# Determination of ingredients in packaged pharmaceutical tablets by energy dispersive Xray diffraction and maximum likelihood principal component analysis multivariate curve resolutionalternating least squares with correlation constraint

Journal Item

Received: 2 July 2020 | Revised: 18 December 2020 | Accepted: 11 January 2021

DOI: 10.1002/cem.3329

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

# Determination of ingredients in packaged pharmaceutical tablets by energy dispersive X-ray diffraction and maximum likelihood principal component analysis multivariate curve resolution-alternating least squares with correlation constraint

**Peter S. Kenny**[1] | **Chiaki Crews**[2] | **Tom Fearn**[1] | **Robert D. Speller**[2]

[1]Department of Statistical Science, University College London, London, UK

[2]Department of Medical Physics & Biomedical Engineering, University College London, London, UK

**Correspondence**
Peter S. Kenny, Department of Statistical Science, University College London, London WC1E 6BT, UK.
Email: peter.kenny.14@ucl.ac.uk

**Funding information**
Engineering and Physical Sciences Research Council, Grant/Award Numbers: EP/G037264/1, EP/M506448/1

## Abstract

Energy dispersive X-ray diffraction (EDXRD) and maximum likelihood principal component analysis multivariate curve resolution-alternating least squares (MLPCA-MCR-ALS) with correlation constraint were used to quantify the composition of packaged pharmaceutical formulations. Recorded EDXRD profiles from unpackaged and packaged samples of ternary mixtures were modelled together in order to recover the concentrations as well as the pure profiles of the constituent compounds. MLPCA was used as a data pretreatment step to MCR-ALS, accounting for the high noise and nonconstant variance observed in the EDXRD profiles and was shown to improve the resolution accuracy of MCR-ALS for the data set. Local correlation constraints were applied in the MCR-ALS procedure in order to model unpackaged and packaged samples simultaneously while accounting for the matrix effect of the packaging materials. The composition of the formulations was estimated with root-mean-square error of prediction for each component, including paracetamol, being approximately 2.5 %w/w for unpackaged and packaged samples. Paracetamol concentration was resolved simultaneously for the unpackaged and packaged samples to a greater degree of accuracy than achieved by partial least squares regression (PLSR) when modelling the contexts separately. By modelling the effects of the packaging and incorporating accurate reference information of unpackaged samples into the resolution of packaged samples, the potential of EDXRD and MLPCA-MCR-ALS for the identification and quantification of packaged solid-dosage medicine in nondestructive screening and counterfeit medicine detection has been raised.

**KEYWORDS**

correlation constraint, counterfeit medicine, energy dispersive X-ray diffraction, multivariate curve resolution, packaged formulations

# 1 | INTRODUCTION

Nondestructive volume characterisation of materials located beneath solid surface layers, such as the contents of packaged or concealed goods, requires high-energy photon flux to overcome attenuation effects. Energy dispersive X-ray diffraction (EDXRD) can meet this requirement and measures the coherent scattering fields, which characterise the samples comprising crystalline and polycrystalline materials. Polycrystalline materials are of particular interest in screening contexts, where investigations have shown the potential of EDXRD in identifying the composition of solid-dosage pharmaceutical formulations, powder-form illicit drugs and plastic explosive materials.[1–4] Furthermore, the technology has shown potential in diagnostic medicine to discriminate between healthy and cancerous tissue in breast cancer screening.[5]

Traditional application of X-ray diffraction in a laboratory setting, either with a synchrotron radiation source or a conventional mono-energetic angular dispersive X-ray diffraction (ADXRD) system, enables high-resolution sample profiles to be recorded. The profiles have high spectral selectivity and can be parametrically fitted and then cross-referenced against databases in order to characterise and identify the samples.[6] When performing in situ EDXRD for screening purposes, however, we must accept lower resolution profiles with broad peaks and high levels of noise in order to achieve surface penetration, rapid screening and portability.[6,7] The resulting profiles often have highly overlapping characteristic peaks making interpretation more difficult.

