

Editorial Manager(tm) for Pharmaceutical Research
Manuscript Draft

Manuscript Number:

Title: Molecular Factor Computing for Predictive Spectroscopy

Article Type: Research Paper

Section/Category: Other

Keywords: Near infrared spectroscopy (NIR); Optical computing; Chemometrics; Multivariate analysis; Genetic algorithm

Corresponding Author: professor Robert Lodder, Ph.D.

Corresponding Author's Institution: University of Kentucky

First Author: Bin Dai

Order of Authors: Bin Dai; Aaron Urbas; Craig Douglas; Robert Lodder, Ph.D.

Manuscript Region of Origin:

Abstract: The concept of molecular factor computing (MFC)-based predictive spectroscopy was demonstrated here with quantitative analysis of ethanol-in-water mixtures in a MFC-based prototype instrument. Molecular computing of vectors for transformation matrices enabled spectra to be represented in a desired coordinate system. New coordinate systems were selected to reduce the dimensionality of the spectral hyperspace and simplify the mechanical/electrical/computational construction of a new MFC spectrometer employing transmission MFC filters. A library search algorithm was developed to calculate the chemical constituents of the MFC filters. The prototype instrument was used to collect data from 39 ethanol-in-water mixtures. For each sample, four different voltage outputs from the detector (forming two factor scores) were measured by using four different MFC filters. Twenty samples were used to calibrate the instrument and build a multivariate linear regression prediction model, and the remaining samples were used to validate the predictive ability of the model. Cross validation yielded a standard error of prediction (SEP) of 0.735% for quantification of ethanol in water. Performance of this MFC-based instrument was compared with

the performance of conventional spectrometer that employed a principal component regression (PCR) calibration model.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Molecular Factor Computing for Predictive Spectroscopy

Bin Dai¹, Aaron Urbas¹, Craig C. Douglas² and Robert A. Lodder*

Department of Pharmaceutical Sciences
University of Kentucky
Lexington, KY 40506-0055

*Author to whom correspondence should be addressed. Email: Lodder@uky.edu
Telephone: 859-257-9232

1. Department of Chemistry, University of Kentucky
2. Department of Computer Science, University of Kentucky

Prepared for submission to *Pharmaceutical Research*.

1
2
3
4
5
6 **ABSTRACT**
7

8
9 The concept of molecular factor computing (MFC)-based predictive spectroscopy was
10 demonstrated here with quantitative analysis of ethanol-in-water mixtures in a MFC-
11 based prototype instrument. Molecular computing of vectors for transformation
12 matrices enabled spectra to be represented in a desired coordinate system. New
13 coordinate systems were selected to reduce the dimensionality of the spectral
14 hyperspace and simplify the mechanical/electrical/computational construction of a
15 new MFC spectrometer employing transmission MFC filters. A library search
16 algorithm was developed to calculate the chemical constituents of the MFC filters.
17 The prototype instrument was used to collect data from 39 ethanol-in-water mixtures.
18 For each sample, four different voltage outputs from the detector (forming two factor
19 scores) were measured by using four different MFC filters. Twenty samples were
20 used to calibrate the instrument and build a multivariate linear regression prediction
21 model, and the remaining samples were used to validate the predictive ability of the
22 model. Cross validation yielded a standard error of prediction (SEP) of 0.735% for
23 quantification of ethanol in water. Performance of this MFC-based instrument was
24 compared with the performance of conventional spectrometer that employed a
25 principal component regression (PCR) calibration model.
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

50 **KEY WORDS:** Near infrared spectroscopy (NIR); Optical computing;
51
52
53 Chemometrics; Multivariate analysis; Genetic algorithm.
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7 **INTRODUCTION**
8

9 Near infrared spectroscopy (NIR) has become an important process analytical
10 method for simultaneous multicomponent chemical analysis. NIR has found many
11 applications in process environments and in measurements in the biotechnology and
12 pharmaceutical industries (1-5), where NIR spectroscopy provides online,
13 nondestructive and noninvasive sensing.
14
15
16
17
18
19

20 Industrial environments are usually less friendly to analytical instrumentation than
21 research laboratories. Filter instruments are usually much more stable and rugged than
22 their dispersive or interferometric counterparts, making them ideally suited for the
23 harsh conditions found in industrial environments.(6, 7)
24
25
26
27
28

29 Multivariate calibration is a well-established tool in chemometrics for analysis of
30 NIR, UV-Visible, and Raman spectra. Conventional measurement of chemical or
31 physical properties from spectra is carried out by constructing a predictive model.(8-
32 10) Two of the most commonly used methods to construct a predictive model are
33 partial least squares (PLS) and principal component regression (PCR). In a
34 conventional spectrometer with typical chemometrics, data collection and processing
35 of raw data can be time consuming and computationally expensive, especially when
36 spatial relationships (image data) are required. Methods for selecting small but highly
37 relevant variables to represent the original data in a reduced coordinate space and
38 methods for integrated sensing and processing (ISP) are therefore receiving much
39 attention.(11, 12)
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54 ISP aims to design and optimize sensing systems that integrate the traditionally
55 independent units of sensing, signal processing, communication and targeting. By
56 employing ISP, computational complexity within traditional sensing system has been
57
58
59
60
61
62
63
64
65

1
2
3
4 substantially reduced through determining efficient low-dimensional representations
5
6 of those sensing problems that were originally posed in high-dimensional settings by
7
8 traditional sensing architecture. Successful ISP is expected to yield entirely new ways
9
10 of designing and operating sensor systems.(13)
11

12
13 One approach currently being investigated to simplify both instrumentation and
14
15 computational analysis involves optical pattern encoding.(14) This technique
16
17 involves tailoring the optical spectrum of filters to encode high level information
18
19 about the samples in sensing stage. Theoretical treatment of this methodology can be
20
21 found in the literature.(15, 16) Myrick et al. have demonstrated some practical
22
23 applications of this methodology in UV-visible and NIR spectroscopy.(17-24)
24
25 Encoding applications were based on the fabrication of thin film solid-state optical
26
27 filters, termed multivariate optical elements (MOEs). MOEs were designed to
28
29 replicate the multivariate regression pattern by transmitting and reflecting weighted
30
31 optical signals over a broad wavelength band.
32
33
34

35
36 Recent publications from our laboratory have offered an alternative approach for
37
38 spectral encoding(25, 26). Molecular absorption filters can be used as mathematical
39
40 factors in spectral encoding to generate a factor-analytic optical calibration in a high-
41
42 throughput spectrometer, which we term molecular factor computing (MFC). The
43
44 molecules in the filter effectively compute the calibration function by weighting the
45
46 signals received at each wavelength over a broad range of wavelengths. One or more
47
48 molecular filters are used in MFC-based spectrometer to produce detector signals
49
50 correlated to desired sample information. Advantages of this new approach over
51
52 conventional spectroscopy include significantly reducing the computational demand
53
54 (the integrated sensing and processing, or ISP, advantage), shorter data collection and
55
56 analysis time with higher signal-to-noise ratio (S/N) (especially for imaging
57
58
59
60
61
62
63
64
65

