ELSEVIER

Chemometrics and Intelligent Laboratory Systems 40 (1998) 101–107

Chemometrics and intelligent laboratory systems

The effect of heteroscedastic noise on the chemometric modelling of frequency domain data

Andrew M. Woodward *, Bjørn K. Alsberg [1], Douglas B. Kell [2]

Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DA, UK

Received 14 March 1997; accepted 15 October 1997

Abstract

The structure of noise in a dataset and, in particular, whether it is homoscedastic or heteroscedastic, can significantly affect the properties of multivariate calibration models. This is particularly true when the data are subjected to a nonlinear transformation prior to the formation of the model. The problems of mathematical modelling in the frequency domain in the presence of heteroscedastic noise are demonstrated using simple, illustrative, synthesised datasets and partial least squares regression. The heteroscedasticity spreads signal-dependent information throughout the spectrum of the signal, removing the localisation seen with band-limited signals with homoscedastic noise. Heteroscedasticity significantly reduces the scope for efficient variable selection to allow modelling on a reduced variable set, with consequences for the production of sparse models which generalise well according to the parsimony principle. However, significant modelling can take place purely on the noise components even when the frequency range of the signal has been completely excluded. Optimal denoising schemes will beneficially take into account the noise structure of a dataset.

Keywords: Heteroscedastic noise; Frequency transform; Mathematical modelling; Chemometrics

1. Introduction

Mathematical modelling is applied widely to frequency-domain information derived from time domain signals by means of a frequency transform. It is commonplace that much of modern spectroscopy is concerned with forming a relationship between a vector corresponding to a wavelength-dependent parameter of a sample and the concentration of a determinant of interest which that sample may contain, a procedure usually referred to as multivariate calibration [1–7]. However, the signal, its spectra and consequently the model, are inevitably corrupted by noise. This unwanted noise on the time signals is usually assumed to be homoscedastic, i.e. to be independent of the signal characteristics.

However, many instruments and processes, if not most, have some component of heteroscedastic noise (i.e. the characteristics of the noise depend on the characteristics of the signal for each datum) [8]. Alternatively, heteroscedastic noise can arise from homoscedastic noise by a preprocessing stage applied to the signal; for example this is the case for absorption

* Corresponding author. Tel.: +44-1970-623111 ext. 4188; fax: +44-1970-622350; e-mail azw@aber.ac.uk.

[1] E-mail: bka@aber.ac.uk.

[2] E-mail: dbk@aber.ac.uk.

0169-7439/98/$19.00
PII S0169-7439(97)00078-6

infrared spectra which are converted from transmission spectra using a non-linear transform [9–11]. The presence of heteroscedastic noise has fundamental implications for a modelling process performed on a frequency-transformed representation of the data.

2. Theoretical considerations

A frequency transformation such as the FFT [12] delocalises information across the whole frequency space, i.e. the frequency components of the noise are spread across high-frequency regions not typically covered by the frequency content of the signal. For homoscedastic noise, these noise components are uncorrelated with the measured reference data and so modelling will not take place on this region. Thus the modelling will be restricted to the region of frequency space containing the representation of the signal.

A signal, $h(t)$, containing only homoscedastic noise can be written as

$$h(t) = \alpha n(t) + s(t) \quad (1)$$

where α is a scalar that determines the size of the noise $n(t)$ and $s(t)$ is the pure signal. This Fourier transforms to

$$H(\omega) = \alpha N(\omega) + S(\omega) \quad (2)$$

such that the signal is independent of the noise and is concentrated only in the region defined by $S(\omega)$. This situation is illustrated in Fig. 1, where the transformed signal is seen to be concentrated into the first 20 or so frequency bins and the rest of the spectrum is predominantly structureless noise.

In the presence of heteroscedastic noise, this situation changes fundamentally. The structure of the noise is now dependent on the structure of the signal and since the signal (in a well designed experiment) contains information correlated with the reference data the structure of the noise will also be correlated with the reference data. This noise will be spread out across the frequency domain in a manner similar to that due to homoscedastic noise, but now this 'noise-only' region will no longer be uncorrelated with the reference and significant modelling can take place on the (correlation between signal and) noise alone.

