Anal. Chem. 1994, *66*, 43–51

Preprocessing of Analytical Profiles in the Presence of Homoscedastic or Heteroscedastic Noise

Olav M. Kvalheim,* Frode Brakstad, and Yi-zeng Liang†
Department of Chemistry, University of Bergen, N-5007 Bergen, Norway

Analytical profiles are commonly normalized to the most intense peak or to constant sum prior to library searches or multivariate analysis. This work examines normalization procedures from a theoretical point of view and their effects on simulated and real data. It is found that while normalization of data with homoscedastic noise to constant sum only introduces the well-known bias toward negative peak correlations, normalization in the presence of heteroscedastic noise has a profound impact on the correlation structure. It follows that heteroscedastic noise in analytical signals should be transformed to homoscedasticity *prior* to the use of any normalization procedure. For cases where the standard deviation of the noise is proportional to the peak intensity, the log-transform followed by row centering is shown to be an effective cure against the spurious correlations induced by constant-sum normalization. For the case of heteroscedastic noise where the relative standard deviation of the noise is decreasing with increasing peak intensity, a modified Box–Cox power transform is developed for removing heteroscedasticity in data prior to normalization to constant sum. The power transform has several attractive features: (i) it can be tailored to cure different patterns of heteroscedastic noise, (ii) it enhances the information in small peaks compared to large ones, and, (iii) it conserves strong linear correlations between signals.

A fundamental objective in many analytical investigations is to reveal relationships between samples and/or variables. A common example is comparison of the spectral profile of a sample with a library of reference spectra.[1] Other examples can be found in the wealth of literature concerning use of various pattern recognition techniques for classification of spectra.[2] The basic rule used when looking for relationships between analytical profiles is that similar profiles imply similar samples. This demands that "size" variation among profiles should be minimized prior to data evaluation. This has led to the common habit of first normalizing the sample profiles to constant sum using all[3] or a subset[4] of the variables. The residual variation in the replicated samples may subsequently be used as a scale or "measure of significance" for the observed relationships.[5]

A basic assumption for reliable interpretation of the correlation patterns in multivariate data is that pretreatment does not disturb the original relationships between the variables. For data with a uniform noise distribution, meaning that the noise is distributed independently of signal intensity, this represents a valid assumption in most practical cases. Thus, a Monte Carlo study by Skala[6] showed no significant distortion of the correlation structure for samples with at least seven variables.

Frequently, analytical profiles exhibit what is known as heteroscedastic noise,[7] meaning that the absolute noise increases with increasing intensity. Heteroscedastic noise may have a profound impact on the interpretation of normalized analytical data. Thus, normalization to constant sum effectively converts the noise from peaks with high intensity into systematic variation. Liang et al.[8] and Keller et al.[9] showed that heteroscedastic noise in data acquired on a liquid chromatograph with diode array detection (LC-DAD) gave rise to spurious components when the data were decomposed by principal component analysis.[10] Heteroscedastic noise is a matter of concern also for other kinds of analytical data. Thus, Toft and Kvalheim[11] showed that while the noise is approximately homoscedastic for transmittance FT-IR, the log transform to obtain absorbances induces heteroscedastic noise into the data. The lesson to learn from these two examples is that exploration of the noise pattern is mandatory before choosing normalization procedure.

Selective normalization,[4] where a few medium-sized peaks are used as internal standards, has been prescribed as a remedy against the entanglement caused by normalization. Unfortunately, it turns out to be impossible to retain the "true" correlation structure in normalized profiles based on signals showing a heteroscedastic noise structure. This observation points to the conclusion that heteroscedastic noise should be transformed to homoscedasticity *prior* to the use of any normalization procedure.

The optimal transformation to change heteroscedastic noise into homoscedastic depends on the structure of the heteroscedasticity in the signals. Results from regression and calibration theory suggest that the log transform is optimal if the standard deviation is proportional to the mean of the signal.[7] If the standard deviation of the noise is proportional to the root of the mean, the square root transform provides a homoscedastic noise pattern.[7]

† On leave from the Department of Chemistry and Chemical Engineering, Hunan University, Changsha, PRC.

(1) Zupan, J., Ed., *Computer-Supported Spectroscopic Databases*; Ellis Horwood: Chichester, UK, 1986.
(2) Varmuza K. *Pattern Recognition in Chemistry*; *Lecture Notes in Chemistry*; Springer: Berlin, 1980; Vol. 21.
(3) Reyment, R. A. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 79–91.
(4) Johansson, E.; Wold, S.; Sjödin, K. *Anal. Chem.* **1984**, *56*, 1685–1688.
(5) Kvalheim, O. M.; Aksnes, D. W.; Brekke, T.; Eide, M. O.; Sletten, E.; Telnaes, N. *Anal. Chem.* **1985**, *57*, 2858–2864.
(6) Skala W. *Math. Geol.* **1977**, *9*, 519–528.
(7) Sokal, R. R.; Rohlf, F. J. *Biometry—The Principles and Practice of Statistics in Biological Research*, 2nd ed.; Freeman and Co.: NewYork, 1981; pp 417–427.
(8) Liang, Y.-z.; Kvalheim, O. M.; Keller, H. R.; Massart, D. L.; Kiechle, P.; Erni, F. *Anal. Chem.* **1992**, *64*, 946–953.
(9) Keller, H. R.; Massart, D. L.; Liang, Y.-z.; Kvalheim, O. M. *Anal. Chim. Acta* **1992**, *263*, 29–36.
(10) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
(11) Toft, J.; Kvalheim, O. M. *Chemom. Intell. Lab. Syst.* **1993**, *2*, 79–91.

0003-2700/94/0366-0043$04.50/0

As will be discussed further on in the Theory section, normalization to constant sum after the log transform has a drawback: It induces artificial differences between profiles. This is easily seen when profiles from replicated samples are compared. Except for noise, replicate profiles differ only by a multiplicative factor. The log transform changes this multiplicative relationship into an additive one, so that subsequent normalization to constant sum induces artificial differences between profiles. Dividing each element of an analytical profile by the geometric mean before applying the log transform represents a possible way of circumventing this problem. The square root transform conserves multiplicative relationships between samples and variables, suggesting that the power transform may be a good candidate for curing different degrees of heteroscedastic noise prior to normalization to constant sum. The Box–Cox power transform,[12] which was developed for regression-type problems, should represent a useful starting point for finding such a preprocessing routine.

