AN ALLAN VARIANCE COMPUTATION FROM VERY LARGE DATASETS

Jan Kunz
Doctoral Degree Programme (2), FEEC BUT
E-mail: xkunzj00@stud.feec.vutbr.cz

Supervised by: Petr Beneš
E-mail: benesp@feec.vutbr.cz

Abstract: This paper deals with calculating an Allan variance from a large datasets, which does not fit into a RAM memory. For this reason a parallel method is implemented in LabVIEW 2015 and tested on four different configurations of personal computer. An execution time was measured for a dataset of 2 754 448 000 samples. The time varies in dependence on calculated samples and the configuration of the computer. For the faster computers execution takes several minutes, whereas on the slower one the longest tested calculation of 100 samples per decade takes about two hours, which is still practical.

Keywords: Allan variance, large file, MEMS, gyroscope

1 INTRODUCTION

An Allan variance, firstly introduced in [1], is a method for evaluating stochastic parameters of sensors. The method calculation (1) together with parameters extraction are for instance described in [1, 2, 3, 4, 5, 6, 7, 8], for this reason the Allan variance calculation will not be further discussed in this paper.

$$\sigma(\tau) = \frac{1}{2(N-2n)} \sum_{k=1}^{N-2n} \left[\overline{\Omega}_{k+1}(\tau) - \overline{\Omega}_k(\tau)\right]^2 \tag{1}$$

where σ is the Allan variance for time τ, N is the number of samples, n is the number of samples in a group Ω for time τ and $\overline{\Omega}$ denotes an average value from the group Ω

The Allan variance was initially intended to use for etalon characteristics.[1] Later, it was established as a standard for frequency stability.[2] Then, as the sensor quality increases, this method became to be important for a characterisation of sensors.[3, 5, 9] The knowledge of the stochastic parameters is extremely important especially for inertial navigation gyroscopes, where its output is integrated to get an angular position, which is essential for orientation, so the noise causes significant error. [4]

Furthermore, the sensors are gradually improved and stochastic parameters are therefore better which makes them more difficult to measure. The only way how to measure them is to collect data so long that the parameters reveal itself. [5, 7, 8] The necessary measurement time varies from several days for MEMS gyroscopes to several weeks for fibre optic gyroscopes. [10, 11] This leads to very long measurements, which generates large datasets.

2 DATA LENGTH

To evaluate the parameters it is necessary to have a data which covers them. On one hand there exist a stochastic parameters such as a quantization or a white noise, which are significant on a higher frequencies so, to measure them it is necessary to sample the data with a high sample rate, usually

several tens of kilohertz. Whereas on the other hand, parameters such as a bias instability or an exponentially correlated noise (rate random walk) needs quite a long time to be significant. For example, in case of MEMS gyroscopes ADXLS61x it is more than 50 000 s.[10, 12] Moreover, for some fibre optics gyroscopes it is more than several hundreds hours. [11]

Furthermore, to calculate them it is necessary to measure ten times, for the Allan variance, and five times, for a total variance, longer.[2, 13] To estimate all aforementioned noises it is required to sample with a high sample rate for a very long time. This leads to several billions measured values. For example, in case of MEMS gyroscope a measurement with sampling frequency f_s = 10 kHz and for t = 300000 s yields in 3 billions samples which in double precision (8 bytes per sample) takes up 24 GB of data. Moreover, the measurement can take much longer [1] or the sampling frequency can be higher to evaluate the high frequency noises. In conclusion, when calculating the Allan variance, it can be necessary to process data which does not fit into the RAM memory therefore, it is necessary to use a different approach.

On the other hand, to calculate the Allan variance it is not necessary to store everything with the same sampling frequency. Data for a calculation of higher cluster times can be sampled with a lower sampling rate, which results in a smaller dataset, which could fit into the RAM memory. However, lowering the sampling rate could lead to some information losses for an other data processing. This together with almost unlimited and cheap storage capacity makes this attitude obsolete so, it will not be further discussed.

3 CALCULATION LIMITS

When processing a large dataset it has to be taken into account that computers are not built for them hence, it brings some issues which has to be solved. The algorithm is developed in LabVIEW because in the same environment is also processed the measurement due its connectivity to a measurement hardware and other advantages mentioned for example in [14].

For example, LabVIEW using for indexing arrays I32 data-type, which has the maximal represented value $2^{31} - 1 = 2\ 147\ 483\ 647$ so, even if that big array fitted into the RAM memory, it would not be possible to access all elements. Of course, this inconvenience can be solved quite easily by splitting the array to several shorter ones or in a two dimensional array, however these limitation has to be taken into account.