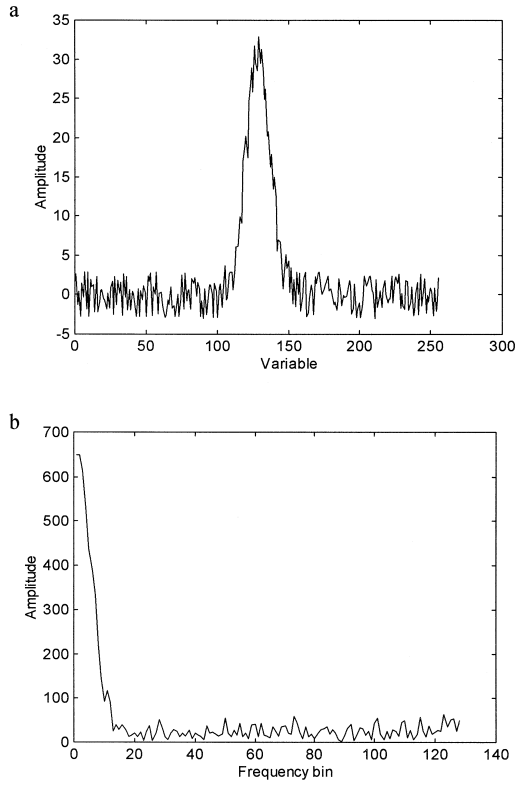

Fig. 1. (a) Gaussian function with added homoscedastic noise such that the amplitude of the noise (after removing its d.c. to produce double-sided noise) was 10% of the amplitude of the noise free Gaussian and (b) its amplitude spectrum.

In what way is the noise spread out across the frequency domain? In order to see this we recognize that the signal containing heteroscedastic noise $h(t)$ can be written as follows:

$$h(t) = \alpha f(s(t))\, n(t) + s(t) \quad (3)$$

where $f(s)$ is the functional dependence of the noise on this signal. A Fourier transform of the noisy signal $h(t)$ produces:

$$H(\omega) = \alpha F[s(t)] \otimes N(\omega) + S(\omega) \quad (4)$$

where $\otimes$ is the convolution operator and all the upper case letters signify the corresponding Fourier transform of the functions in the time domain (written in lower case letters). The spectrum of the heteroscedastic noise can thus be regarded as the convolution of the spectrum of the homoscedastic noise

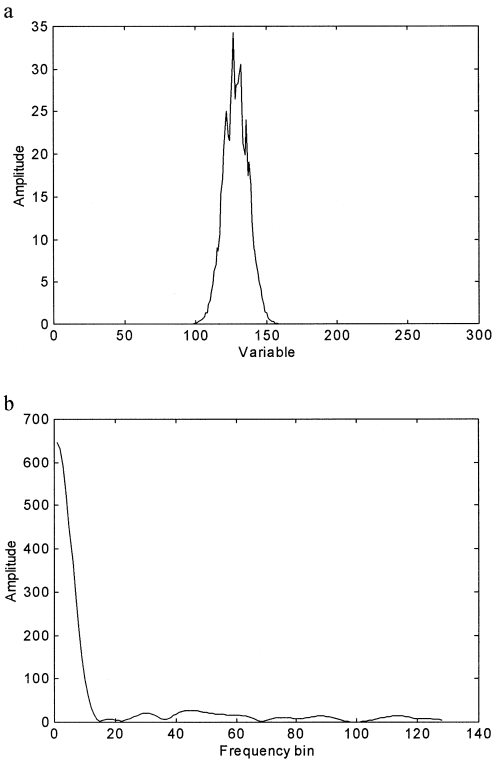

Fig. 2. (a) Gaussian function with added heteroscedastic noise. The noise of Fig. 1 was multiplied by the signal Gaussian to produce heteroscedastic noise and (b) its amplitude spectrum.

with that of the signal. Accordingly, all frequencies in the spectrum will contain information related directly to the signal, $s(t)$. This situation is illustrated in Fig. 2.