The aim of this work is to investigate the effect of normalization procedures on analytical profiles with different noise pattern and to tailor preprocessing procedures to the different situations. The task is attacked by the combined effort of a theoretical investigation and analysis of simulated and real data. In order to be able to look at possible influencing factors simultaneously, simulations were carried out by means of two-level factorial designs.[13]

THEORY

The Introduction has summarized some of the reasons, methods, and pitfalls for normalization of analytical profiles prior to correlation analysis. In essence, the discussion has pointed to the significance of revealing the noise pattern of the analytical method prior to selecting the preprocessing procedure. The outline of the Theory section is as follows: First, aims of normalization procedures and measures of performance are discussed. Subsequently, the concept of normalization is generalized by introducing what is called weighted normalization, and effects of the noise distribution on the results of normalization are discussed. Finally, possible candidates for pretreatment in the presence of homoscedastic and heteroscedastic noise are examined.

Normalization, Information Content, and Interpretation. The total variance S^2, for all samples and all variables, in a sample set can be partitioned into structure and noise:

$$S^2 = S^2_{\text{structure}} + S^2_{\text{noise}} \tag{1}$$

The ratio between structural and noise variance decides how much information can be extracted from a data set. If the noise variance is large compared to the structural variance, little information can be gained. On the other side, if the ratio between structural and noise variance is large, the potential information content in the data is correspondingly large. Procedures for pretreatment should enhance the information content in the analytical data, for instance, by minimizing effects of noise in signals and differences between samples due to, for example, differences in analyzed amounts. Normalization is performed in order to eliminate systematic "size" differences between samples. A successful normalization should have the following characteristics: (i) The difference between replicated samples should be minimal since this will act as lower limit for significant information, and (ii) the correlation between the original analytical signals should be retained as well as possible in the normalized data. The latter requirement can never be fully attained, simply because of the closure relation

$$\sum_k z_{ik} = 1 \tag{2}$$

leading to the well-known bias in the direction of negative correlations between peaks.[3,14] This is easily seen for the correlation between only two signals normalized to constant sum where the correlation coefficient is always −1.

Weighted Normalization. Normalization is defined as the transformation

$$\mathbf{z}_i^{\mathrm{T}} = 1/N_i \mathbf{x}_i^{\mathrm{T}} \tag{3}$$

where $\mathbf{z}_i^{\mathrm{T}}$ and $\mathbf{x}_i^{\mathrm{T}}$ represent the normalized and raw intensity vectors of sample i (superscript T implies transposition of a column vector into a row vector). The scalar N_i is determined from

$$N_i = \mathbf{x}_i^{\mathrm{T}} \mathbf{w} \tag{4}$$

The vector $\mathbf{w}$ is a column vector of weight factors, and thus eq 2 defines weighted normalization. Although, in principle, other choices are possible, the elements of this vector are here assumed to take only the values 0 or 1. At least one element of $\mathbf{w}$ must differ from zero.

The introduction of the weight vector $\mathbf{w}$ provides us with the opportunity to define all the common procedures for normalization by the same equation. Normalization to constant sum is defined by $\mathbf{w} = 1$. Normalization to internal standards or selective normalization[4] is performed by choosing the elements of $\mathbf{w}$ corresponding to the selected peaks as 1 and 0 for the others.

Normalization and Noise. In order to be able to assess the effect of the noise pattern on the normalization procedure, the intensity of the signal k of a sample i is expressed in terms of the "true" signal $\tilde{x}_{ik}$ plus a noise term:

$$x_{ik} = \tilde{x}_{ik} + e_{ik}; \quad \forall k \tag{5}$$

For normalization to constant sum, the normalization constant N_i is obtained as

$$N_i = \sum_k x_{ik} = \sum_k \tilde{x}_{ik} + \sum_k e_{ik} \tag{6}$$

If the noise is homoscedastic, the expectation value of the error term, $\langle \Sigma_k e_{ik} \rangle$, is zero and normalization to constant sum only introduces the trivial weak negative correlations discussed above. If, however, the noise is increasing with increasing signal intensity, the total noise term in eq 6 is dominated by the noise in the largest signals and this induces additional spurious correlations. This can easily be demonstrated for replicated samples where, by definition, the total variance is equal to the noise variance (see eq 1). If we assume that the replicates are dominated by a signal m and, furthermore, that its noise variance is much larger than the noise variance of all the other signals, the expectation value

(12) Box, G. E. P.; Cox, D. R. *J. R. Stat. Soc., Ser. B* **1964**, *26*, 211–243.
(13) Box, G. E. P.; Hunter, W. G; Hunter, J. S. *Statistics for Experimenters*; Wiley & Sons: New York, 1978; pp 505–508.

of the noise term in eq 6 is approximately equal to e_{im} so that

$$N_i = \sum_k x_{ik} = \tilde{x}_{im} + \sum_{k \neq m} \tilde{x}_{ik} + e_{im} \quad (7)$$

Inserting eq 7 in eq 3 shows that when e_{im} is large and positive, the intensities of the small signals are systematically suppressed after normalization to constant sum. Oppositely, when e_{im} is large and negative, the intensities of the small signals are systematically enlarged. Thus, in the presence of heteroscedastic noise, normalization to constant sum effectively converts noise from the largest signal into systematic, but false positive correlations between the smaller signals. When the small signals are suppressed, the large signal are enlarged and vice versa, leading to false negative correlations between the major variable and the others.

One might think that selective normalization[4] including only the signals of smallest size and variance should be a simple solution to the problem of heteroscedastic noise. However, selective normalization may be difficult to apply in practice. There are two main reasons for this. The variance of every signal is usually composed of contributions from structural differences as well as noise (eq 1). This often makes it difficult and sometimes impossible to find a suitable set of signals in the low-intensity or midintensity range with approximately the same noise variance. Another less serious drawback of the procedure is the subjective element connected with the selection of the appropriate signals. A theoretical objection that can be raised against selective normalization is that the samples are actually not normalized to the same size since only the subset of the variables included in the normalization sums to constant sum.