To overcome the collinearity and overlap within profiles, well-established multivariate calibration methods using orthogonal latent variables, such as principal components regression (PCR) and partial least squares regression (PLSR), have been used in EDXRD studies to quantify target compounds within mixtures.[1,3,8] A disadvantage of this approach is that the latent variables resolved cannot be easily interpreted since they do not represent a pure instrumental response of a single compound.

Multivariate curve resolution-alternating least squares (MCR-ALS) is a powerful and popular soft-modelling method, which aims to resolve simultaneously the concentration profiles and the individual pure instrumental responses for many or all compounds found within a mixture set of samples.[9] MCR-ALS has been used in a broad range of applications to achieve successful resolution when the data adheres approximately to an underlying bilinear model.[10] By incorporating constraints within the alternating least squares routine, solutions with meaningful physical and chemical interpretations can be recovered for quantification and identification.

In this study, ternary mixtures of synthetic samples comprising common pharmaceutical ingredients pressed into tablets have been analysed to resolve their concentrations and the pure instrumental responses in both unpackaged and packaged scenarios. We build on the study by Crews et al., in which the bilinearity of the data set was demonstrated and unpackaged and packaged tablets were modelled separately using PLSR.[1] In the present study, unpackaged and packaged formulations are modelled together in a multiset structure to resolve concentrations for all samples for a given compound as well as a pure instrumental response for each compound. We believe this presents a more useful, realistic and flexible framework, in which accurate reference information for compound concentrations in unpackaged samples is used to resolve unknown concentrations for packaged sample compounds during screening analysis.

Lyndgaard et al. first developed the implementation of local calibration models within MCR-ALS and applied this method to Raman spectroscopy data from ternary mixtures of synthetic tablets in blister packaging to successfully quantify their paracetamol content.[11] In the present study, the simultaneous resolution of the profiles from unpackaged and packaged tablets—modelled in a multiset structure using the local correlation constraint—is achieved to determine the concentrations of paracetamol from the EDXRD data.

# 2 | THEORY

## 2.1 | Energy dispersive X-ray diffraction

In the EDXRD technique, a polychromatic X-ray source is used to irradiate samples and an energy dispersive profile of the scattered X-rays is recorded by a fixed-angle detector. As with the more commonly used ADXRD technique, the objective is to differentiate and characterise crystalline and polycrystalline materials by the interplanar spacings of their crystal planes. Bragg's law defines that constructive interference of coherently scattered X-rays occurs when the X-ray wavelength, $\lambda$, scattering angle, $\theta$, and the planar spacing, $d$, meet the following condition:

$$n\lambda = 2d\sin\theta \tag{1}$$

where $\lambda$ is inversely proportional to the energy of the scattered X-ray and $n$ is any integer. By using a wide-band source, an energy dispersive detector, and keeping the diffraction angle fixed, the locations of peaks in energy space correspond to the planes of the crystal. In order to compare profiles between EDXRD systems of different arrangements and to compare between the two modes, ADXRD and EDXRD, it is useful to convert angle and energy units into units of momentum transfer, $x$, common to all systems. By rearranging Equation 1 we find momentum transfer, $x$:

$$x = \frac{1}{2d} = \frac{1}{\lambda}\sin\theta = \frac{E}{hc}\sin\theta \tag{2}$$

where $E$ is the energy of the X-ray photon, $h$ is Planck's constant and $c$ is the speed of light in vacuo, and we use the relationship between photon energy and wavelength:

$$\lambda = \frac{hc}{E} \tag{3}$$

There are several advantages of EDXRD over ADXRD in screening scenarios. Firstly, its capability is to collect data rapidly by using a polychromatic source which has greater flux. Secondly, its lack of moving parts allows the equipment to be more portable and easier to use outside of a laboratory environment, and lastly, its use of higher energies for overcoming the effect of beam attenuation in thicker and shielded samples.

