



ASSESSMENT OF REGULARITY IN PROTEIN MASS SPECTRA BY WAVELET-BASED TOOLS WITH APPLICATION IN DIAGNOSTICS OF OVARIAN CANCER

Kumbit Hwang

Taewoon Kong

Haesong Choi


- **Background and Objective**
- **Data Set**
- **Methodology**
- **Analysis**
- **Results**

Background

- 21,880 estimated new cases/ The 9th most common cancer among women
- 13,850 estimated deaths/ The 5th cancer leading to the death
- The relative five-year survival rate : 46%

▼ Estimated New Cases in US in 2016


Females



Breast	207,090	28%
Lung & bronchus	105,770	14%
Colon & rectum	70,480	10%
Uterine corpus	43,470	6%
Thyroid	33,930	5%
Non-Hodgkin lymphoma	30,160	4%
Melanoma of the skin	29,260	4%
Kidney & renal pelvis	22,870	3%
Ovary	21,880	3%
Pancreas	21,770	3%
All Sites	739,940	100%

▼ Estimated Death Cases in US in 2016

Females



Lung & bronchus	71,080	26%
Breast	39,840	15%
Colon & rectum	24,790	9%
Pancreas	18,030	7%
Ovary	13,850	5%
Non-Hodgkin lymphoma	9,500	4%
Leukemia	9,180	3%
Uterine Corpus	7,950	3%
Liver & intrahepatic bile duct	6,190	2%
Brain & other nervous system	5,720	2%
All Sites	270,290	100%

Background

- Unlike other cancers, mortality rates for OC have declined only slightly since 1971(War on cancer)
 - ∴ the unavailability of early detection tests and treatments*
(Only 15 percent of ovarian cancer patients are diagnosed early.)


Objective

- To develop and explore a new testing modality for an early detection of OC

Samples

- NCI PBS data banks
- **162** Ovarian cancer(OC) patients and **91** Control subjects
- Protein mass spectral data obtained by PBSII SELDI-TOF mass spectrometer
- Data: Intensities at distinct M/Z values
(the range of M/Z: 0.0000786 to 19995.513)

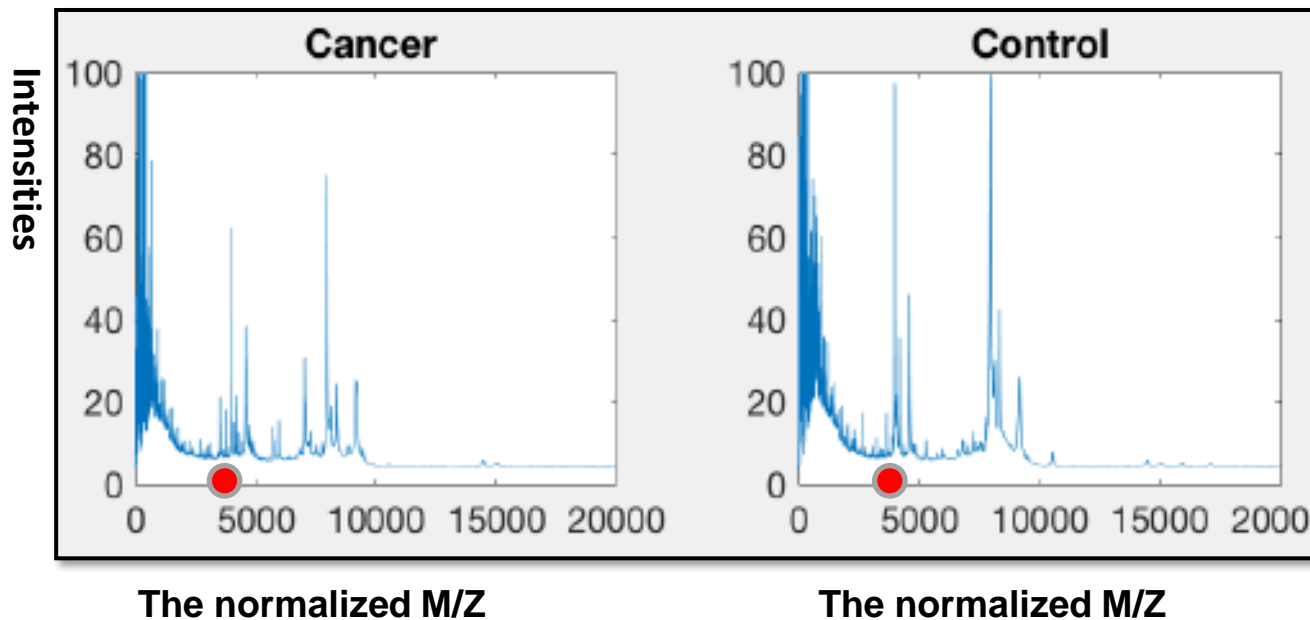
15,154 Intensity x (162 OC patients/ 91 control subjects)