According to the considerations following eqs 6 and 7, the log ratio transformation of Aitchison,[14] which have found widespread use in, for example, geology and geochemistry, may be vulnerable to spurious correlations in the presence of heteroscedastic noise. Aitchison assumes that normalization to constant sum is always performed as the first operation. The normalized data are subsequently log transformed and row centered, i.e., subtracting the mean of the log transformed profile:

$$\mathbf{y}_i^T = \ln \mathbf{z}_i^T - 1/M \sum_k \ln z_{ik} \quad (8)$$

In eq 8, M is the total number of variables in the analytical profile. Log transform and row centering cannot correct for the spurious correlations introduced by constant-sum normalization in the presence of a heteroscedastic noise structure, and thus, the correlation structure may be severely distorted. Furthermore, since eq 8 shows that the centered log ratio procedure actually represents a normalization of every peak to the geometric mean of all the peaks, the question arises whether a prior constant-sum normalization is indeed needed. Both size differences between profiles and heteroscedasticity are taken care of by applying the transformation defined by eq 8 directly on raw profiles.

In addition to the detoriative influence of constant-sum normalization on the interpretation of the variable correlations in the presence of heteroscedastic noise, the redistribution of noise from the large signals into the smaller ones produces a larger spread between replicates. If the variables are standardized to equal variance prior to the correlation analysis, this effect is further enlarged.[15] This reduces the possibility of resolving small structural differences in the data.

According to the considerations above, normalization to constant sum or the major signal is the best procedure for chemical measurement techniques showing a homoscedastic noise behavior since the expectation value of the error term is 0. In the presence of heteroscedastic noise, other procedures have to be found.

Correcting for Heteroscedasticity prior to Normalization. Since heteroscedastic noise has such a destructive influence on normalization to constant sum, it is proposed that the data should be transformed to homoscedasticity prior to normalization. The choice of which transformation to use depends upon the degree of heteroscedastic noise in the data and whether signals are strongly correlated or not due to characteristics of the measurement technique. If the standard deviation of the noise, $s(x)$, increases linearly with the mean signal size $\bar{x}$,

$$s(x)/\bar{x} = \text{constant} \quad (9)$$

a homoscedastic noise pattern can be obtained by means of the logarithmic transformation.[7] Signals showing a linear relationship between the standard deviation of the noise and the square root of the size of the signal

$$s(x)/\bar{x}^{1/2} = \text{constant} \quad (10)$$

obtain a homoscedastic noise pattern through the square root transformation.

Equation 9 describes a situation with strong heteroscedasticity in the noise structure, i.e., a situation where the relative standard deviation of the noise is the same for the smallest and largest signals. On the other hand, eq 10 describes a situation where the relative standard deviation of the noise is reduced by a factor of 10 when the signal size increases with a factor of 100, i.e., a weak heteroscedastic behavior. Box and Cox[13] have deviced a transformation that can cover these and all intermediate cases of heteroscedastic noise:

$$y = (x^\lambda - 1)/\lambda; \quad \lambda \neq 0 \quad (11a)$$

$$y = \ln x; \quad \lambda = 0 \quad (11b)$$

The Box–Cox power transform was derived for the analysis of variance and so it needs some modification to be useful in connection with normalization.

If we disregard for the moment the case of $\lambda = 0$ and just look at the cases where $0 < \lambda < 1$, it is clear that the denominator in the transformation is superfluous when the transformation is accompanied by a nomalization procedure. Furthermore, as mentioned in the Introduction and shown in the next paragraph, the power transform conserves the multiplicative relationships between profiles and variables and so the subtraction of 1 in the numerator of eq 11a induces changes between profiles. Therefore, we propose to use simply the transformation x^λ to cover cases where the absolute noise is increasing but the relative noise is decreasing with signal

(14) Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman & Hall: London, 1986.

(15) Kvalheim, O. M. *Anal. Chim. Acta* **1985**, *177*, 71–79.

Table 1. Two-Level Factorial Design Used for Simulation Studies

condition	peak ratio[a]		noise pattern		% noise	
1	10	(−)	homoscedastic	(−)	2	(−)
2	1000	(+)	homoscedastic	(−)	2	(−)
3	10	(−)	heteroscedastic	(+)	2	(−)
4	1000	(+)	heteroscedastic	(+)	2	(−)
5	10	(−)	homoscedastic	(−)	5	(+)
6	1000	(+)	homoscedastic	(−)	5	(+)
7	10	(−)	heteroscedastic	(+)	5	(+)
8	1000	(+)	heteroscedastic	(+)	5	(+)

[a] Ratio between major and minor peaks.

size. In fact, by induction it can be shown that if the noise increases with $(1 - \lambda)$ power of the signal mean, $0 < \lambda < 1$, the λ power transform provides homoscedastic noise. For instance, the $^1/_4$ power transform cures cases where the relative standard deviation of the noise increases with the $^3/_4$ power of the signal mean. This corresponds to a case where the relative standard deviation of the noise is reduced by a factor of ~3 with an increase in the signal size by a factor of 100. Accordingly, the simplified power transform can be tailored to fit the noise structure of the analytical method.

Correlation and Transformations. The log transform, which is so excellent for taking care of strong heteroscedastic noise, has some undesired influence on linear correlations between samples and variables. This can easily be shown on profiles from replicated samples. Disregarding noise, replicated profiles differ only by a multiplicative factor a so that replicate profiles obey relations like $\mathbf{x}_i^T = a\mathbf{x}_j^T$. Constant-sum normalization after log transform destroys this relationship since $\log a$ appears as an additive constant in every signal and thus differences are induced between profiles. A straightforward way to taking care of this problem is to use row centering to remove the size factor $\log a$. Note, however, that in the presence of noise small size differences exist between replicates even after row centering. The power transform for $0 < \lambda < 1$ conserves the multiplicative relationship and so constant-sum normalization can be used after the power transform to take care of all cases of heteroscedastic noise where the relative standard deviation is decreasing with signal size.