The trade-off for rapid screening and volume scanning is that greater peak broadening and higher noise levels are observed in EDXRD profiles, which is largely due to uncertainty in the true scattering angle.[7] In samples comprising multiple polycrystalline compounds, recorded scattering profiles from EDXRD are the result of scattering events from all constituent compounds, which form linear and additive contributions to the recorded profiles. There is significant peak overlap observed in the profiles of mixtures and little or no selectivity in the momentum transfer space variables. These properties motivate the use of factor analysis methods for recorded profiles to determine their composition and pure response in order to enable quantification and identification. An advantage of EDXRD over spectroscopic methods such as Raman and NIR is that profile features, including peak shape, location, intensity and attenuation, can be readily related to the underlying physical phenomena of the samples and the experimental setup. MCR-ALS is a good candidate for modelling the EDXRD data as it is flexible to the incorporation of physical and chemical information into the modelling procedure without requiring full parametrisation as with hard modelling methods.

## 2.2 | MCR-ALS

MCR-ALS is based on the bilinear decomposition of a data matrix $\boldsymbol{D}$ into the concentrations of constituent compounds and their corresponding pure instrumental profiles:

$$\boldsymbol{D} = \boldsymbol{C}\boldsymbol{S}^T + \boldsymbol{E} \tag{4}$$

where $\boldsymbol{C}$ is a matrix of concentrations of the chemical constituents, $\boldsymbol{S}$ is a matrix of pure instrumental responses of the constituents and $\boldsymbol{E}$ is the residual matrix of variation not explained by the bilinear model. The first step of the decomposition is to reproduce the data matrix $\boldsymbol{D}$ by calculating abstract factors which span a subspace with rank equal in number to the varying chemical components in the system. Typically, the rank can be estimated by finding and using the number of significant singular values in a singular value decomposition, or if the chemical rank is already known, then by simply using the known chemical rank. For standard MCR-ALS, the data are reproduced by projecting the data matrix into principal component space of the determined rank, to form a matrix $\boldsymbol{D}_{PCA}$. In this study, to account for high levels of nonhomoscedastic noise, maximum likelihood principal component analysis (MLPCA) has also been used in the reproduction of the data matrix, forming $\boldsymbol{D}_{MLPCA}$, and compared to the performance of standard MCR-ALS.[12]

Where pure samples or pure variables are present in the data set, the problem of resolution is simplified greatly, and a classical least squares calculation can recover the corresponding pure spectra or concentrations, respectively. Often such selectivity in the data is not present, as is the case in the present study, and an iterative procedure is followed to estimate $\boldsymbol{C}$ and $\boldsymbol{S}$ in which known physical and chemical information about the system is included to transform the abstract factors into meaningful chemical factors. Before initialising the iterative procedure, initial solutions of either the concentration matrix or the pure spectrum matrix must be estimated. In this study the purest variable selection method SIMPLISMA is used, which determines the purest samples or purest momentum transfer space variables in the data set.[13] Where initial purest samples for $\boldsymbol{S}$ are determined by SIMPLISMA, the concentration matrix is estimated by a least squares application:

$$\boldsymbol{C} = \boldsymbol{D}_{\boldsymbol{PCA}} \left(\boldsymbol{S}^{T}\right)^{+} \tag{5}$$

where $+$ denotes the Moore-Penrose pseudoinverse of a matrix. Conversely, the pure instrumental response matrix is calculated by least squares if purest momentum transfer variables are found:

$$\boldsymbol{S} = \boldsymbol{C}^{+} \boldsymbol{D}_{\boldsymbol{PCA}} \tag{6}$$

Following the calculation of initial solutions, the alternating least squares procedure is initiated, and the matrices $\boldsymbol{C}$ and $\boldsymbol{S}$ are calculated iteratively using constrained least squares applications of Equations 5 and 6. In this study, the non-negativity constraint is applied in both directions using the *lsqnonneg* function in MATLAB. For the concentration direction, the correlation constraint (Section 2.3) is applied to the paracetamol concentration vector using reference values from a calibration set of samples. Equality constraints are applied for the concentration vectors of caffeine and microcrystalline cellulose—which are the remaining compounds in the synthetic tablets that were analysed—by substituting reference concentration values of samples in the calibration set.