M/Z	-7.86E-05	2.18E-07	9.60E-05	0.000366	0.00081	0.001429	0.002221	0.003188	0.004329	0.005644	0.007133
Intensity	4.1689291	4.13273	4.096531	4.154852	4.054299	3.971845	3.919558	4	4.034188	4.040221	4.106586

Samples

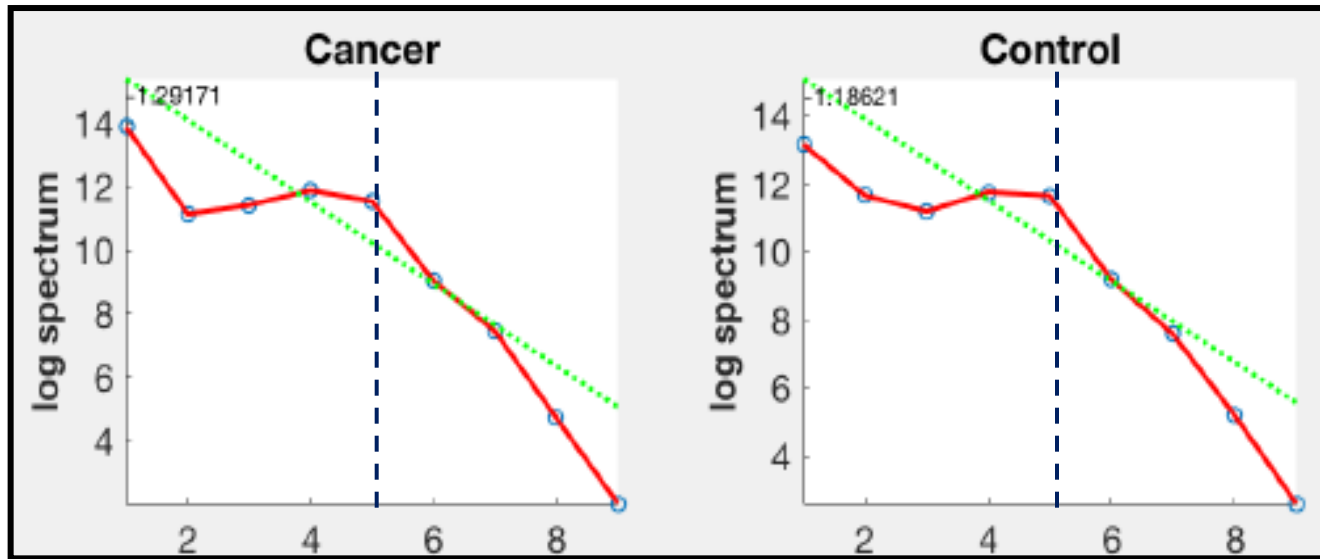
- But, there are lots of fluctuation of the intensities between 0 to 4000 of M/Z
→ we use the intensity data from 4001 to 13001.