This spectrum is clearly different in form to that of Fig. 1b, showing more obvious structure in the noise region, which now contains information relating to the signal.

3. Software and methods

The modelling process chosen to illustrate the effects of heteroscedastic noise was partial least squares regression because it is well known, widely applicable to many areas of modelling and easily performed by many commercial and in-house software packages. The nomenclature used below is in accordance with multivariate modelling literature [7].

All simulations and PLS1 models were performed using Matlab 4.2 (The MathWorks, 24 Prime Park Way, Natick, MA), Models on synthesised data are created, with either heteroscedastic or homoscedastic noise added. The *x*-data (signal) was a set of 31 Gaussian curves with amplitudes varying linearly between 0 and 30 arbitrary units. Gaussians were chosen in order to illustrate the effect since they are simple functions with simple, known, standard Fourier transforms and can represent a single spectral peak.

The *y*-data (reference) was a simple linear vector of values 0, 1, ..., 30, reflecting the amplitudes of the Gaussians in the *x*-data. For all following modelling purposes this dataset was split into odd and even samples; where the odd samples were used as the calibration set and the even samples as the test set to verify the predictive ability of the PLS model, the odd number of total samples ensuring that the extreme spectra are in the training set to avoid any need for the model to extrapolate in predicting the test set. The precision of the predictions is shown as the root mean squared error of prediction (RMSEP), expressed as a percentage of the mean *y*-value of the prediction set.

4. Results

The noise-free data were Fourier transformed to produce their frequency spectra and PLS modelling was performed on the odd samples of these spectra, producing a prediction of the even samples.The RMSEP for this prediction was zero, showing perfect representation of the noise-free dataset. Similarly the prediction on the time domain data is also perfect, as expected.

Note that all predictions in this development showed optimum prediction for one PLS factor unless otherwise stated.

Homoscedastic noise was added to the dataset with Matlab's (single sided) random number generator such that the amplitude of the (flat-spectrum) noise, after removing its d.c. to produce double-sided noise, was 10% of the amplitude of the median Gaussian in the noise-free dataset. This amplitude is constant for all Gaussians in the dataset and independent of the ordinate of the Gaussian.

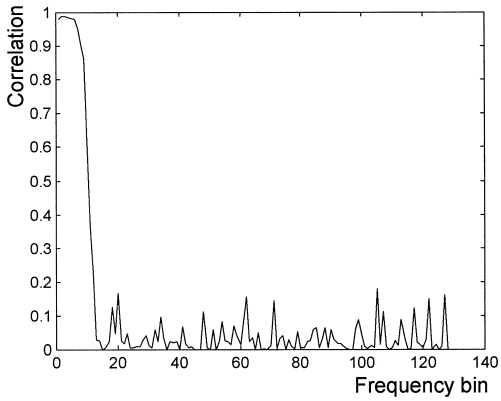

Fig. 3. Correlation of each frequency in the dataset including homoscedastic noise with the known reference *y*-data.

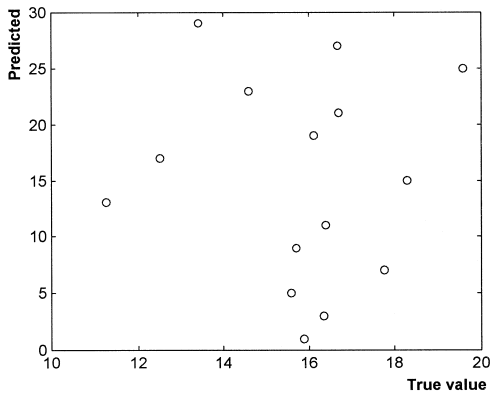

Fig. 4. Frequency domain prediction on homoscedastic noise variables only. The gradient and intercept of the regression line through this data are −0.1322 and 17.0874. respectively, and the correlation coefficient is −0.0318.

