEEE International Parallel & Distributed Processing Symposium*, Miami, FL, April 2008.

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