1
2
3
4 spectrometry, through the Fellgett advantage), higher optical throughput (the
5
6 Jacquinot advantage), and more rugged instrumentation with a considerably lower
7
8 cost.
9

10
11 This report describes the instrumentation and application of a molecular factor
12
13 computing-based spectrometer. Such a spectrometer may be particularly useful in
14
15 applications where real-time video analyses of remote sensing data are required. In
16
17 such cases, molecular filters placed in front of near-IR cameras would produce images
18
19 in which the intensities were proportional to the factor scores, without the need for
20
21 additional computation. Ethanol in water mixtures were selected as training and
22
23 validation samples to design molecular filters that would test the concept of MFC-
24
25 based spectroscopy. Ethanol is used in liquid pharmaceuticals to enhance solubility,
26
27 for example. Ethanol is also sometimes abused in the general population. Sensing
28
29 alcohol in the environment is necessary in such an application to evaluate the
30
31 effectiveness of pharmacotherapy or other therapies for alcohol abuse.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

MATERIALS AND METHODS

Traditional NIR Training Spectra Collection. The ethanol was reagent grade, obtained from AAPER (Shelbyville, KY). The water was distilled in house. Quantitative mixtures of alcohol and water were prepared by volume using grade-A volumetric flasks and burettes. Twenty samples were prepared with ethanol concentrations ranging from 0% to 14%. Sample solutions were placed in a quartz cuvette with a path length of 1 mm. Conventional near-infrared transmission spectra for comparison with MFC were collected using a dispersive spectrometer (Ocean Optics NIR256 temperature-regulated NIR, Dunedin, FL) over the wavelength range of 900-2500 nm to acquire a total of 256 data points per spectrum. Data analysis was limited to 1400-2200 nm to avoid the short wavelength region, which was not used for the MFC tests. As a result, the selected transmission spectra included 117 data points. The sample temperature remained constant at 25 °C during the data collection period. Each recorded spectrum was the average of 10 scans, with the total integration time ca. one second. The transmission spectra are shown in Figure 1a. To calculate the required composition of MFC filters for ethanol determinations, full spectra of ethanol/water mixtures over the wavelength range of interest must be available. For maximum accuracy, these spectra must represent the optical characteristics of the MFC spectrometer, not the conventional instrument. As a result, the transmission spectra of the dispersive spectrometer were convolved with the transmission spectra of a 1400-nm long pass filter, the emission spectrum of the tungsten NIR source, and the response curve of the InGaAs photodiode in the prototype instrument to give a corrected representation of the MFC instrument response. These corrected transmission spectra were used as training spectra for MFC

1
2
3
4 filters selection and multivariate analysis. The corrected spectra are presented in
5
6 Figure 1b.

7
8 ***MFC-based High Throughput NIR Spectrometer.*** A graphic representation of
9
10 the instrumental setup is given in Figure 2. A 12V, 100W tungsten-halogen
11
12 broadband source (model 621, McPherson Inc., Chelmsford, MA) with 1400-nm long
13
14 pass filter (Thorlabs, Newton, NJ) was used as the source of broadband NIR light.
15
16 The tungsten-halogen light source has more intense radiation in the shorter NIR
17
18 wavelength region. To avoid saturating the detector with short wavelength NIR
19
20 radiation that contains little chemical information about the samples, the 1400 nm
21
22 long pass filter was used to block the short wavelength radiation. The source beam
23
24 was modulated with an optical chopper (Model SR540, Stanford Research Systems
25
26 Inc., Sunnyvale, CA) at a frequency of 280 Hz. The light beam was focused onto an
27
28 InGaAs photodiode (Fermionics Opto-Technology, Simi Valley, CA) through a
29
30 convex lens after passing through the molecular filter cuvette and sample cuvette. A
31
32 step-indexed sliding cuvette tray was constructed in-house that permitted manual
33
34 selection of cuvettes in the beam path. All cuvettes used for holding the liquid MFC
35
36 filters were 2 mm path length optical glass. The sample cuvette had 1 mm path length.
37
38 A two-factor spectrum from a sample consisted of four data points because the
39
40 positive and negative factor loadings were represented by separate molecular filter
41
42 mixtures. Thirty-nine ethanol-in-water mixtures were scanned with the MFC-based
43
44 spectrometer. Twenty samples were used to calibrate the instrument and build a
45
46 multivariate linear regression prediction model, and the remaining samples were used
47
48 to validate the predictive ability of the model. To avoid possible false responses due
49
50 to instrument drift, samples were measured in a random order. The sample
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 temperature was held constant at 25 °C during the data collection period. A 3-second
5
6 integration was employed at each MFC filter.
7

8 **Data Analysis.** All data analysis was carried out using Matlab 7.0 (Mathworks,
9
10 Inc., Natick, MA). The PLS toolbox v3.51 for Matlab (Eigenvector Research, Inc.
11
12 Wenatchee, WA) was used for multivariate analysis. A genetic algorithm and direct
13
14 search toolbox for Matlab were used to perform the NIR library search to generate
15
16 combinations of liquids for use as MFC filters.
17
18

19
20 **Theory.** As illustrated in Figure 2, using the MFC approach, traditional bulky
21
22 multi-channel wavelength selection devices such as gratings and moving mirrors are
23
24 replaced with simple MFC filters. Only a light source, detector and MFC filters are
25
26 needed to construct a minimal MFC-based spectrometer. The weighted combination
27
28 of spectral responses from the filters is designed to match the regression vector from
29
30 transmission spectra-based factor methods like PCR or PLS calibration. Because a
31
32 multivariate regression vector can be positive or negative while all transmission
33
34 spectra of MFC filters are positive, two distinct MFC filters are employed to represent
35
36 accurately the multivariate regression vector. Depending on the complexity of the
37
38 regression vector and availability of MFC filter materials, an exact match of the
39
40 regression vector and availability of MFC filter materials, an exact match of the
41
42 regression vector to the filter might be very difficult. Fortunately, an exact match is
43
44 not absolutely necessary, for reasons that are addressed in MFC filters selection. For
45
46 each MFC filter, the signal produced at the detector is a dot product of the filter
47
48 transmission spectrum and the sample transmission spectrum, with a signal offset
49
50 v_{offset} in practice(23).
51
52

$$v_{out} = G \times \vec{s} \cdot \vec{f} + v_{offset} \quad (1)$$

53
54
55 v_{out} is the output voltage, G is the constant amplifier gain, f represents the MFC filter
56
57 spectrum vector, and s represents the corrected sample spectrum vector.
58
59
60
61
62
63
64
65

1
2
3
4 For m samples and n filters, V_{out} (m by n) is output voltage matrix.