The RMSEP for the equivalent PLS prediction on this data is 3.3%. The prediction on the time domain data gives an RMSEP of 3.63%.

Performing a correlation of each individual frequency variable in the amplitude spectra with the *y*-data gives the function shown in Fig. 3. This shows that the signal region up to frequency bin 15 is highly correlated with the reference, whereas the noise region above this is essentially uncorrelated except for a few variables which correlate by chance. A few from any set of random variables will always correlate to some degree with any chosen reference. The correlation provides essentially similar information to the first-factor loadings plot in the PLS model.

Performing a PLS prediction on only the noise variables (variable 20 upwards) shows that there is negligible modelling ability in this region as shown by the RMSEP figure of 56.9% and depicted in Fig. 4.

The noise matrix created above was then multiplied by the corresponding individual signal Gaussian to produce linear heteroscedastic noise. The amplitude of this noise is consequently linear with the ordinate of the *x*-data. Linearity was chosen as being a simple form of heteroscedasticity with which to illustrate the effect. However the effect is general to any relationship between noise and signal as theoretically predicted above and as will be seen later. This noise was then added to its respective signal Gaussian to form a dataset with heteroscedastic noise, i.e. the noise is identical for the homo- and heteroscedastic cases except for the multiplying function $f(s(t))$ which is, in this case, linear.

PLS predictions of the full dataset give an RMSEP of 2.1%. with two factors optimal. This is slightly better than the homoscedastic prediction but requires more factors reflecting the now more complicated data structure. The prediction on the time domain data gives an RMSEP of 3.08% with two factors optimal. However, the correlation function equivalent to that of Fig. 3 shows much more correlation at the noise frequencies as indicated in Fig. 5 and as might be expected from this, the PLS prediction on noise alone, using the same variable range as

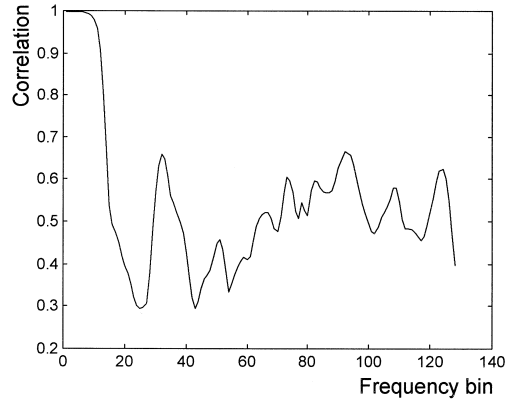

Fig. 5. Correlation of each frequency in the dataset including heteroscedastic noise with the known reference *y*-data.

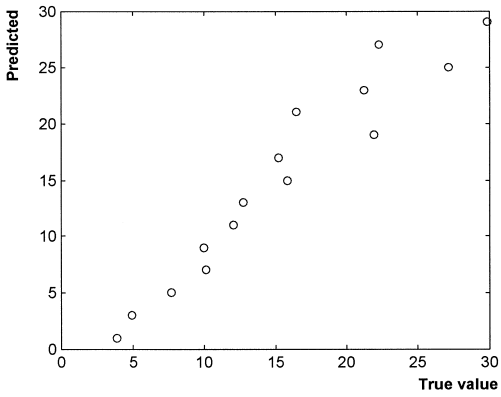

Fig. 6. Frequency domain prediction on heteroscedastic noise variables only. The gradient and intercept of the regression line through this data are 1.1047 and −2.0570, respectively, and the correlation coefficient is 0.9624.

for the homoscedastic case, is quite respectable with an RMSEP of 16.3%, with two factors optimal, as shown in Fig. 6.

To eliminate the possibility of this prediction occurring on purely a chance correlation, the above analysis was repeated ten times. The results are shown in Table 1. Table 1 also shows equivalent results for several different forms of heteroscedastic relationship with the total noise power adjusted to be constant for all noise types. It consistently shows negligible modelling capability (high RMSEP value) on homoscedastic noise, but significantly better modelling ability (low RMSEP value) on heteroscedastic noise with the simplest heteroscedastic relations giving the best modelling. The fact that the predictions are carried out on the same noise matrix for both hetero- and homoscedastic cases also precludes the possibility of chance correlations since a chance correlation in the heteroscedastic dataset would also produce the same chance correlation, and consequently similarly good prediction, in the homoscedastic dataset.

