Mark One Lifestyle, Inc.

Scan to Scan Variation for the StellarNet Spectrometer

Data Collection

- Date: 2015-04-22
- Ticket: AS-295
- Git Repo: data.mkone.co/var/git/science/vessyl/as-295.git
- Git Branch: master

Analysis

- Ticket: VA-93
- Git Repo: git.mkone.co:vessyl-algorithms/algorithms.git
- Git Branch: va-93

Table of Contents

Revision Control

Revision Number	Revision Date	Notes	Owner
1.0.0	2015-05-08	First draft release	Ehson Ghandehari

Executive Summary

- Variance increases with signal amplitude and wavelength
- Coefficient of Variation is between 0.5-5%
- Variance is not normally distributed. As compare to a normal distribution, less values were shifted towards the sample mean. The distribution was symmetrical, with very tails.

Introduction

Mark One Lifestyle has a StellarNet spectrometer. The purpose of this study was to examine the variation within a trial (128 scans) at each wavelength.

Methods

The data was collected under section 3.5.1 of the Wavelength Selection DOE.

The setup used in the data collection was the red golden setup. The red golden setup consisted of a light source (Ocean Optics DH-2000; serial number 005400821), an optical fiber (Thorlabs; M200L02S-UV), an optical switch (Ocean Optics; INLINE-TTLS), a custom bifuricated fiber optic reflection probe (Thorlabs; 6 fibers from the light source and 1 fiber to the spectrometer; the length from the light source to the cuvette or the spectrometer to the cuvette is 1 m each; the fiber used in this reflectance probe is a Thorlabs FG550UEC), a custom refurbished cuvette holder (Mark One Lifestyle Inc.), a quartz glass cuvette (Thorlabs; CV10Q3500F) with a custom aluminum mirrored surface, and a spectrometer (StellarNet; Silver Nova model, serial number 15040704). The light source was connected to the optical switch by M200L02S-UV fiber. The optical switch was connected to the right side of cuvette holder by FG550UEC probe, and the left side of cuvette holder was connected to the spectrometer by the same probe. The spectrometer was placed in an environmental chamber (Espec BTL-433) at a temperature of 15 C and 70% relative humidity. The spectrometer responded to a wavelength range from 178.28 nm to 1121.7 nm. A linear silicon CCD array detector was able to distinguish between 2036 wavelengths.

Data Collection

The data was collected under section 3.5.1 of the Wavelength Selection DOE. In summary, a single operator inserted an empty cuvette into the setup, took a trial of 128 scans.

The cuvette was cleaned by rinsing it three times with distilled water using a squirt bottle. Then it is rinsed once with isopropyl alcohol (IPA) and blown dry with pressurized air. After this initial cleaning, the cuvette was not cleaned again. The clean and empty cuvette was placed in the holder and clamped into position. The mirrored surface was contralateral to the fibre cable. A series of 128 scans, a trial, was collected and stored.

The operator was Ehson.

Data Analysis

The data was converted from XML format into tidy formatted comma separated values and all trials were concatenated together and compressed. This was done

using a custom bash script that calls gsxmltidy.py and the output is concatenated. This output is then compressed using gzip. These compress files were read into R version 3.1.2 (2014-10-31). Within R a four stage cleaning process was performed. All stages are optional and can be performed in any order. They include a conversion phase where the data is converted into the correct data type. This phase is often performed, in part, during the read operation. The transform phase is used to manipulate the values of the data and correct any errors. During the filtering phase, unneeded or erroneous data is removed from the data set. In the final phase the data is transformed into tidy format, if required.

After the cleaning process, the data is run through a series of checks. These checks are independent of the cleaning phase and they test the quality of the data and assumptions that are being made. These would include, but not limited to, checks to see if the data is the correct data type, if there are missing values, the correct number of levels exist in a factor, values are in range and to make sure that all data is present.

The light source produces several very large spikes in the spectrometer data. This causes saturation in some wavelengths and bleed over into the adjacent wavelengths. Therefore, if the sample data was clipped, sample measures greater than 58890, in any trial it was removed for all trials collected by a given operator. To remove the bleed over, wavelengths that were +/- 1 nm an either side were also removed.