$$V_{out} = G \times SF^T + V_{offset} \quad (2)$$

5
6
7
8
9 where F (n by k) is the transmission spectra matrix of MFC filters, and S (m by k) is
10 transmission spectra matrix of samples.
11

12
13
14 The vector of concentration values, Y (m by l), of the training samples are
15 predicted by multivariate linear regression (MLR) according to Eq. 3:
16

$$\hat{Y} = V_{out} C + E = G \times SF^T C + Offset \quad (3)$$

17
18
19
20
21 where C (n by l) are the regression coefficients, E is a scalar, m is the number of
22 training samples, and n is the number of MFC filters.
23
24

25
26 After MFC filters were selected and the regression coefficients R obtained,
27

$$R = F^T C \quad (4)$$

28
29
30 this R works in a similar fashion to PCR loadings .
31

$$\hat{y}_i = G \times S_i F^T C + offset = G \times S_i R + offset \quad (5)$$

32
33 For m training samples, the root-mean-square error of calibration (RMSEC)(23) is
34

$$RMSEC = \left[\sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{m} \right]^{1/2} = \left[\sum_{i=1}^m \frac{(G \times S_i R + offset - y_i)^2}{m} \right]^{1/2} \quad (6)$$

35
36
37
38
39
40
41
42
43 The minimum RMSEC is reached by searching a NIR spectral library to select the
44 best molecules for MFC filters. G and $offset$ are the parameters adjusted after the
45 MFC filters have been chosen. While one could select MFC filter molecules to match
46 a regression vector that provides a fixed RMSEC specified *a priori*, searching the NIR
47 library to find a combination of MFC filters that minimizes the RMSEC is usually
48 more desirable. A perfect spectral match may require a large number of different
49 filters molecules or filter molecules that are not available in the library.
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 ***Spectral Region Selection.*** Theoretically, the MFC-based spectrometer approach
5
6 should function in any spectral region where molecular filters are available. For this
7
8 research, the NIR spectral region was used because NIR spectrometry is a widely
9
10 employed PAT and ethanol has a significant absorbance between two water
11
12 absorbance bands in the NIR region between 1400 and 2200 nm.
13
14

15 ***Radiometric Correction.*** The multivariate prediction of analyte concentration
16
17 using MFC is inherently radiometric in nature. Radiometric measurement is based on
18
19 a detector response that is directly related to sample transmission instead of
20
21 absorbance. Of course, sample concentration is linearly related to absorbance when
22
23 Beer's law holds and transmission is logarithmically related to sample concentration.
24
25 In a low absorbance regime, transmission relates to concentration approximately
26
27 linearly, however, in a higher absorbance regime, the nonlinear relationship between
28
29 concentration and transmission predominates. To model both regimes in transmission
30
31 mode, extra principal components or latent variables have to be used in a linear
32
33 multivariate calibration model. (27)
34
35
36
37

38 Before using transmission spectra to perform a library search for MFC filter
39
40 constituent selection, the transmission spectra were corrected for unique optical
41
42 characteristics of the MFC spectrometer. Data provided by manufacturers' test sheets
43
44 were used to form the correction factors. In the experimental MFC system, the
45
46 radiometric correction was performed by convolving the transmission spectra with the
47
48 emission spectrum of the source lamp, the transmission spectra of the 1400 nm long
49
50 pass filter, and the response curve of the InGaAs photodiode in the prototype
51
52 instrument. Thus, the corrected transmission spectra represented an unbiased detector
53
54 response as a function of wavelength. The corrected spectra in Figure 1b revealed
55
56 that the transmission of the spectrometer is not completely cut off at 1400 nm. The
57
58
59
60
61
62
63
64
65

1
2
3
4 transmission of the actual 1400-nm long pass filter employed was approximately 25%
5
6 at 1400 nm. However, the transmission was much lower at shorter wavelengths and
7
8 was less than 1% at 1370 nm. Because the variation of the sample spectra from 1370
9
10 nm to 1400 nm was small, the effects of the slightly wider bandpass on prediction of
11
12 sample composition were negligible.
13
14

15 **MFC Filter Selection.** The chemicals chosen as MFC filters were found by
16
17 searching a library of near-IR transmission spectra containing 1923 compounds (John
18
19 Wiley & Sons, Inc.). The library consisted of two spectra of each compound
20
21 collected over slightly overlapping regions, 952-1587 nm and 1388-2630 nm. Because
22
23 the coverage of the MFC system is 1400-2200 nm, only the spectra from 1388-2630
24
25 nm were used in the library search. Molecular factor scores were calculated from the
26
27 product of the transmission spectra from the NIR spectral library and the corrected
28
29 transmission spectra of ethanol / water mixtures:
30
31

$$32 \quad U_{m \times l} = S_{m \times k} L_{l \times k}^T \quad (6)$$

33
34 where U is the score matrix, L is the transmission spectra of the NIR library, S is the
35
36 corrected transmission spectra of training samples, l is number of compounds in the
37
38 library ($l=1923$), m is number of training spectra ($m=20$), and k is the number of
39
40 wavelength values in the spectra ($k=117$). A modified genetic algorithm (28) was
41
42 used to search the score space to find four filters that yielded a predictive model with
43
44 the lowest root mean square error of cross validation (RMSECV). The RMSECV
45
46 function was used as the fitness function of the genetic algorithm. The genetic
47
48 algorithm library search was performed 50 consecutive times. Due to the indefinite
49
50 nature of the genetic algorithm, each time the search routine produced a somewhat
51
52 different MFC filter combination, but roughly the same RMSECV. Four common
53
54 chemicals were selected as molecular filters: water, methanol, ethanesulfonic acid,
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 and 2,2-diethoxypropane. The transmission spectra of these chemicals are shown in
5
6 Figure 3. These four MFC filters gave an adequate calibration model ($r^2=0.995$,
7
8 RMSEC=0.229%, RMSECV=0.339%, $p=0.05$ by f test). This result is slightly better
9
10 than a corresponding PCR calibration model based on corrected transmission spectra
11
12 ($r^2=0.993$, RMSEC=0.359%, RMSECV=0.551%, $p=0.05$ by f test). The PCR
13
14 regression vector and simulated regression vector based on MFC filters are both
15
16 presented in Figure 4. It is evident in Figure 4 that these two regression vectors do
17
18 not match. The search for a regression vector by genetic algorithm is intended to
19
20 reach a minimum on a multidimensional response surface. The PCR regression vector
21
22 is one of many such minima, and it can be visualized as a point in a reduced
23
24 orthogonal p-factor space that describes a linear relationship between the spectra and
25
26 concentration. Due to the stochastic nature of the genetic algorithm, it is possible to
27
28 obtain several essentially equivalent solutions to the optimization problem. Therefore,
29
30 it is not surprising that the regression vector generated by MFs did not match a
31
32 predefined PCR regression vector. Such a pattern match is unnecessary. Indeed, the
33
34 fact that multiple solutions exist makes it easier to find molecular filters that are stable
35
36 and compatible with other molecules in the filter system.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