Spectroscopic techniques usually provide several correlated signals for each analyte. Under quantitative conditions, each resolved signal from a particular analyte is, disregarding noise, perfectly correlated, i.e., meaning that $x_k = ax_m$, for the two variables k and m. After the power transform with $0 < \lambda < 1$, the signals for variables k and m are still perfectly correlated, while the log transform partially destroys the linear correlation. However, correlations that are not perfect are weakened by the power transform.

Multicomponent Analysis and Transformations. Deconvolution of analytical profiles acquired for multicomponent samples represents an important area for correlation analysis. A typical example is provided by peak purity assessment by correlation analysis of a two-way profile (retention time and spectral directions) obtained from multidetection chromatography.[8,9] There are usually two different goals of such an analysis: (i) to determine the number of analytes under a profile and (ii) to resolve the multicomponent profile into spectra and chromatograms of the individual analytes. Correlation analysis plays a crucial role in the first step. Thus, the task is commonly solved by principal component analysis of the whole profile or, better, smaller regions.[8,9,11] The number of analytes is estimated as the number of principal components with eigenvalues significantly larger than the noise level.

As shown in previous work,[9] heteroscedastic noise may have a detoriative effect on the result from this kind of correlation analysis. Due to the undesired effect of the log and power transforms on linear correlations, we appear to be left with no useful alternative. However, both the log and power transforms provide correct estimation of number of analytes in regions where only *one* analyte is eluting. A one-component region $\mathbf{X}_1$ can be described as a product of a chromatographic profile $\mathbf{c}_1$ and a spectral profile $\mathbf{s}_1$, i.e., $\mathbf{X}_1 = \mathbf{c}_1\mathbf{s}_1^T$. It follows directly that the power transformed data obey the relation $\mathbf{X}_1^\lambda = \mathbf{c}_1^\lambda(\mathbf{s}_1^T)^\lambda$. The log transform provides a sum of contributions from the chromatographic and spectral profiles, i.e., $\log \mathbf{X}_1 = \log \mathbf{c}_1\mathbf{1}^T + \mathbf{1} \log \mathbf{s}_1^T$. It appears that principal component analysis will imply two analytes in the one-component regions. However, row centering after log transform reduces the rank by one, and therefore, a one-component region is described by one principal component also after the log row centring transform. Both power and log transforms should thus have some potential for detecting selective regions under a multicomponent profile.

For the general case of mixture analysis, with several coeluting analytes, the prospects are not so good, however. It may be necessary to compromise between heteroscedastic noise and preservation of linear correlations. Normalization of spectra with total absorbance above a chosen threshold to constant total absorbance reduces the heteroscedasticity in the data and may provide a reliable assessment of the number of analytes under an unresolved profile.[8,9] Subset selection[4] represents another alternative if a good criterion can be found for the selection procedure.

Normalization Procedures Adapted to Noise Structure. Our theoretical analysis has shown that the following normalization procedures should be appropriate:

homoscedastic noise,

normalization to constant sum (eqs 3 and 4)

$$\mathbf{z}_i^T = \mathbf{x}_i^T / \sum_k x_{ik} \tag{12a}$$

heteroscedastic noise,

relative standard deviation constant (eq 9)

$$\mathbf{z}_i^T = \ln \mathbf{x}_i^T - 1/M \sum_k \ln x_{ik} \tag{12b}$$

heteroscedastic noise,

relative standard deviation decreasing with signal size (e.g., eq 10)

$$\mathbf{z}_i^T = \mathbf{x}^\lambda / \sum_k x_{ik}^\lambda \tag{12c}$$

Noise and Structural Variance after the Power Transform. Equation 5 can be rewritten and power transformed to obtain

$$x_{ik}^\lambda = \bar{x}_{ik}^\lambda (1 + e_{ik}/\bar{x}_{ik})^\lambda \tag{13}$$

Assuming that $e_{ik}/\tilde{x}_{ik} << 1$, eq 13 approximates to

$$x_{ik}^{\lambda} \cong \tilde{x}_{ik}^{\lambda}[1 + \lambda(e_{ik}/\tilde{x}_{ik})] \quad (14)$$

The expectation value of the term $e_{ik}/\tilde{x}_{ik}$ is equal to the relative error of the untreated signal. For the transformed signal this error term is reduced by the factor λ (see eq 14), showing that the power transform reduces the noise variance by a factor of λ^2. On the other hand, the transformation also reduces the structural variance (as defined by eq 1) in a signal, and this has to be counterbalanced against the gain in noise variance. Similar reduction of noise variance accompanies the log transform.

EXPERIMENTAL SECTION

Simulated Data. Preliminary studies of data suggested that three factors might be important with respect to choosing the most appropriate preprocessing procedure. The three factors were the ratio between the intensities of major and minor signals, the noise structure (homoscedastic vs heteroscedastic), and the standard deviation of the noise. Thus, we decided to simulate data sets with two major and two minor peaks for 10 samples. The data sets were generated by full two-level factorial design. With three factors varying between two levels, this provides eight different conditions for each investigated transformation. The ratio between the major and minor signals was chosen as either 10 (–) and 1000 (+) with the size of the minor signals always being equal to 10 except for added noise. The noise pattern was varied as either homoscedastic (–) or heteroscedastic (+). Homoscedastic noise was added as 2 or 5% relative to the smallest peaks. Similarly, heteroscedastic noise was added as either 2 or 5% relative to the individual peaks. The design is given in Table 1. The design was executed for two different procedures for pretreatment: (i) normalization to constant sum (eq 12a) and (ii) log transform prior to row centering (log row centring, eq 12b). The eight conditions for each transformation were repeated three times and the correlation matrices calculated. This gives a total of 48 data sets (eight conditions repeated three times for two different transformations), each one simulating 10 replicated samples. Prior to pretreatment, the peaks in a data set are expected to be almost uncorrelated since the 10 samples differ only by added noise.

Three responses were calculated from the correlation matrices: (i) the correlation coefficient between the two major peaks, (ii) the correlation coefficient between the two minor peaks, and (iii) the average of the four correlation coefficients between the major and minor peaks. Each of the three responses was reported as the average from the three replicated runs generated for each combination of factors defined by the design. In principle, we should use the differences between the correlation coefficents before and after transformation, but as the expectation value is zero for all the correlation coefficients before pretreatment, we simply used the results after pretreatment of the simulated data. Also, in order to check the power transform combined with normalization to constant sum, a further simulation was carried out where noise proportional to the $^3/_4$ power of the signal mean was added to an analytical profile.