Results

The operator performed a single trial, and each trial consisted of 128 scans. All scans for a setup were combined and the mean, per wavelength, was plotted. The thickness of the line, at each wavelength, was plus or minus one standard deviation.

Variance Between Wavelengths

A general assumption in statistical modelling is that the measured value (Y) is the sum of the signal (X) and the error (e). X, the sample, would be a function of the wavelength, X(lambda). The error,e, would not be a function of X or lambda and is completely random with a zero mean. That is, if the error is assumed to be independent of the wavelength or any other measurable factor. Figures 1 and 2 are a scatter plots of the deviance that each trial has from the mean value within a wavelength. The black line at Deviance=0 represents the mean measured value within a wavelength. Each blue dot represents the amount that a measure value deviated from the sample mean. The area near 680 nm, that lacks observations, are the result of removing data because the signal exceeded the dynamic range of the spectrometer. The darker the blue areas, means that more points are sitting on top of each other. This is most notable at near either end of the spectrum for both spectrometers.

The green lines in represent the 25th and 75th quantile. The line has been smoothed across wavelengths using a General Additive Model (GAM). The yellow line represents the 2.5th and 97.5th quantile with the same type of smoothing. If the variance was truly random and therefore independent of the wavelength, this would be a straight line that is parallel to the black line. This is evidence that variation within each wavelength is a function of the wavelength itself.