RESULTS AND DISCUSSION

Analysis of Ethanol in Water Mixtures. Multivariate Analysis of Absorbance and Transmission Spectra. In order to compare the results of MFC measurements with the results of multivariate analysis using a conventional spectrometer, PCR was performed on the same training set used for selection of MFC filters. First, the PCR calibration was performed with corrected transmission spectra. The optimum predictive model was defined as the model with lowest RMSECV by leave-one-out cross validation. Four principal components were required to build a calibration model with optimum predictive ability. Theoretically, two principal components should be sufficient to model the ethanol/water mixtures. The extra principal components were included due to the nonlinear response between the transmission spectra and concentration. The RMSEC was 0.359%, corresponding to 2.56% error relative to the range of the calibration set. The four PCs model was validated by leave-one-out cross validation, and the RMSECV was 0.551%, or 3.93% relative to the range of the calibration set. Next, a PCR calibration was carried out on absorbance spectra, which were calculated from original transmission spectra (using $A=1/\log T$). Three principal components were required to build an optimum calibration model. The RMSEC was 0.309%, corresponding to 2.20% error relative to the range of the calibration set. The three PCs model was validated by leave-one-out cross validation, and the RMSECV was 0.494%, or 3.53% relative to the mean of the calibration set. Compared to the PCR model based on corrected transmission spectra, the PCR model based on absorbance spectra required fewer principal components and had a slightly lower RMSEC and RMSECV. The better performance of the model based on absorbance spectra is expected because of the linear response between absorbance and concentration.

1
2
3
4 *Expectation from Simulation.* As described in the MFC filters selection section,
5
6 the simulation study for the MFC filters predicted a RMSEC of 0.229% and
7
8 RMSECV of 0.339% with corrected transmission spectra. . The result showed that the
9
10 PCR model is not necessarily the best model. MFC filters outperform the traditional
11
12 scanning PCR model in terms of RMSECV. The simulation result in Figure 5 shows a
13
14 plot of the predicted ethanol concentrations versus the actual ethanol concentrations
15
16 using a MLR model based on 4 MFC filters and a 4-component PCR model based on
17
18 corrected transmission spectra.
19
20

21
22 *Determination of Ethanol with the MFC Approach.* The voltage output from the
23
24 detector was recorded for each of 39 samples through each MFC filter. The samples
25
26 were split into two groups for cross validation, and 20 samples were used to calibrate
27
28 the MFC-based instrument, while the other 19 samples were used as the validation
29
30 dataset. The 20 calibration samples were different from those samples used as training
31
32 samples for the MFC filters selection, but were prepared at the same nominal
33
34 concentrations. Additional calibration was necessary because the correction factors
35
36 used with the training spectra were all obtained from manufacturer's test datasheets
37
38 and set-ups, and might be different once assembled in the prototype instrument. The
39
40 optimal correlation between detector output voltage and ethanol concentration were
41
42 obtained by following equation.
43
44
45
46

$$47 \hat{Y} = \begin{bmatrix} v_{out[1,1]}v_{out[1,2]}v_{out[1,3]}v_{out[1,4]} \\ \dots\dots\dots \\ \dots\dots\dots \\ v_{out[m,1]}v_{out[m,2]}v_{out[m,3]}v_{out[m,4]} \end{bmatrix} \begin{bmatrix} -28567 \\ -14368 \\ 27997 \\ 21164 \end{bmatrix} - 34 \quad (7)$$

48
49
50
51
52
53
54
55 Where \hat{Y} was the predicted ethanol concentration, and v_{out} was the voltage output
56
57 of each sample for each MFC filter. The RMSEC of the model was 0.748%, and the
58
59
60
61
62
63
64
65

1
2
3
4 RMSEP by data splitting was 0.735%. Figure 6 shows a plot of predicted ethanol
5
6 concentrations versus actual ethanol concentrations of all 39 samples.
7

8
9 The estimated RMSEP (0.735%) of the MFC-based measurement was not as good
10 as the RMSEP (0.339%) predicted by the simulation. Still, the actual MFC result
11 shows that the MFC instrument is able to produce a useful numerical concentration
12 result. The difference between the simulated instrument and the actual instrument
13 results was due to several factors:
14
15
16
17
18

19
20 1. Sampling noise arose from the positioning of molecular filter cuvettes and/or
21 sample cuvettes that did not exist in the simulation.
22

23
24 2. The transmission spectra in the NIR library were obtained with a path length of
25 2.5 mm, while cuvettes with a path length of 2 mm were used as MFC filters in the
26 prototype instrument. The difference in the profile of transmission spectra due to the
27 different path length likely increased prediction error.
28
29
30

31
32 3. Instrumental limitations prevented obtaining the exact transmission spectrum
33 of the 1400 nm long pass filter, the emission spectrum of the light source, and the
34 detector response curve in the prototype MFC-based instrument to correct the training
35 transmission spectra. Alternative correction factors were obtained from
36 manufacturers' datasheets. Better results might be expected if each individual optical
37 component in the prototype instrument were carefully calibrated.
38
39
40
41
42
43
44
45
46

47 4. Although an optical chopper and lock-in-amplifier were used to reduce noise
48 and thermal drift, the MFC-based prototype instrument was shown to have a
49 significant instrument drift. Simple studies with the light source (e.g. 1400 nm long
50 pass filter in place but without MFC chemicals or sample cell present) exhibited
51 signal drift as high as 4% relative over 20 minutes, which was roughly the time
52 required to scan all 39 samples in the MFC instrument. This significant drift could
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 contribute to the high RMSEP. Future studies will utilize a double-beam design to
5
6 eliminate this drift.

7
8 5. Using the genetic algorithm-based MFC filters selection algorithm, only the
9
10 predictive ability of the MLR model was considered in the fitness function. The
11
12 sensitivity of each individual MFC filter to changes in ethanol concentration was not
13
14 taken into account. PCR is a regression method based on orthogonal principal
15
16 components that maximize variance. However, MLR only aims to minimize the sum
17
18 of the squared errors, and variance maximization for dependent variables is not taken
19
20 into account. Therefore, the genetic algorithm-based MFC filters selection could
21
22 select MFC filters with high prediction ability but low sensitivity, which results a
23
24 hypothetical low RMSEP in the simulation study that is difficult to achieve with real,
25
26 physical filters. (A new search algorithm that takes both prediction and sensitivity
27
28 into account is currently being investigated.)
29
30
31
32