Replicated Mass Spectra. The noise structure in mass spectral data was investigated by injecting eight replicates of 0.1% (w/v) 1,2,3,4-tetrachlorinated benzene (98% GC, Aldrich, Catalog No. 13, 187–189, M_w 215.89) to an ordinary quadrupole GC/MS system, equipped with a capillary column. The instrumental operating conditions are the same as described in ref 16.

Programs Used for Data Generation and Evaluation. Mathlab (The Math Works Inc., 1991) was used for generating simulated data sets with added noise. SIRIUS for Windows (Pattern Recognition Systems Ltd., 1993) was used for evaluation of data.

Table 2. Correlation Distortion of Transformed Data

		correlation distortion[a]					
		constant sum			log row centering		
condition	factor levels	r(l,l)	r(l,s)	r(s,s)	r(l,l)	r(l,s)	r(s,s)
1	– – –	–0.22	–0.43	–0.04			
2	+ – –	–0.33	–0.43	–0.27			
3	– + –	–0.99	–0.04	–0.08	–0.48	–0.31	–0.43
4	+ + –	–1.00	0	0.34	–0.19	–0.44	–0.20
5	– – +	–0.18	–0.48	0.17			
6	+ – +	–0.14	–0.46	0.06			
7	– + +	–0.98	–0.07	0.39	–0.25	–0.39	–0.11
8	+ + +	–1.00	0	0.47	–0.35	–0.32	–0.30

[a] r(l,l), r(l,s), and r(s,s) imply correlation coefficients between large peaks, between large and small peaks (average of four pairs), and between small peaks, respectively.

RESULTS AND DISCUSSION

This section starts with a visual examination of results in order to reveal the most obvious correlation distortions accompanying the examined preprocessing procedures. We then proceed with a more detailed evaluation of the results by computing regression models for predicting the correlation distortions from the varied factors and thus revealing significant factors and factor interactions. Finally, the implications from the observations of the factorial designs are investigated with respect to gas chromatography/mass spectrometry.

Correlation Distortion, Noise Pattern, and Normalization. The results of the factorial designs are displayed in Table 2 for the data sets using constant-sum normalization and log row centering at the eight conditions defined in Table 1. Note that, for the log row centering procedure, the results for homoscedastic noise have been excluded since we had shown in a previous paper[11] that the log transform induces heteroscedastic noise when applied to analytical profiles with a homoscedastic noise pattern.

Table 2 shows that constant-sum normalization induces strong negative correlations between major peaks in the presence of heteroscedastic noise, while only a weak negative correlation is observed for data showing a homoscedastic noise behavior. This is as expected from the analysis in the Theory section. An opposite picture is found for the correlations between major and minor peaks: homoscedastic noise shows a stronger tendency toward inducing negative correlations between major and minor peaks than heteroscedastic noise. However, the negative correlations are as expected from the closure relation for homoscedastic noise (see Theory section and ref 14). The zero correlation between major and minor peaks in case of heteroscedastic noise is more surprising, but

(16) Brakstad, F. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 87–100.

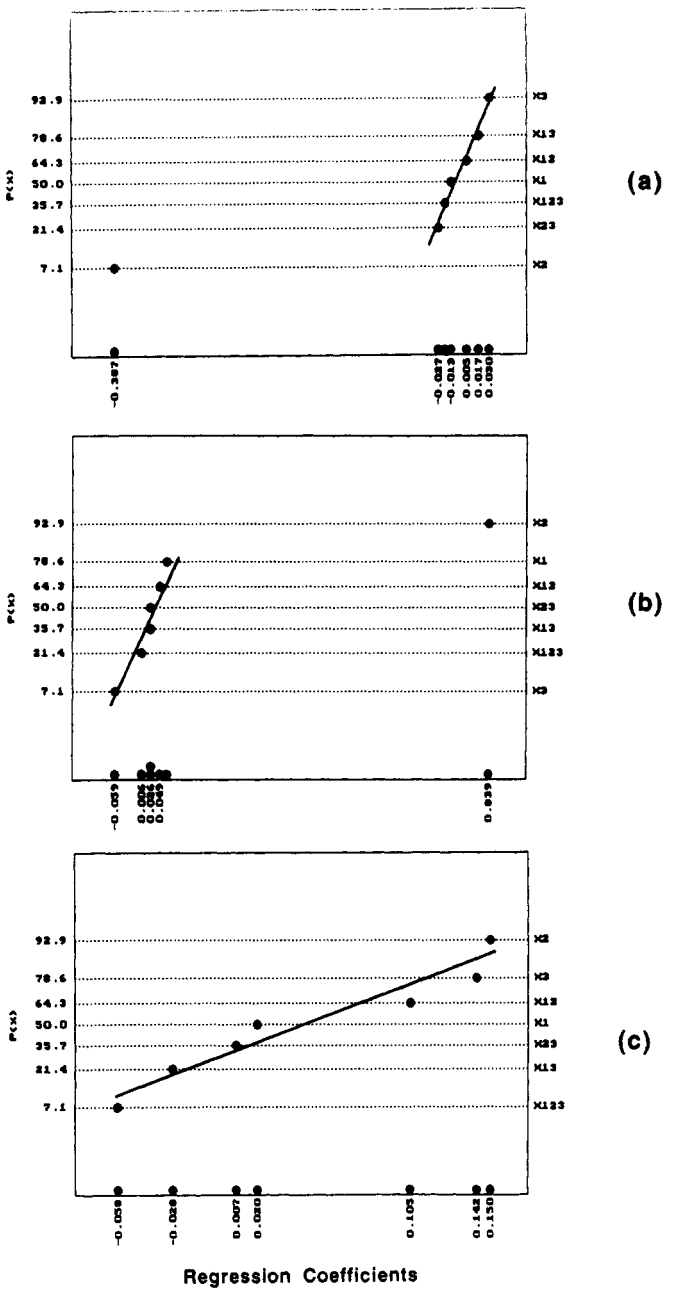

Figure 1. Normal probability plots for the calculated effects of the three factors peak ratios, noise structure, and noise level and their interactions on the correlations (a) r(l,l), (b) r(l,s), and (c) r(s,s), respectively, obtained after constant-sum normalization.