The matrices $\boldsymbol{C}$ and $\boldsymbol{S}$ that are resolved at a particular iteration of MCR-ALS are compared to the original data matrix, $\boldsymbol{D}$, by calculating their matrix product:

$$\hat{\boldsymbol{D}} = \boldsymbol{C}\boldsymbol{S}^{T} \tag{7}$$

Following which a goodness of fit to the original data matrix is evaluated by a percentage lack-of-fit measure:

$$LOF\,(\%) = 100 \times \sqrt{\frac{\sum_{i,j}\left(d_{i,j} - \hat{d}_{i,j}\right)^{2}}{\sum_{i,j} {d_{i,j}}^{2}}} \tag{8}$$

where $i$ denotes the sample and $j$ denotes the instrumental response variable for the original data matrix element $d$ and the element of the reproduced model data matrix $\hat{d}$. For constraints that are appropriately selected and applied for the system being studied, the *LOF* should decrease between subsequent iterations of MCR-ALS until convergence is achieved. The concentrations and pure signal matrices in the final iteration are deemed to be the optimal solutions for the algorithm implemented. Additional figures of merit to evaluate the performance of the procedure used in this study are the root-mean-square error of prediction (RMSEP), for determining prediction accuracy for each compound, and $R^2$ for the predicted against reference values to calculate the goodness of fit to the model for each compound.

Solutions which result in the optimal LOF may not be unique. Indeed, there may be a range of solutions of $\boldsymbol{C}$ and $\boldsymbol{S}$ which fit equally well to the data matrix, and the solution in this case is ambiguous.[14] However, it has been revealed that when a correlation constraint is applied to a chemical component, and in the case where all components that are present outside of the calibration set are also present within the calibration set, a unique solution is obtained for the constrained component.[15]

## 2.3 | Correlation constraint

The correlation constraint is used to scale the values in a concentration vector into real concentration units by incorporating known reference values into the procedure thus improving the accuracy and interpretability of the resolved concentrations. The constraint can be applied to some or all of the component vectors within the concentration matrix. For each component to which it is applied, an internal calibration model is formed for the samples in a calibration set in order to update the concentrations of the samples in the test set by aligning their scale with that of known reference concentrations.

Reference and resolved concentration values, $\boldsymbol{c}^*$ and $\boldsymbol{c}$ respectively, are modelled by a simple linear regression:

$$\boldsymbol{c} = b_1\boldsymbol{c}^* + b_0 + \boldsymbol{e} \tag{9}$$

The concentrations in the calibration set which have been recovered by least squares, according to Equation 5, under the constraint of nonnegativity, are regressed against real-scaled and known concentrations:

$$\boldsymbol{c}_{cal} = b_1\boldsymbol{c}^*_{cal} + b_{\boldsymbol{0}} \tag{10}$$

The regression coefficients recovered are then used to recalculate the concentrations of the samples in the test set, which are those samples not included in the calibration set:

$$\boldsymbol{c}^*_{\boldsymbol{test}} = (\boldsymbol{c}_{test} - b_0)/b_1 \tag{11}$$

The vector formed by $\left[\boldsymbol{c}^*_{cal};\boldsymbol{c}^*_{test}\right]$ is merged into the concentration matrix $\boldsymbol{C}$ of all components for the subsequent iteration. This procedure has been applied to the paracetamol concentration vector when resolving unpackaged and packaged sets of EDXRD profiles separately. At each subsequent iteration, the resolved concentrations in the calibration set approach those of the reference concentrations in the case when the constraint is applied appropriately. In this case, the coefficient $b_0$ tends to zero and the coefficient $b_1$ tends to unity. Furthermore, an appropriate use of the correlation constraint should not result in a large increase in the percentage lack-of-fit of the MCR-ALS model to the original data matrix when compared to MCR-ALS applied without the correlation constraint. This rule of thumb should be applied to the application of any constraint in MCR-ALS given that the LOF criterion closely corresponds to the objective function in least squares optimisation.