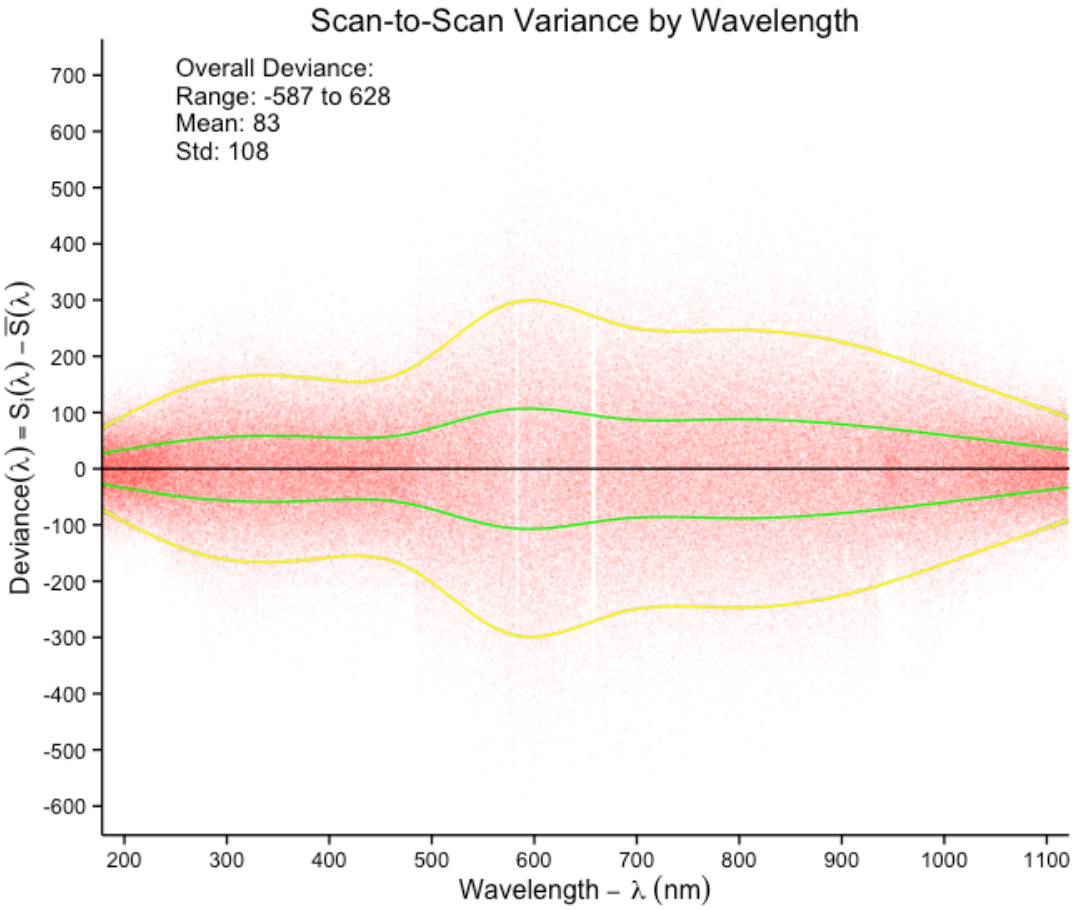

Figure 1: Scatter plot of the variation of observed measurements with respect the the mean sample within each wavelength. The black line represents the mean deviance. The green lines are the 25th and 75th quantile values that have been smoothed across wavelengths. The yellow line is the 2.5th and 97.5th percentile with the same smoothing. The error is not independent of the wavelength.

For The absolute mean deviance across all wavelengths was 83 counts for spectrometer. The standard deviation was 108. The observed range for spectrometer was -587 to 628.

For the spectrometer, the observed range was well outside the three standard deviations. However, there is very little skew (0.01), and the distribution is not

concentrated around the mean (kurtosis=0.93). This means that the data is not normally distributed but it is symmetric, with long tails.

Figures 1 demonstrated that the variance near extremes of the examined wavelengths was the smallest. It was hypothesized that the error in the signal may have been related to the magnitude of the sample measured. An examination of Figure 2 showed that the smallest signal in sample was near the extremes in the examined wavelengths; the same regions that demonstrated the smallest variance in the deviation measurement. The sharp signal drops in the extremes (wavelength below 200 nm or above 1100 nm) were due to the temperature compensation feature of Silver Nova spectrometer.

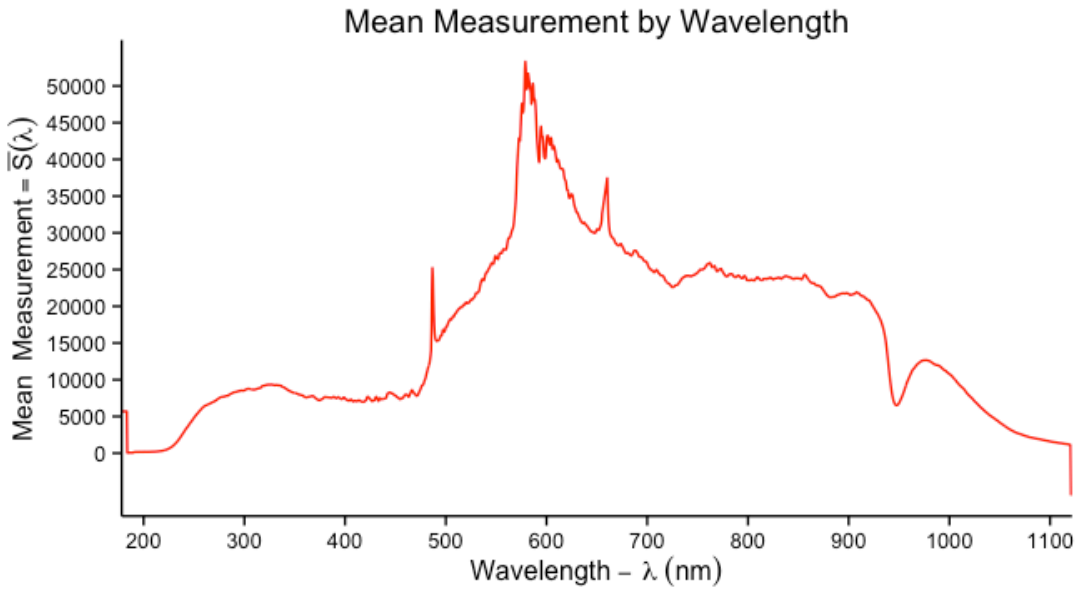

Figure 2: Mean measurement for each wavelength. The smallest mean measurements happened in the same regions where the smallest deviance was observed.