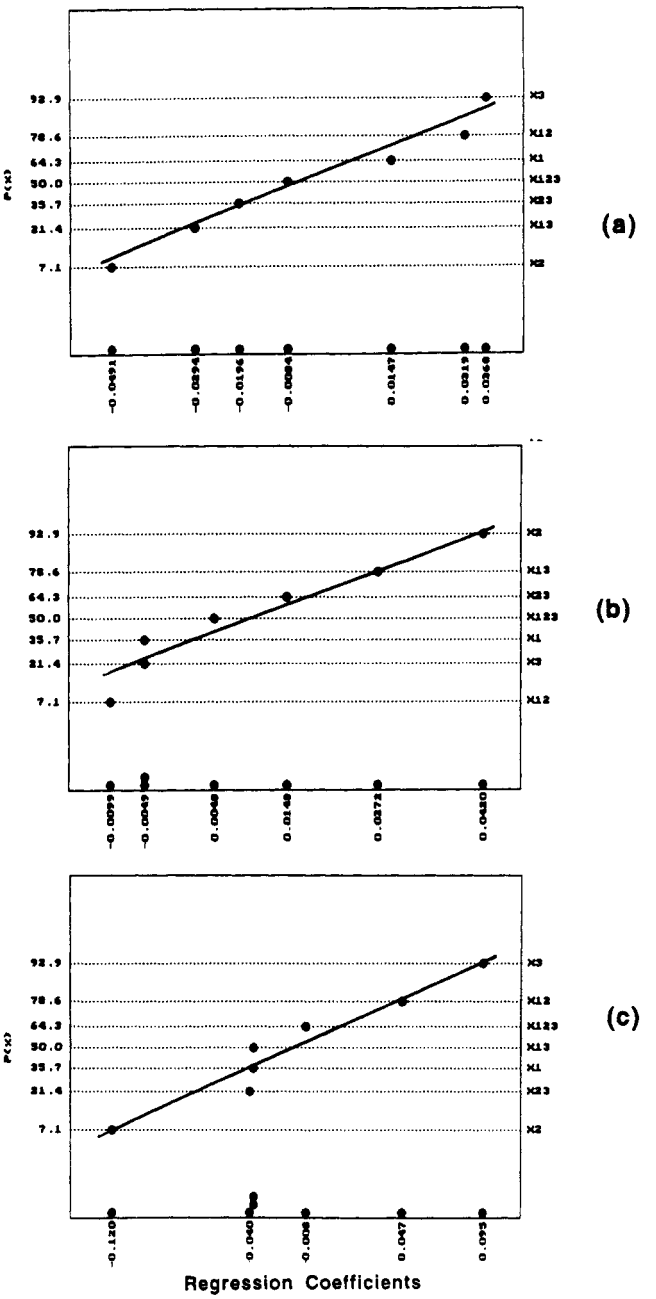

Figure 2. Normal probability plots of the calculated effects of the three factors (peak ratios, noise structure, noise level) and their interactions on the correlations (a) r(l,l), (b) r(l,s), and (c) r(s,s), respectively, obtained after constant-sum normalization for data with homoscedastic noise and log row centering for data with heteroscedastic noise.

easily explainable. The strong negative correlation between the major peaks implies that the correlations with one minor peak must be counterbalanced by a correlation of the same size but with opposite sign for the other minor peak. The overall sum is then approximately zero, a result that was confirmed by detailed examination of the correlation structure of these data. There is a tendency for weak positive correlations between minor peaks in the case of heteroscedastic noise as also reported by Skala.[17] Thus, constant-sum normalization seems to induce only the trivial negative bias due to closure for data with a homoscedastic noise structure, while the same normalization procedure has a much more profound effect on data with a heteroscedastic noise structure. Since normalization to constant sum is always performed as the first operation in Aitchison transformation,[14] his method will, inevitably, be influenced when applied to heteroscedastic data.

Log row centering of heteroscedastic data is seen to provide negative bias of the same size as found in the four runs with homoscedastic noise normalized to constant sum. Thus, this transform seems to perform well for data with heteroscedastic noise. Deliberately, we have not shown the results obtained for log row centering of data with homoscedastic noise. Log row centering induces strong positive correlations between the major peaks in the case of homoscedastic noise for the simulated data. This effect may seem surprising, but is easily explainable. The log transform reduces major peaks more

(17) Skala, W. *Chem. Geol.* **1979**, *27*, 1–9.

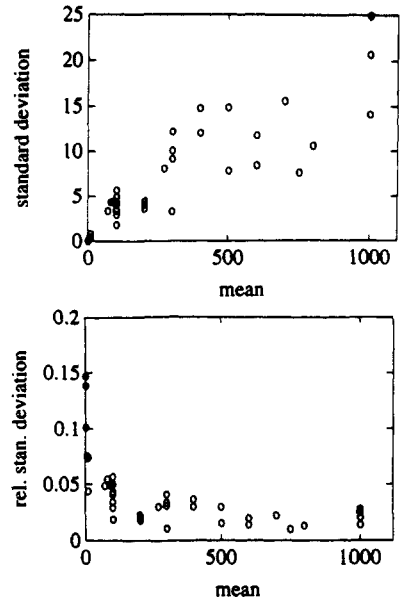

Figure 3. Standard deviation and relative standard deviation of the noise plotted vs the signal mean for a set of simulated replicated profiles.

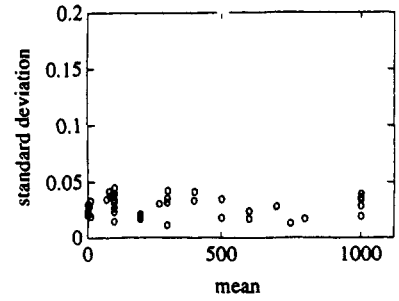

Figure 4. Standard deviation of the noise plotted vs the signal mean for the same simulated replicated profiles as in Figure 3, but after four-root transformation.