## 2.4 | Multiset modelling with local calibration models

The objective of this study is to determine the feasibility of modelling EDXRD profiles from unpackaged and packaged samples of synthetic pharmaceutical formulations simultaneously. MCR-ALS is flexible to modelling multiple subsets of data simultaneously in the case that there is correspondence between the subsets in the concentration direction, spectral direction, or in both directions, and is known as multiset modelling.

When modelling both sets together, the first step is to define an augmented data matrix, in accordance with Equation 4, which in this study is composed of data from unpackaged and packaged samples over the same range of momentum transfer space:

$$\boldsymbol{D} = \begin{pmatrix}\boldsymbol{D}_u\\ \boldsymbol{D}_p\end{pmatrix} = \begin{pmatrix}\boldsymbol{C}_u\\ \boldsymbol{C}_p\end{pmatrix}\boldsymbol{S}^T + \begin{pmatrix}\boldsymbol{E}_u\\ \boldsymbol{E}_p\end{pmatrix} \tag{12}$$

where $u$ denotes the unpackaged subset and $p$ the packaged subset. Both subsets are decomposed with the same instrumental response matrix $\boldsymbol{S}$. When modelling unpackaged and packaged subsets together, a fourth component is included in $\boldsymbol{C}$ and $\boldsymbol{S}$ in order to model the scattering and attenuation contribution of the packaging material. This component was constrained by nonnegativity in the concentration and instrumental response direction during MCR-ALS but is otherwise freely modelled.

This study applies the methodology developed by Lyndgaard et al. who adapted the correlation constraint to work in a multiset structure in which a local calibration model is formed for each subset.[11] This adaptation is beneficial when there are differing matrix effects present in the subsets, meaning that the linear correspondence between the real **C** for a subset and the measurement matrix $\boldsymbol{D}$ are unequal between subsets.

To account for matrix effects, separate internal calibration models using the correlation constraint are applied using the calibration samples within each subset. The two local models are defined as follows:

$$\begin{aligned} \boldsymbol{c}_u &= b_{u,1}\boldsymbol{c}_u^* + b_{u,0} + \boldsymbol{e_u} \\ \boldsymbol{c}_p &= b_{p,1}\boldsymbol{c}_p^* + b_{p,0} + \boldsymbol{e_p} \end{aligned} \tag{13}$$

where $\boldsymbol{c}^*$corresponds to the correctly scaled reference concentrations and $\boldsymbol{c}$ to the concentrations resolved by the least squares step.[5] Calibration coefficients $b$ are calculated using the resolved concentrations and reference concentrations in the calibration sets for the two subsets:

$$\begin{aligned} \boldsymbol{c}_{u,\ cal} &= b_{u,1}\boldsymbol{c}_{u,cal}^* + b_{u,0} \\ \boldsymbol{c}_{p,\ cal} &= b_{p,1}\boldsymbol{c}_{p,cal}^* + b_{p,0} \end{aligned} \tag{14}$$

For the remaining samples, which form the test sets for the two subsets, the concentrations are recalculated as follows:

$$\begin{aligned} \boldsymbol{c}_{u,test}^* &= (\boldsymbol{c}_{u,test} - b_{u,0})/b_{u,1} \\ \boldsymbol{c}_{p,test}^* &= \left(\boldsymbol{c}_{p,test} - b_{p,0}\right)/b_{p,1} \end{aligned} \tag{15}$$