33
34 In addition to the multivariate regression model for ethanol concentration, an
35
36 estimate of the detection limit for binary mixtures of ethanol and water was also
37
38 calculated. The estimate was based on an extension of the BEST metric for sub-
39
40 cluster detection with sample populations that has been described previously (29, 30).
41
42 The experimental MFC data were then analyzed to estimate the limits of detection of
43
44 each component in binary mixtures of two components. This was performed by
45
46 translating the sample population mean of 1% ethanol in water sample towards pure
47
48 water sample population's mean until the two clusters could not be differentiated
49
50 using the BEST subcluster detection algorithm. The estimate of the detection limit for
51
52 ethanol in water determined by this procedure is 0.26%. The dynamic range for
53
54 ethanol detection by MFC was a factor of 57. The extended BEST metric provided
55
56 lower errors than traditional regression approaches because it took both changes in
57
58
59
60
61
62
63
64
65

1
2
3
4 sample cluster location as well as scale into account. However, to achieve its better
5
6 results the extended BEST requires multiple replicates of the same sample, which can
7
8 be impractical in real-life remote sensing applications.
9

10
11 In order to assess of the long-term stability of molecular filters, the molecular
12
13 filters were directly exposure to the near-IR light beam for 10 hours. For each of those
14
15 four molecular filters, the signal was continually monitored and variations in signal
16
17 level of 4% were observed in this study (the same range as the variation in the light
18
19 source intensity). The molecular filters were also sealed in cuvettes over two-month
20
21 period, and there appeared no visible degradation of these molecular filters over that
22
23 time. It is worth noting that, for some other molecular filters that were not used in
24
25 this study, severe degradation of MFs can be observed. Thus, it is necessary to
26
27 compile a spectral library using only stable molecules for MFC.
28
29

30
31 The susceptibility of MFC-based spectroscopic measurement to complex matrix
32
33 interference in samples is not well understood. Theoretically, the MFC-based
34
35 instrument should be able to precisely measure the specific chemical species of
36
37 interest as long as the potential interferences were introduced and modeled in the
38
39 training set. Future research will include determination of ethanol containing other
40
41 alcohols as interferences that are not in the training set to evaluate the susceptibility of
42
43 MFC to this sort of interference.
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

CONCLUSION

A prototype MFC-based spectrometer was designed, constructed, and tested for the analysis of ethanol-in-water mixtures. The concept of molecular factor computing was demonstrated. The results obtained from an MFC-based measurement were compared to PCR calibration based on conventional scanning spectrometry. Although the actual results from MFC-based prediction in the first prototype were slightly worse than from conventional PCR prediction, the MFC simulation study suggested that a better prediction model could be built based on MFC. A double-beam MFC instrument under construction may achieve the superior results predicted by the simulation. Advantages of the MFC approach over conventional spectroscopy include significantly reducing the computational demand (the integrated sensing and processing, or ISP, advantage), shorter data collection and analysis time with higher signal-to-noise ratio (S/N) (especially for imaging spectrometry, through the Fellgett advantage), higher optical throughput (the Jacquinot advantage), and more rugged instrumentation with a considerably lower cost. The high optical throughput of an MFC system could offer improved analytical ability in systems with a weak signal.

Problems with reproducibility in positioning of filter cuvettes and samples cuvettes increased measurement noise in the MFC-based prototype spectrometer. The effect will be reduced by using aperture control and through better design of slides for holding filters and samples.

A new library search algorithm should be developed to select the optimal MFC filters. Prediction ability and sensitivity of MFC filters both should be taken into account in the fitness function of genetic algorithm-based searches.

The number of potential filter materials is huge. Solutions and solid-state mixtures could both be used as molecular filters. The use of organic solvents as MFC filters

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

introduces some ruggedness problems for process analysis. To simplify the instrument and improve the system stability, solid-state MFC filters constructed from materials such as polymers may offer a good alternative to liquid filters (6).

MFC offers users a simpler ISP instrument with significant reduction of computational complexity and processing time at the cost of some experimental flexibility. In other words, MFC-based instruments are not general-purpose research tools. Instead, the MFC approach is for practical measurement in the real world where fast results are needed and achieved by integrating the processing into the sensing stage.

In addition to applications of this technique as a process analytical technology (PAT), MFC-based remote NIR imaging for real-time surveillance has gained interest. A MFC-based NIR imaging system for remote ethanol sensing is currently under construction in our laboratory. The range of possible applications is likely to expand when imaging systems are available.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation through CNS-0540178 and by the National Institutes of Health through N01AA 33003 and T32 HL072743.

References

1. R. J. Dempsey, D. G. Davis, R. G. Buice, and R. A. Lodder. Biological and medical applications of near-infrared spectroscopy. *Applied Spectroscopy* **50**: 18A-34A (1996).
2. J. K. Drennen and R. A. Lodder. Nondestructive near-infrared analysis of intact tablets for determination of degradation products. *J Pharm Sci* **79**: 622-7 (1990).
3. A. S. El-Hagrasy, H. R. Morris, F. D'Amico, R. A. Lodder, and J. K. Drennen, 3rd. Near-infrared spectroscopy and imaging for the monitoring of powder blend homogeneity. *J Pharm Sci* **90**: 1298-307 (2001).
4. A. Urbas, M. W. Manning, A. Daugherty, L. A. Cassis, and R. A. Lodder. Near-infrared spectrometry of abdominal aortic aneurysm in the ApoE-/- mouse. *Anal Chem* **75**: 3318-23 (2003).
5. T. D. Ridder, S. P. Hendee, and C. D. Brown. Noninvasive Alcohol Testing Using Diffuse Reflectance Near-Infrared Spectroscopy. *Applied Spectroscopy* **59**: 181-189 (2005).
6. M. R. Fischer and G. M. Hieftje. Near-IR multiplex bandpass spectrometer utilizing polymer filters. *Applied Spectroscopy* **50**: 1246-1252 (1996).
7. A. Fong and M. G. Hieftje. Near-IR Multiplex Bandpass Spectrometer Using Liquid Molecular Filters. *Applied Spectroscopy* **49**: 493-498 (1995).
8. K. R. Beebe and B. R. Kowalski. Introduction to multivariate calibration & analysis. *Anal Chem* **59**: 1007A-1017A (1987).
9. H. Martens and M. Martens. Multivariate Analysis of Quality An Introduction. (2001).
10. H. Martens and T. Naes. Multivariate Calibration. (1989).
11. R. Leardi. Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics* **14**: 643-655 (2000).
12. R. Leardi. Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets. *Data Handling in Science and Technology* **23**: 169-196 (2003).
13. C. Schwartz. Integrated Sensing and Processing <http://www.darpa.mil/dso/thrust/math/isp.htm>.
14. O. Soyemi, D. Eastwood, L. Zhang, H. Li, J. Karunamuni, P. Gemperline, R. A. Synowicki, and M. L. Myrick. Design and testing of a multivariate optical element: The first demonstration of multivariate optical computing for predictive spectroscopy. *Analytical Chemistry* **73**: 1069-1079 (2001).
15. S. E. Bialkowski. Species discrimination and quantitative estimation using incoherent linear optical signal processing of emission signals. *Analytical Chemistry* **58**: 2561-2563 (1986).
16. A. M. C. Prakash, C. M. Stellman, and K. S. Booksh. Optical regression: a method for improving quantitative precision of multivariate prediction with single channel spectrometers. *Chemometrics and Intelligent Laboratory Systems* **46**: 265-274 (1999).
17. F. G. Haibach, A. E. Greer, M. V. Schiza, R. J. Priore, O. O. Soysmi, and M. L. Myrick. On-line reoptimization of filter designs for multivariate optical elements. *Applied optics* **42**: 1833-1838 (2003).
18. F. G. Haibach and M. L. Myrick. Precision in multivariate optical computing. *Applied optics* **43**: 2130-2140 (2004).