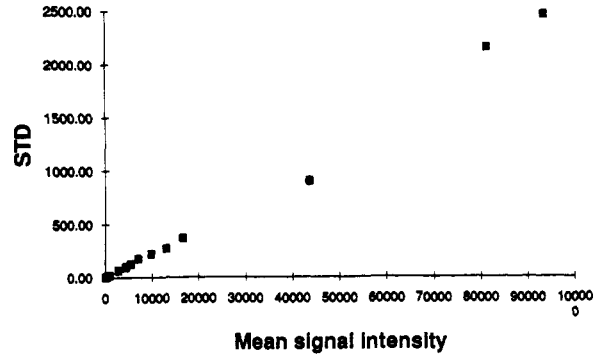

Figure 5. Standard deviation of the noise plotted versus the signal mean for randomly selected fragments from a set of replicated mass spectra.

than small peaks, thus inducing heteroscedastic noise on signals with a homoscedastic noise behavior.[11] After log transformation, the smaller signals have obtained a larger relative noise. Row centering of the transformed data redistributes noise from the small signals into systematic increase and decrease of the major signals, thus producing positive correlations between the major peaks.

Constant-Sum Normalization. The 2^3 factorial design generated for the constant-sum normalization was analyzed by calculating regression models using the correlation coefficients between the two major peaks, r(l,l), the average of the four correlation coefficients between the major and minor peaks, r(l,s), and the correlation coefficients between the two minor peaks, r(s,s), respectively, as responses. In addition to the three varied factors (Table 1), all interactions were included in the modeling.

Figure 1 shows the calculated effects of the three factors and their interactions on the correlations r(l,l), r(l,s), and, r(s,s), respectively, displayed in normal probability plots. Clearly, only the noise pattern (x_2) influences significantly the correlations between large peaks and between large and small peaks. Variation in peak ratios and noise level is found to have no influence on these correlation distortions. For the correlation between the small peaks, all factors and interactions fall close to a straight line so that the null hypothesis of a normal distribution, and thus no significant factors, cannot be rejected in this case.

Regression models for r(l,l) and r(l,s) including only the noise pattern (x_2) account for 98.1% of the variance in both r(l,l) and r(l,s). Furthermore, the predictions from these regression models show a correlation coefficient with the actual values in Table 2 of 0.992 for both r(l,l) and r(l,s). These results confirm beyond doubt the profound impact of the noise structure (x_2) upon the results of constant-sum normalization. The regression models can be written as

$$r(l,l) = -0.60 - 0.39x_2 \qquad (15a)$$

$$r(l,s) = -0.24 + 0.21x_2 \qquad (15b)$$

The best pretreatment is the one influencing the correlation structure least. Since the data were generated with only random correlation, this requirement is fulfilled when eqs 15a and 15b provide correlations close to zero. For the correlation between major peaks, homoscedastic noise ($x_2 = -1$) only induces the trivial negative bias expected from the closure relation (eq 1), while constant-sum normalization of the data with heteroscedastic noise ($x_2 = 1$) distorts this correlation toward perfect negative correlation.

Equation 15b suggests an opposite picture for the correlation bias between major and minor peaks. Thus, the bias is apparently smallest after constant-sum normalization of data with heteroscedastic noise. As discussed above, detailed inspection of the correlations contributing to r(l,s) shows that strong correlations are induced between the minor peaks and the major peaks, but with opposite sign and thus summing to zero. Thus, the results are consistent and constant-sum normalization is the optimal choice for data with homoscedastic noise.

Log Row Centering vs Constant-Sum Normalization. In order to confirm that log row centering could really replace constant-sum normalization in the case of heteroscedastic noise, a new 2^3 design was constructed by combining the four runs with homoscedastic noise and constant-sum normalization with the four runs with heteroscedastic noise and preprocessed by log-row centering. The calculated regression coefficients are plotted as normal probability plots in Figure 2 for the three responses r(l,l), r(l,s), and r(s,s) of the combined design. In all cases, the effects fall on a straight line, consistent with the null hypothesis of no significant factors or interactions. Thus, for the minimization of spurious negative correlations

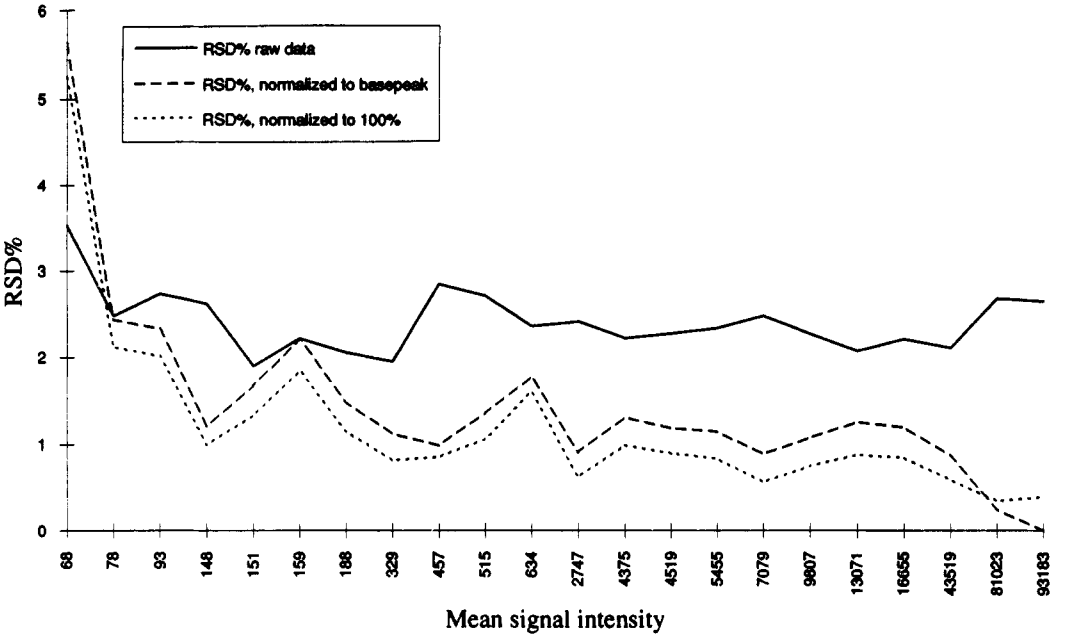

Figure 6. Relative standard deviation of the noise for different normalization procedures plotted versus the signal mean for the same fragments and spectra as shown in Figure 5.