The reduced prediction efficiency on noise variables with the more complicated nonlinear heteroscedastic relations is most likely due to the fact that PLS is not optimal for modelling these nonlinearities. Accordingly a neural net model was formed using NeuralDesk (Neural Computer Sciences, Lulworth Business Centre, Nutwodd Way, Totton, Southampton, Hampshire, UK) on the logarithmic dataset in Table 1 (the dataset with the highest RMSEP value). This improved the modelling of pure noise variables to an RMSEP of 25.27%, proving this hypothesis.

This has very important ramifications for variable selection procedures that are widely used in order to select only those variables most relevant to a modelling process and for modelling in general. In the presence of only homoscedastic noise, the modelling process is improved in both speed and precision by eliminating the (uncorrelated) noise variables. The result of progressively removing the high frequency noise variables on the RMSEP of a multivariate model (one factor optimal as above) on the dataset including homoscedastic noise is shown in Fig. 7. As might be expected, the prediction begins to degrade markedly only when the variables pertaining to the signal begin to be removed (at about bin number 10).

However, in the presence of heteroscedastic noise, the (now correlated) noise variables contribute to the modelling to some degree and variable selection has

Table 1
%RMSEP values for ten consecutive homo- and heteroscedastic noise predictions for several different heteroscedastic noise types

Heteroscedastic noise function	%RMSEP
α (homoscedastic)	67.50 ± 7.48
$\alpha 1 * s(t)$ (linear)	18.03 ± 3.14
$\alpha 2 * s(t)^2$ (squared)	25.68 ± 6.78
$\alpha 3 * s(t)^3$ (cubed)	30.43 ± 3.15
$\alpha 4 * \exp(s(t)/10)$ (exponential)	29.97 ± 5.30
$\alpha 5 * \ln(s(t) * 50)$ (logarithmic)	42.11 ± 7.24
$\alpha 6 * (1/(1 + \exp(-s(t))))$ (sigmoid)	38.06 ± 3.07

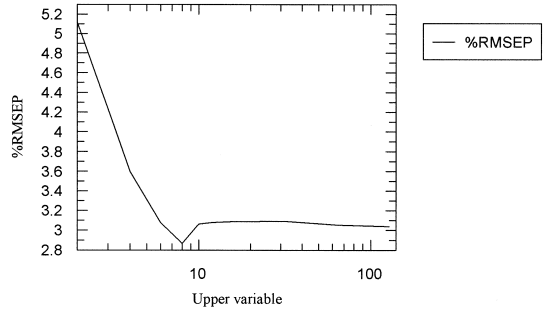

Fig. 7. Result of progressively pruning high frequency variables on the prediction of the dataset including homoscedastic noise.

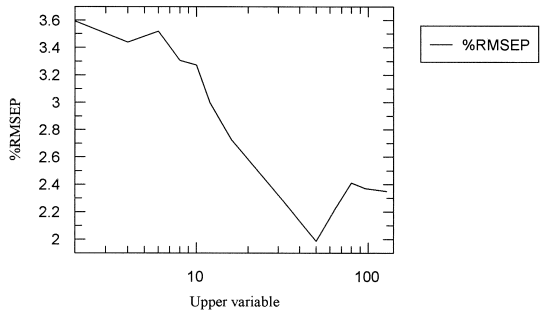

Fig. 8. Result of progressively pruning high frequency variables on the prediction of the dataset including heteroscedastic noise.

less clearcut benefits. The process may end up as a trade off between improved modelling speed and degraded modelling precision as the now correlated noise variables are progressively rejected. The corresponding result of progressively removing the high frequency noise variables on the RMSEP of a multivariate model (two factors optimal as above) on the dataset including linear heteroscedastic noise is shown in Fig. 8. Here the prediction begins to degrade when many noise variables are still present.