19. M. L. Myrick, O. Soyemi, J. Karunamuni, D. Eastwood, H. Li, L. Zhang, A. E. Greer, and P. Gemperline. A single-element all-optical approach to chemometric prediction. *Vibrational Spectroscopy* **28**: 73-81 (2002).
20. M. L. Myrick, O. Soyemi, H. Li, L. Zhang, and D. Eastwood. Spectral tolerance determination for multivariate optical element design. *Fresenius' Journal of Analytical Chemistry* **369**: 351-355 (2001).
21. M. L. Myrick, O. O. Soyemi, F. Haibach, L. Zhang, A. Greer, H. Li, R. Priore, M. V. Schiza, and J. R. Farr. Application of multivariate optical computing to near-infrared imaging. *Proceedings of SPIE-The International Society for Optical Engineering* **4577**: 148-157 (2002).
22. M. L. Myrick, O. O. Soyemi, M. V. Schiza, J. R. Farr, F. Haibach, A. Greer, H. Li, and R. Priore. Application of multivariate optical computing to simple near-infrared point measurements. *Proceedings of SPIE-The International Society for Optical Engineering* **4574**: 208-215 (2002).
23. O. O. Soyemi, F. G. Haibach, P. J. Gemperline, and M. L. Myrick. Nonlinear optimization algorithm for multivariate optical element design. *Applied Spectroscopy* **56**: 477-487 (2002).
24. O. O. Soyemi, F. G. Haibach, P. J. Gemperline, and M. L. Myrick. Design of angle-tolerant multivariate optical elements for chemical imaging. *Applied Optics* **41**: 1936-1941 (2002).
25. L. A. Cassis, B. Dai, A. Urbas, and R. A. Lodder. In vivo applications of a molecular computing-based high-throughput NIR spectrometer. *Proc. SPIE-Int. Soc. Opt. Eng.* **5329**: (2004).
26. L. A. Cassis, A. Urbas, and R. A. Lodder. Hyperspectral integrated computational imaging. *Analytical and Bioanalytical Chemistry* **382**: 868-872 (2005).
27. P. Geladi and B. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta* **185**: 1-17 (1986).
28. E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M.-H. Tsou, C.-F. Horng, A. B. E. S. Iversen, M. Liao, C.-M. Chen, M. West, J. R. Nevins, and A. T. Huang. Gene Expression Predictors of Breast Cancer Outcomes. *Lancet* **361**: 1590-1596 (2003).
29. R. A. Lodder and G. A. Hieftje. Detection of Subpopulations in Near-Infrared Reflectance Analysis. *Applied Spectroscopy* **42**: 1500-1512 (1988).
30. Y. Zou, et al. Making Your Best Case - near-IR Spectral Identification of Soil. *analytical chemistry* **65**: A434-A439 (1993).

1
2
3
4 **FIGURE CAPTIONS.**
5

6
7 Figure 1.a Raw, uncorrected transmission spectra of 20 ethanol / water mixtures
8 acquired on a conventional dispersive NIR spectrometer.
9

10 Figure 1.b. Corrected spectral response function. These data are based on the
11 transmission spectra in Figure 1a, convolved with following radiometric vectors:
12 radiance spectrum of tungsten light source, the transmission spectrum of 1400 nm
13 long pass filter, and the response curve of the InGaAs photodiode.
14

15
16 Figure 2. A graphical representation of the MFC-based high throughput
17 spectrometer.
18

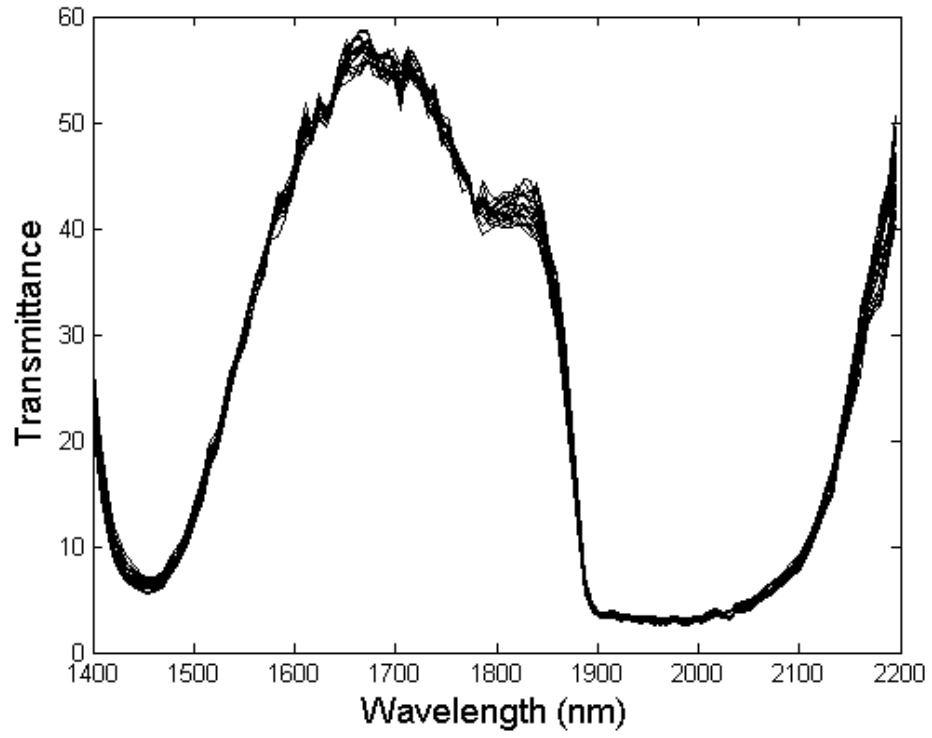
19 Figure 3. The transmission spectra of the selected MFC filters.
20

21 Figure 4. The regression vectors versus wavelength. The solid line shows the
22 PCR regression vector, and the dashed line shows the regression vector based on
23 MLR calibration of the MFC filter.
24

25
26 Figure 5. A plot of the predicted ethanol concentrations versus the actual ethanol
27 concentrations using a MLR model based on 4 simulated MFC filters and a PCR
28 model based on corrected transmission spectra. Stars: PCR model based on corrected
29 transmission spectra, RMSEC=0.359%, RMSECV= 0.551%. Circles: MLR model
30 based on 4 simulated MFC filters, RMSEC=0.229%, RMSECV=0.339%.
31
32

33 Figure 6. A plot of the predicted ethanol concentrations versus the actual ethanol
34 concentrations of all 39 samples. Diamonds: calibration samples, $r^2=0.968$,
35 RMSEC=0.748%. Crosses: validation samples, RMSEP=0.735%. Significant at
36 $p=0.05$ by f test.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Figure 1a.



b.