▲ Protein mass spectra for the case(left) and control(right)

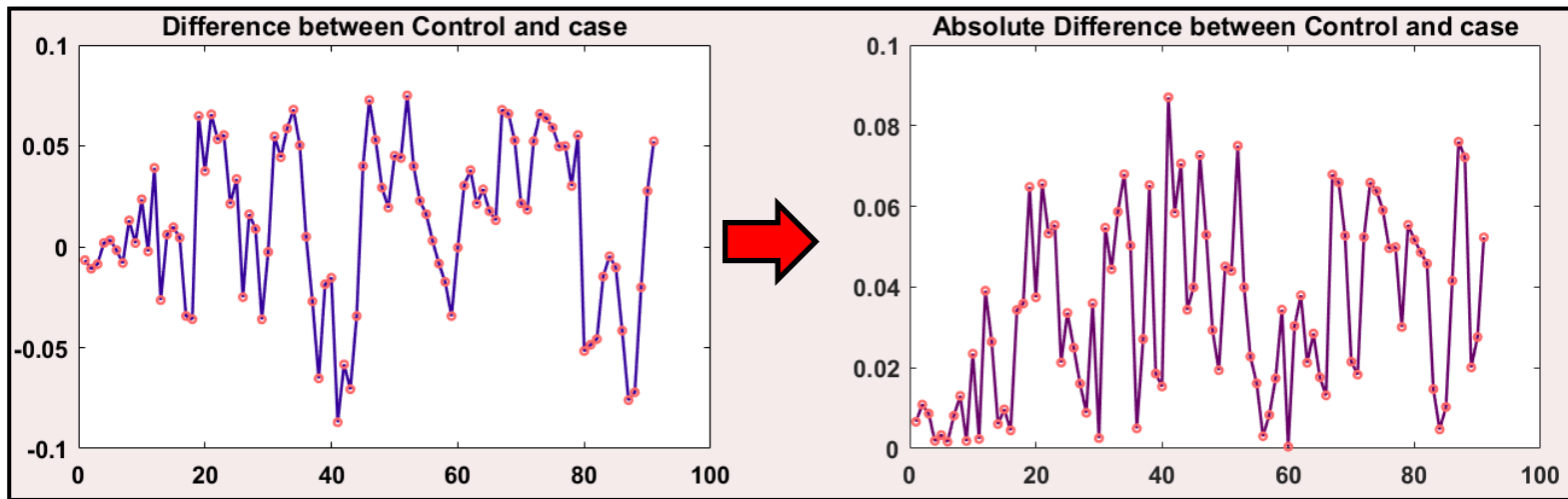
SLOPE USING WAVELET SPECTRUM

1. Using Wavelet based log-spectrum with a specific filter, we estimated the slopes for 2^{10} data packet
We created a set of 91 slope estimations for each samples(91slopes*(162+91))
(window size : 100)
2. Log-spectrums are regularly decreasing after dyadic level 5.
 \therefore we used the log spectrum from level 5 to 9.



↑ Slope estimation with wavelet levels from $k_1=5$ to $k_2=9$

3. After obtaining the Hurst exponents matrix, we calculated the absolute difference of means between Cancer and Control case.
4. We ran a logistic regression with 10 variables selected based on the absolute difference.