Furthermore, as a part of the Allan variance calculation average values from the input data are calculated.[1, 2] The average can be calculated, in case of a total variance, from a whole input array.[13] In that case, a proper calculation can be limited by a dynamic range of a double number representation, which is used for calculation and has a dynamic range of 52 bit, or 313 dB [2].[15] Furthermore, for a proper summation of two numbers of the same bit length the result has to be a one bit longer. Using the same logic a sum of $2^{10} = 1024$ elements requires the result to be ten bits longer. That means that summation of three billions ($\approx 2^{33}$) samples, which has a 20 bit precision, needs a dynamic range 53 dB which exceeds the double precision dynamic range. On the other hand, this inconvenience can be solved quite easily by doing several consecutive averages and then a final average, however it needs to be considered.

[1] For instance, data for Allan variance for aforementioned Fibre optic gyroscope should be measured at least for 1000 hours, which yields, with the sampling frequency f_s = 10 kHz, in 288 GB dataset.

[2] DR = $20 \cdot log_{10} 2^{52} = 20 \cdot \frac{log_2 2^{52}}{log_2 10} \doteq \frac{20 \cdot 52}{3.32} = 313\ dB$

4 PROCESSING POSSIBILITIES

To calculate the Allan variance from a large dataset can be used the same attitude as with normal dataset length. This means that the variance value is calculated point-to-point, so all data are processed for one cluster length and then for the other. This method is simple and needs just a minor changes in the implementation, however it requires to load a whole dataset for each calculated sample. For example calculation of 100 samples from aforementioned data would require reading of 2,4 TB of data, which is extremely time consuming, considering a disk read speed 80 MB/s the reading would take more than 8 hours and therefore, ineffective.

4.1 PARALLEL COMPUTATION

To improve the computation speed and decrease the amount of data to read in general it is better to use a parallel calculation. Moreover, this method allows iteration parallelism which also greatly improve the performance. The parallelism is provided automatically by the LabVIEW environment.[16] For this reason, it is enough to just enable this feature to speed-up the computation.

In contrast to the previous method this one loads the data only once. From each data part loaded for processing a partial average for all groups is calculated. This attitude requires some additional calculations such as number of samples remaining to finish each group and the number of completed elements in each group. Despite that these additional calculations requires some time it is still faster and more convenient for the hard drive than the previous method.

The algorithm can be described (fig. 1) in three steps for each loaded data part. Firstly, an average calculation for all uncompleted groups is finished, the result is subtracted from the previous one, the difference is squared and added to the previous sum.[1, 2] In case that there is not enough data to complete a particular group, whole actual samples are summed and the result together with an actual count of addends are passed to the next iteration. Then, an average from whole complete groups available in the remaining data are calculated, the difference is done and summed. Finally, the rest of samples for each group is pre-calculated for the completion in the next iteration. After all the data are processed the Allan variance is calculated for each group from the sum of the squared differences.

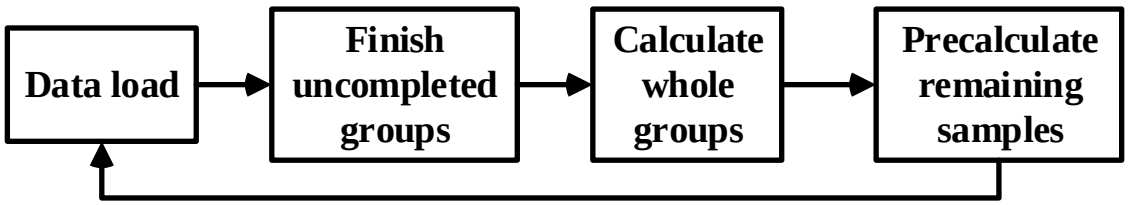

Figure 1: Parallel computation algorithm of an Allan variance

This algorithm was implemented in LabVIEW 2015 on a computer with Windows 10 64-bit operating system. The algorithm is able to calculate the Allan variance from a very large files in a reasonable time, which varies in dependence on the calculated samples. For testing was used a 20 GB file and the durations for different number of calculated samples per decade are in table (1).

The different duration for different computer configurations (tab. 1) shows that the speed of computation is highly dependent on a disc speed for less calculated samples, whereas for more calculated samples processors play more important role. For all cases it is approximately twice faster to use 64-bit LabVIEW version, which can be explained by wider registers and better usage of memory.[17] On the other hand, the 32-bit version offers wider range of tool-kits and hardware drivers so this is often the only installed LabVIEW version.[18] Hence, the installation of 64-bit LabVIEW version

Table 1: Table of durations of an Allan variance calculation from 2 754 448 000 samples in dependence on samples per decade and different computer configurations. PC1 has processor AMD Phenom II X4 840 3,2 GHz, 8 GB DDR3 1333 MHz RAM and 500 GB WD blue HDD. PC2 has processor Intel i5-4460 3,4 GHz, 8 GB DDR3 1600 MHz RAM and 1TB Seagate Desktop HDD. PC3 has processor Intel i5-4460 3,4 GHz, 8 GB DDR3 1600 MHz RAM and 240 GB Intel SSD 535 Series SSD. PC4 has processor Intel i5-7600K 3,8 GHz, 16 GB DDR4 2400 MHz RAM and 1 TB Seagate 7200 rpm HDD.