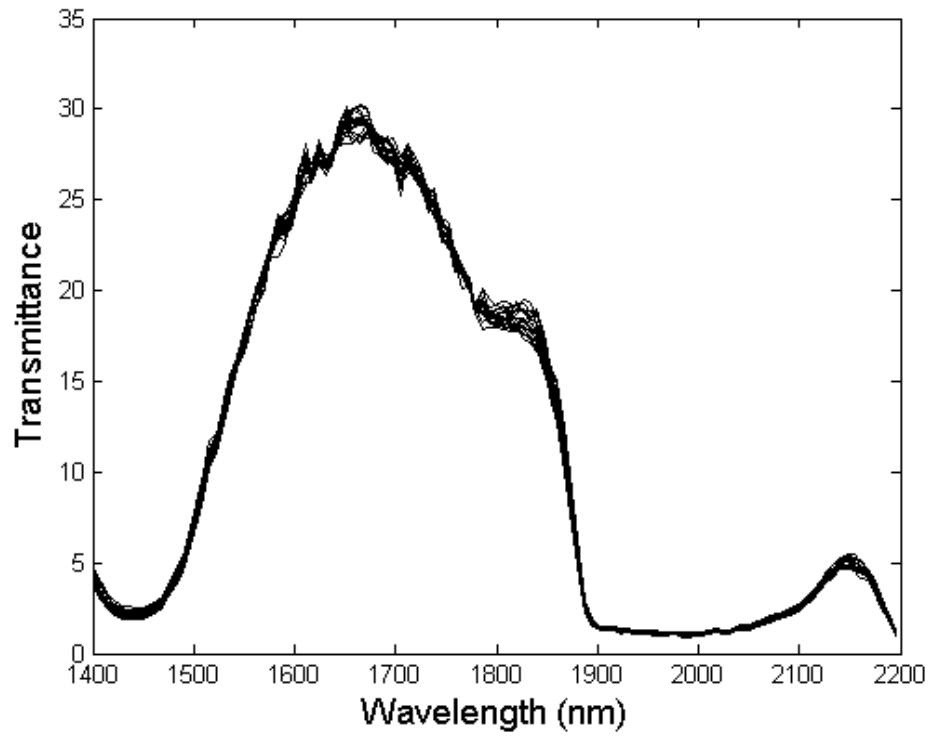


Figure 2.

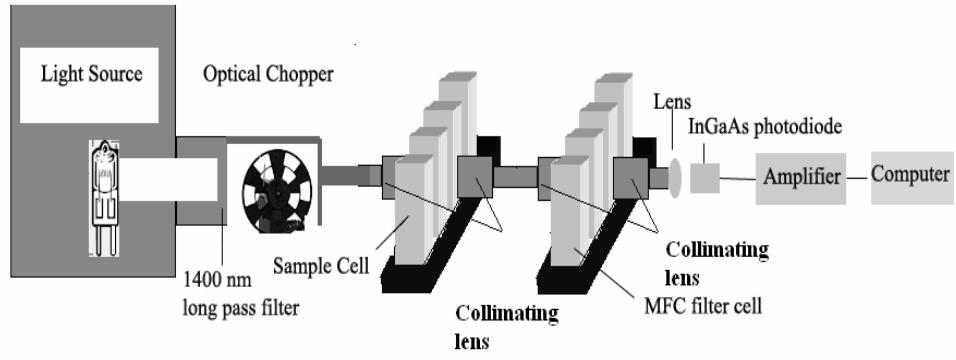
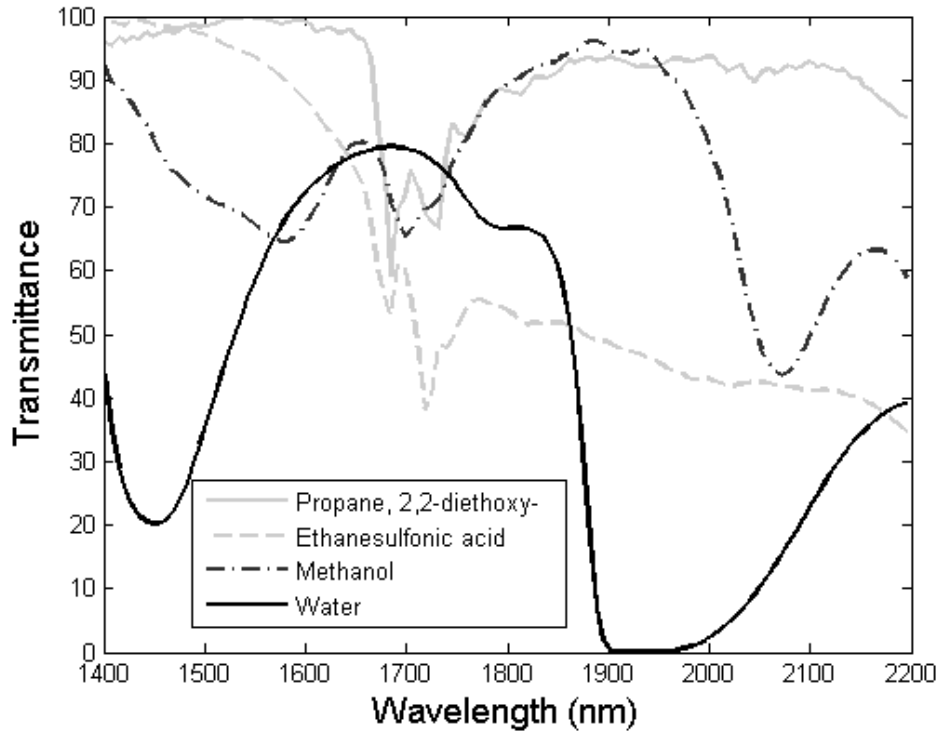
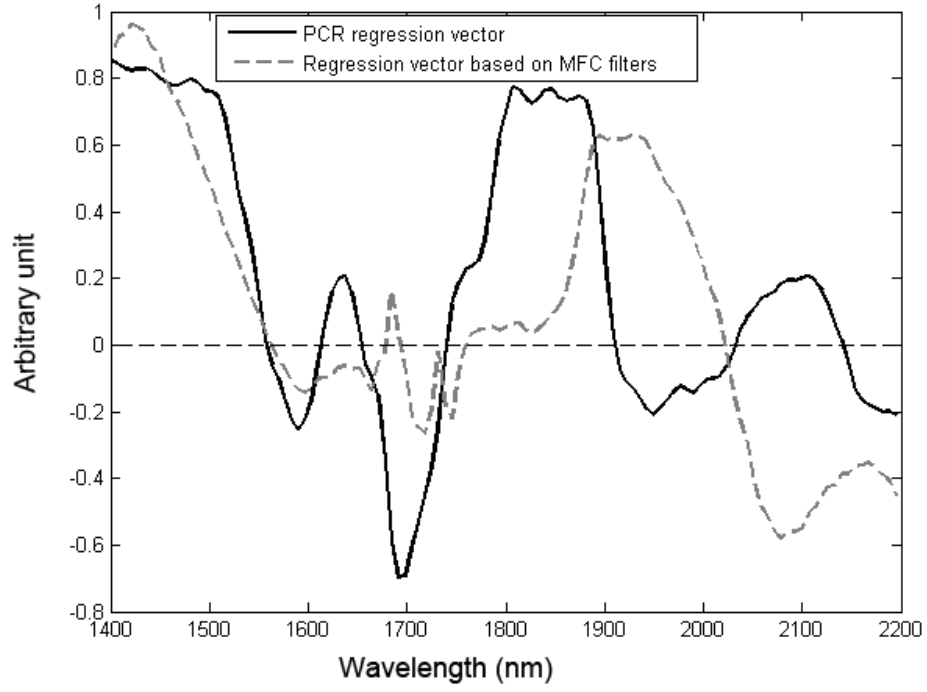