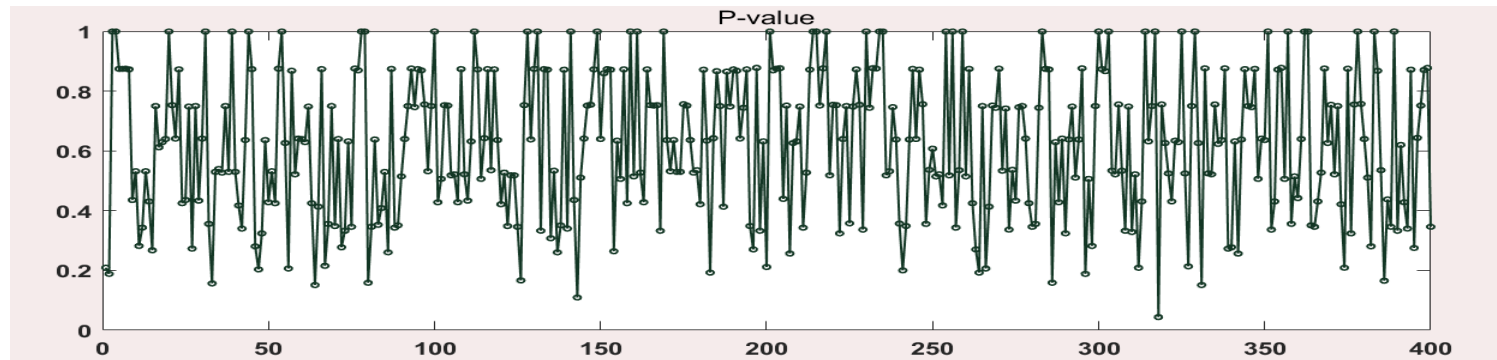


5. We randomly selected 67% of the data as a training data in order to create a classifier. Then we used the remaining 33% of the data to test performance
6. We fitted a logistic regression model on training data and classified the test data 1000 times

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
Intercept	-42.34	15.338	-2.761	0.0057***
v1	15.64	7.553	2.071	0.0383**
v2	-14.939	20.57	-0.726	0.4676
v3	-35.274	10.628	-3.319	0.0009***
v4	23.604	10.137	2.328	0.0198**
v5	42.003	23.764	1.767	0.0771*
v6	46.256	14.345	3.225	0.0012***
v7	-22.621	7.793	-2.903	0.0036***
v8	-25.019	34.985	-0.715	0.4745
v9	17.168	34.244	0.501	0.6161
v10	1.886	4.188	0.45	0.6525
(1%***, 5%** ,10%*)				

7. Additionally, we needed to test whether the true responses (Cancer:1 and Control:0) are same as the fitted values, through the Wilcoxon Sum Rank procedure 1000 times

TRUE	1	0	1	1	1	0	0
FITTED	1	0	0	1	1	1	0



Obtained 1000 p-values, the mean of p-values are $0.601 > 0.05$
∴ We failed to reject the null-hypothesis.
∴ We can conclude that the fitted values are good enough.

8. We tested 5 filter with 1000 steps, and chose the Haar filter.

	Mean Accuracy	Mean Sensitivity	Mean Specificity	Mean p-val
Haar	0.9059	0.901	0.9045	0.6205
Symmlet6	0.8984	0.8987	0.8999	0.6173
Symmlet8	0.8779	0.8794	0.8772	0.5795
Daubechies6	0.8996	0.8943	0.9107	0.6033
Daubechies8	0.8322	0.8537	0.7965	0.5911

		True	
		1	0
Predicted	1	46	3
	0	4	32

Confusion Matrix – Validation Sample

Mean Accuracy = 90.59%
 Mean Sensitivity = 90.1%
 Mean Specificity = 90.45%
 Logit link for binomial data with 0.65

Conclusion

- We achieved 90.6% accurate classification rate.
- We conducted classification analysis based on the wavelet based log-spectrum.
- From the logistic classification model, we found that Hurst exponent has the ability to discriminate cancer and control subjects.
- We can validate this method to find patterns of reproducible diagnostic value and contribute to the Ovarian cancer analysis.

1. Ovarian Cancer National Alliance

<https://ocrfa.org/wp-content/uploads/2016/08/OCRFA-Statistics-2016>

2. JUNG, Y. Y., PARK, Y., JONES, D., ZIEGLER, T., and VIDAKOVIC, B. (2010). Self-similarity in NMR Spectra: An Application in Assessing the Level of Cysteine, Journal of Data Science, 8,1, 1 – 19.

3. NICOLIS, O., RAMÍREZ, P., and VIDAKOVIC, B. (2011). 2-D Wavelet-Based Spectra with Applications. Computational Statistics & Data Analysis, 55, 1, 738–751.

Thank you for your attention!