5. Conclusion

It can be seen that when a frequency transform is applied to heteroscedastic noise, the resulting noise in the frequency domain can be strongly correlated with the signal. Since the signal is correlated with the reference in any well designed experiment the noise will also be correlated with the reference. This contrasts with the uncorrelated noise produced by homoscedastic processes. *Any modelling process used on the frequency domain may consequently form a significant part of its model on the noise variables*. Modelling will occur across the entire frequency space. This is again in complete contrast to the homoscedastic case in which modeling on these noise variables is poor.

It is also worth mentioning that the datasets including heteroscedastic noise require more PLS factors for optimal modelling than does that including homoscedastic noise, since the noise and signal are no longer independent, but interact in a manner similar to that noticed previously in evolving principal components analysis [9].

As can be seen from the results in Figs. 7 and 8, the task of satisfying the parsimony principle [13], which seeks to select only the most relevant variables from which to form a model in order to increase the efficiency and precision of the modelling process, is complicated by the presence of heteroscedastic noise.

There is much current interest in the removal of noise from noisy signals and spectra, such techniques being commonly referred to as 'denoising' [14–18]; it is evident from the present analysis that the optimal exploitation of such methods will depend greatly upon whether the noise structure itself is homo- or heteroscedastic and upon the form of that structure.

In spectra depending on several components, it will be problematic to identify regions dependent on only one component since information on all components is convolved across all frequencies. The problem of component-separation is complicated by the presence of heteroscedastic noise.

Finally, there is also the intriguing possibility that signals could be modelled on their 'noise' components alone even if the detector used is not capable of receiving the signal frequencies themselves. The downside of this is that signals deliberately rejected by detector filters could still bleed through to the modelling process via their noise.

Acknowledgements

We thanks the Chemicals and Pharmaceuticals Directorate of the UK BBSRC, Glaxo-Wellcome and Bruker Spectrospin for financial support.

References

[1] C. Chatfield, A.J. Collins, Introduction to Multivariate Analysis, Chapman and Hall, London, 1980.
[2] P. Bhandare, Y. Mendelson, E. Stohr, R.A. Peura, Appl. Spectrom. 48 (1994) 271.
[3] H.R. Bjorsvik, E. Bye, Appl. Spectrosc. 45 (1991) 771.
[4] M.K. Alam, J.B. Callis, Anal. Chem. 66 (1994) 2293.
[5] D.M. Haaland, Practical Fourier Transform Infrared Spectroscopy, vol. 395, 1990.
[6] J.H. Linn, K.L. Hanley, Appl. Spectrosc. 47 (1993) 2102.

[7] H. Martens, T. Næs, Multivariate Calibration, John Wiley, Chichester, 1989.
[8] R.R. Sokal, F.J. Rolfe, Biometry: The Principles and Practice of Statistics in Biological Research, 2 ed., Freeman and Co., New York, 1981.
[9] O.V. Kvalheim, F. Braksted, Y. Liang, Anal. Chem. 66 (1994) 43.
[10] C. Ritter, J.A. Gilliard, J. Cumps, B. Tilquin, Anal. Chim. Acta 318 (1995) 125.
[11] J. Toft, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 19 (1993) 65.
[12] G.D. Bergland, IEEE Spectrum 7 (1969) 41.
[13] M.B. Seasholtz, B. Kowalski, Anal. Chim. Acta 277 (1993) 165.
[14] R.R. Coifman, M.V. Wickerhauser, Opt. Eng. 33 (1994) 2170.
[15] D.L. Donoho, I.M. Johnstone, C. R. Acad. Sci. Ser. A: 319 (1994) 1317.
[16] D.L. Donoho, I.M. Johnstone, J. Am. Stat. Assoc. 90 (1995) 1200.
[17] G.P. Nason, J. R. Stat. Soc. Ser. B: 58 (1996) 463–479.
[18] C.R. Vogel, M.E. Oman, SIAM J. Sci. Comput. 17, 227.