Figure 3



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Figure 4.



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Figure 5.

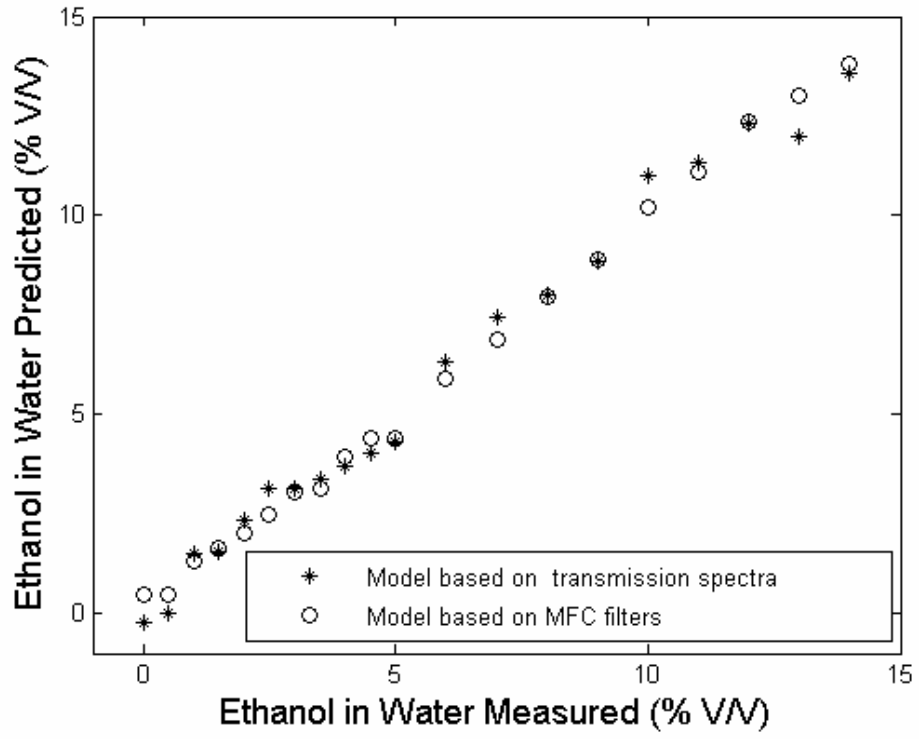
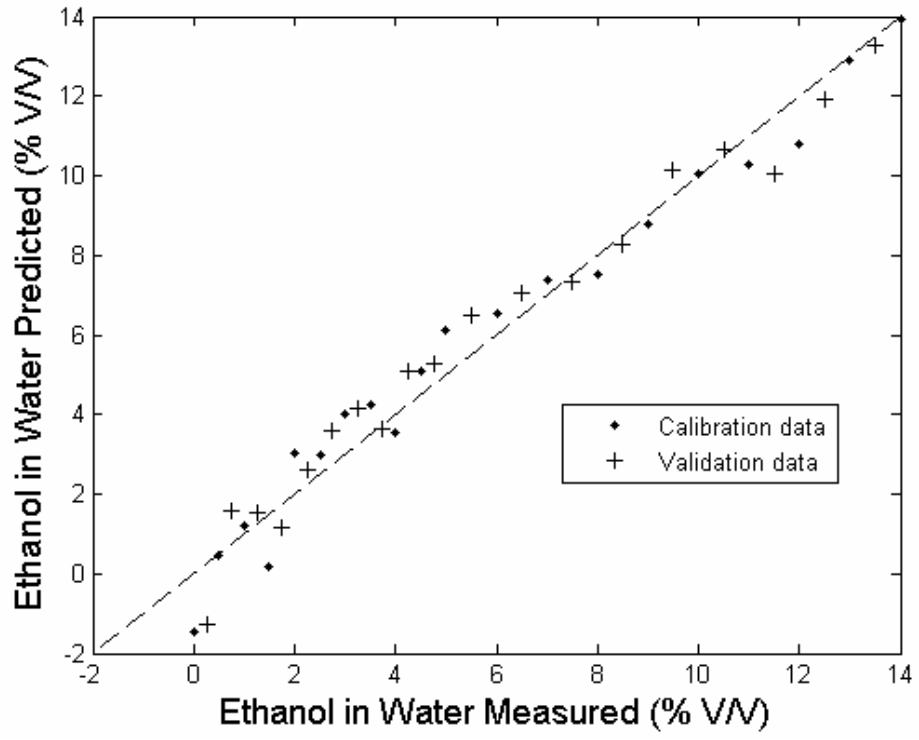


Figure 6



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Suggested Reviewers

Name: Dave Wetzel

Affiliation: Kansas State University, Dept. of Grain Science and Industry

Telephone: 785-532-6005

Email address: dwetzel@ksu.edu

Name: Gary Ritchie

Affiliation: US Pharmacopoeia

Email address: ger@usp.org

Name: Bill Fateley

Affiliation: Kansas State University, Chemistry Dept.

Telephone: 785-532-6298

Email address: bisnbil@ksu.edu

Name: Karl Norris

Affiliation: National Academy of Engineering

Telephone: 301 937 7547

Email address: knnirs@verizon.net