To explore the observation that the variance changes as a function of wavelength, the coefficient of variation (CV) was computed and shown in Figures 3. It was noticed that due to low signals at extreme sides of the spectrum, very high values of CV were calculated for wavelengths below 250 nm or above 1100 nm. In Figure 3, the CV values for wavelength 250-1100 nm are shown.

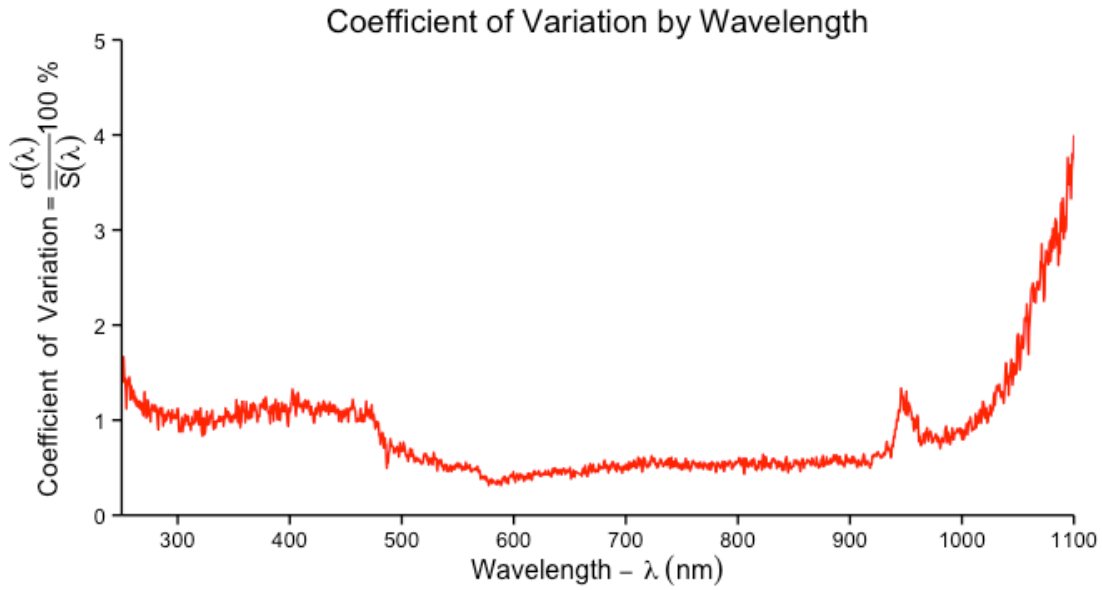

Figure 3: The Coefficient of Variation shows that the normalized variance is highest when the signal is low. It also shown that the variance is proportional to the measurement once it reaches a certain threshold. This is seen by the nearly horizontal line through the mid-ranges of the wavelengths.

Figures 2 demonstrated that the variance is related the the measured sample. If the variance was constant, then the Coefficient of Variation should vary with the mean value. That is, if there was no relationship between variance within a wavelength and the mean measurement than Figures 3 was a scaled version of Figures 2. However, this was not the case. Figures 3 demonstrated that when the mean measurement was low, at the two extremes of the examined wavelengths, then the Coefficient of Variation was relatively high (max 4-5%). In the midsection of the wavelength spectrum, the Coefficient of Variation is around 0.5-1%. That suggests that the variance in the signal is proportional to the measured sample.

This difference may be due to the "dark" noise dominating the low amplitude signal. The increased variance when the signal is larger, may be due to an amplification of the noise as the true signal increases.

Conclusion

This study has shown that the variance is a function of the amplitude of the measured sample and also the wavelength. The dependence on the wavelength may be a function of the "dark" noise. The distribution of the variance between is not normal. However, it is symmetrically distributed with a lower number of samples closer to the expected value than in a normal distribution. This suggests that repeated measures would produce consistent sample means as the spectrometers

has high precision. The observed long-tails of distribution suggests that individual scans may produce inaccurate results.