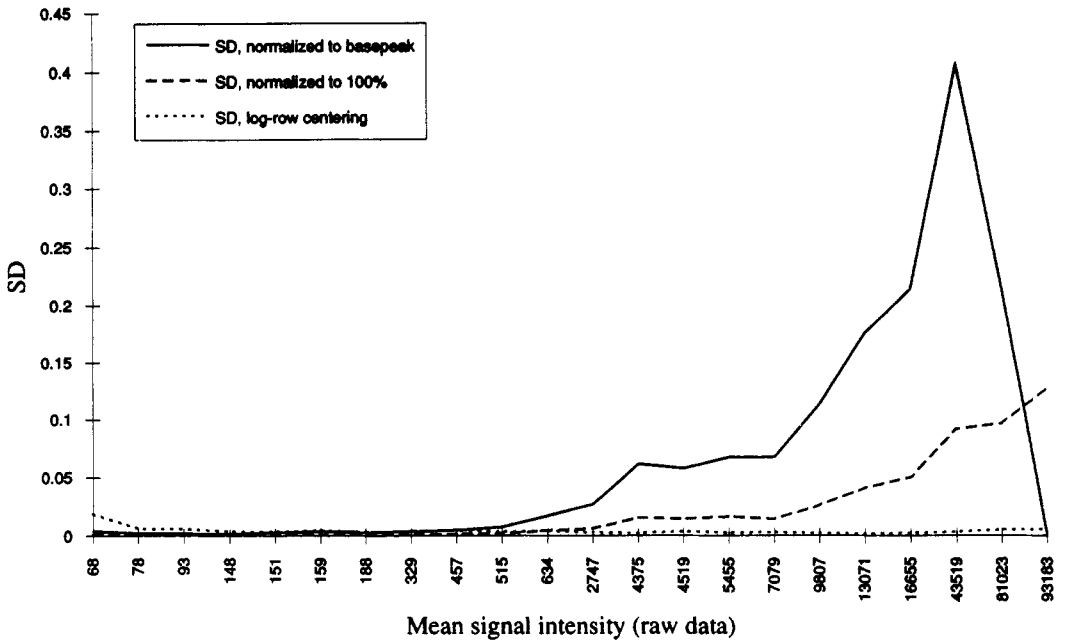

Figure 7. Standard deviation of the noise for different normalization procedures plotted versus the signal mean for the same fragments and spectra as shown in Figure 5.

between major and minor peaks, constant-sum normalization of data is the preferred choice with homoscedastic noise, while log row centering of data is the best choice for data showing a heteroscedastic behavior.

Power Transform and Normalization to Constant Sum. A task that remains to be done is to check the performance of the power transform for correcting cases with weaker heteroscedastic noise, i.e., cases where the relative standard deviation of the noise is decreasing with mean signal size. Figure 3 shows the standard deviation and the relative standard deviation of the noise plotted vs the signal mean for a set of simulated replicated profiles (see the Experimental Section). A heteroscedastic behavior of the noise is obvious, and the decreasing relative standard deviation shows that a power transform should be appropriate. Calculations showed that the standard deviation of the noise increases proportionally to the $^3/_4$ power of the signal mean, suggesting a fourth-root transformation as appropriate in order to achieve a homoscedastic noise structure. Figure 4 leaves no doubt that the standard deviation of the noise has become constant after this transformation. Thus, we were able to make the right diagnosis and to prescribe the right cure in this situation also.

Normalization of Mass Spectra. In order to check our procedures and recommendation on real data, eight replicates of tetrachlorinated benzene analyzed on GC/MS were studied (see the Experimental Section). Figure 5 shows the standard deviation of selected fragments from the replicated raw spectra plotted versus mean intensity. Clearly the noise pattern is heteroscedastic. Figure 6 shows the relative standard deviation of the raw spectra together with the relative standard deviation

of the spectra normalized to constant sum and base peak. As the relative standard deviation of the noise is proportional to the mean intensity, log row centering should be the best preprocessing procedure. The normalization to base peak and constant sum shows similar results: a slightly decreasing relative standard deviation implying that these procedures reduce the heteroscedasticity in the profiles. As shown in Figure 7, however, the noise is still heteroscedastic after these transformations, while log row centering transforms the noise to perfect homoscedasticity.

CONCLUSIONS

The significance of revealing the noise pattern of the response variable prior to linear calibration has long been recognized by analytical chemists.[18] Less well-known is the deteriorative influence caused by heteroscedastic noise on the results of normalization of analytical profiles to constant sum or to the largest signal prior to correlation analysis. As shown in this work, the correlation pattern is profoundly influenced by noise pattern when normalization to base peak or constant sum is performed. This knowledge may be utilized, for example, for extracting more meaningful information from analytical profiles and to increase the success-to-failure ratio when spectral comparisons in a library search are performed. The preprocessing methods advocated in this work should have a significant effect on improving the resolution power of principal component analysis[10] and other correlative techniques, meaning that these methods can be applied to solve problems where the structural variance in the data is approaching the noise level.

Furthermore, the findings of this work should not only be useful in exploratory analysis of analytical data but should indeed also be useful in confirmatory regression-type analysis such as calibration and structure–property modeling where analytical profiles frequently are used as predictors. This follows since regression analysis using signal intensities as predictors assumes that the noise in the intensities is small compared to the error in the response variable(s). This may not be so for large signals showing a heteroscedastic behavior, and thus pretreatment may be necessary to adapt to the assumptions of regression modeling. In addition, a chemical interpretation of the regression model is desirable in most cases, and this can only be achieved through the (preprocessed) analytical profiles.

ACKNOWLEDGMENT

The Royal Norwegian Council for Scientific and Industrial Research (NTNF) is thanked for a Ph.D. grant to F.B. The Norwegian Research Council for Science and Humanities (NAVF) is thanked for a visitor grant to Y.-z.L.

Received for review August 25, 1993. Accepted October 20, 1993.[•]

(18) Agterdenbos, J. *Anal. Chim. Acta* **1979**, *108*, 315–323.

• Abstract published in *Advance ACS Abstracts*, November 15, 1993.