Sample per decade [N]	LabVIEW 2015 64-bit				LabVIEW 2015 32-bit			
	PC1 [s]	PC2 [s]	PC3 [s]	PC4 [s]	PC1 [s]	PC2 [s]	PC3 [s]	PC4 [s]
1	366	172	129	156	393	206	164	183
2	407	186	141	168	458	255	207	215
5	594	237	186	203	791	357	313	282
10	935	311	250	256	1287	478	435	376
20	1606	475	374	357	2223	733	702	551
50	3455	765	749	645	4910	1459	1440	1074
100	6373	1311	1294	1094	9102	2685	2655	1928

is convenient only if the Allan variance is calculated often, otherwise the 32-bit version satisfy the needs. Moreover, in many cases it is enough calculate the Allan variance with 10 or 20 samples per decade which is done within several minutes. In case that higher amount of samples per decade is necessary the Allan variance can be easily calculated on the slowest tested computer overnight, which is enough as this type of calculation is performed only in special cases.

5 CONCLUSION

This paper presents a method of an effective calculation of the Allan variance from a very large datasets which does not fit into a personal computer RAM memory. Also processing of that amount of data are beyond the limits of LabVIEW 2015 programming environment, where the algorithm is programmed.

Two possible ways of the algorithm implementation are considered, however the sequential one is not implemented due its extreme demand on data to load, which is very time demanding and therefore ineffective. For this reason, only parallel implementation is explained, programmed and tested on several different computer configurations.

From the result is visible (tab. 1), that the execution time is highly dependent on LabVIEW version, where the 64-bit version is approximately twice faster than the 32-bit version. Moreover, the time is for less samples per decade dependent mostly on the disc speed, whereas for more samples per decade the time is dependent more on the processor speed than on the disc. On the other hand, the execution times are for all tested configuration still reasonable. In conclusion, the implemented method is able to calculate the Allan variance from a very large files (tens of gigabytes) within a reasonable time on a normal personal computer, which is very practical and convenient for further research.

ACKNOWLEDGEMENT

The completion of this paper was made possible by the grant No. FEKT-S-17-4234 - „Industry 4.0 in automation and cybernetics" financially supported by the Internal science fund of Brno University of Technology.

REFERENCES

[1] David W Allan. Statistics of atomic frequency standards. *Proceedings of the IEEE*, 54(2):221–230, 1966.

[2] James A Barnes, Andrew R Chi, Leonard S Cutler, Daniel J Healey, David B Leeson, Thomas E McGunigal, James A Mullen, Warren L Smith, Richard L Sydnor, Robert FC Vessot, et al. Characterization of frequency stability. *IEEE transactions on instrumentation and measurement*, 1001(2):105–120, 1971.

[3] David W Allan and Judah Levine. A historical perspective on the development of the allan variances and their strengths and weaknesses. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 63(4):513–519, 2016.

[4] Priyanka Aggarwal. *MEMS-based integrated navigation*. Artech House, 2010.

[5] Naser El-Sheimy, Haiying Hou, and Xiaoji Niu. Analysis and modeling of inertial sensors using allan variance. *IEEE Transactions on instrumentation and measurement*, 57(1):140–149, 2008.

[6] Miroslav Matejček and Mikuláš Šostronek. Computation and evaluation allan variance results. In *New Trends in Signal Processing (NTSP)*, pages 1–9. IEEE, 2016.

[7] Jintao Li and Jiancheng Fang. Not fully overlapping allan variance and total variance for inertial sensor stochastic error analysis. *IEEE Transactions on Instrumentation and Measurement*, 62(10):2659–2672, 2013.

[8] Jintao Li and Jiancheng Fang. Sliding average allan variance for inertial sensor stochastic error analysis. *IEEE Transactions on Instrumentation and Measurement*, 62(12):3291–3300, 2013.

[9] Ieee standard specification format guide and test procedure for coriolis vibratory gyros. *IEEE Std 1431-2004*, pages 1–78, Dec 2004.

[10] Analog Devices. Adxrs610.

[11] iXblue. Ixblue ultimate-performance fiber-optic gyroscope.

[12] Analog Devices. Adxrs613.

[13] David A Howe. Total variance explained. In *Proc. 13th European Frequency and Time Forum*, pages 1093–1099, 1999.

[14] National Instruments. Whitepaper: Advantages of labview in academic research.

[15] Dan Zuras, Mike Cowlishaw, Alex Aiken, Matthew Applegate, David Bailey, Steve Bass, Dileep Bhandarkar, Mahesh Bhat, David Bindel, Sylvie Boldo, et al. Ieee standard for floating-point arithmetic. *IEEE Std 754-2008*, pages 1–70, 2008.

[16] National Instruments. Whitepaper: Improving performance with parallel for loops.

[17] National Instruments. Whitepaper: Announcing 64-bit labview.

[18] National Instruments. Whitepaper: National instruments product compatibility for microsoft windows 10.