These predicted concentrations are recorded at each iteration and at the final iteration represent the optimum solutions for both subsets. However, the local calibration models present a different relationship between the concentration vector and the data matrix, and therefore do not correspond to a common instrumental response matrix, as is required for the defined bilinear decomposition in the multiset structure. To preserve the bilinear relationship, and to ensure a good fit of the model to the data, one of the subsets must be rescaled in order to conform to the scale of the other subset. In this study, the packaged concentrations are rescaled according to the unpackaged regression model before the ALS procedure continues to the next iteration:

$$\begin{aligned} \boldsymbol{c}_{p,cal,rescaled}^* &= \left(b_{p,1}\boldsymbol{c}_{p,cal}^* + b_{p,0} - b_{u,0}\right)/b_{u,1} \\ \boldsymbol{c}_{p,test,rescaled}^* &= \left(b_{p,1}\boldsymbol{c}_{p,test}^* + b_{p,0} - b_{u,0}\right)/b_{u,1} \end{aligned} \tag{16}$$

$[\boldsymbol{c}_u^*; \boldsymbol{c}_{u,test}^*; \boldsymbol{c}_{p,cal}^*; \boldsymbol{c}_{p,test}^*]$ is saved as the resolved concentrations vector for the iteration and is then rescaled to $[\boldsymbol{c}_u^*; \boldsymbol{c}_{u,test}^*; \boldsymbol{c}_{p,cal,rescaled}^*; \boldsymbol{c}_{p,test,rescaled}^*]$ to be passed to the next iteration of the alternating least squares procedure.

## 2.5 | Heteroscedastic noise

EDXRD data have high levels of nonhomoscedastic noise, but MCR-ALS does not account for heteroscedastic error structure within the data matrix. The objective function minimises the total squared error by equally penalising errors across profiles and variables. The optimal solution to the underlying bilinear profiles in the data may therefore not be obtained as some elements of $\boldsymbol{D}$ will effectively be overfitted and others underfitted, according to their inherent variability. Two methods to overcome this problem which have been implemented in previous work within the MCR-ALS scheme are weighted alternating least squares (MCR-WALS) and maximum likelihood PCA (MLPCA-MCR-ALS).[16,17] Both methods account for a heteroscedastic error structure. A previous study by Dadashi et al. has shown that when applying MCR-ALS, both methods result in very similar solutions, with similar performance.[17] In

both methods, the variances of the data points must be known or estimated in advance, however MCR-WALS incorporates error weighting into each iteration of the alternating least squares scheme. Conversely, MLPCA-MCR-ALS undertakes a pretreatment step to reproduce the data matrix with an approximately homogenous error structure prior to the initiation of the ALS algorithm. Due to the similar reported performances of MCR-WALS and MLPCA-MCR-ALS and given that MLPCA need only be applied as a pretreatment step, MLPCA has been favoured for use in this study.

## 3 | EXPERIMENTAL

The EDXRD system setup, sample preparation and data acquisition were described in detail by Crews et al.[1] The relevant information for multivariate analysis is presented here.

### 3.1 | Sample preparation

The samples resolved using MCR-ALS were binary and ternary mixtures of powder-form paracetamol, caffeine and microcrystalline cellulose at eight concentration levels (0 %w/w–80 %w/w) according to the design shown in Figure 1. The mixtures were weighed and pressed into cylindrical tablets with diameter 1.3 cm, mass 0.4 g, thicknesses between 0.25 and 0.30 cm and their concentrations were recorded for reference. A pure sample for each compound was also prepared and the corresponding EDXRD profiles were recorded and used for validation purposes. The packaging comprised card, aluminium foil and polyvinyl chloride (PVC). The nominal thicknesses of the materials were 0.075 cm, 0.002 and 0.025 cm, respectively. They were cut to size and positioned such that the tablets had aluminium foil and card on one side, and PVC and card on the other.

### 3.2 | Data acquisition

The X-ray source was operated at 60-kV peak voltage. All samples were irradiated for 300 s, both in and out of packaging with raster scanning employed to mitigate preferred-orientation artefacts. Each event recorded by the fixed-angle detector was binned into one of 512 energy channels using a multichannel analyser, and these recorded energy profiles were subsequently transformed into momentum transfer space according to Equation 2. Triplicate measurements were taken on all samples, both in and out of packaging, and the measurements were averaged for each sample prior to

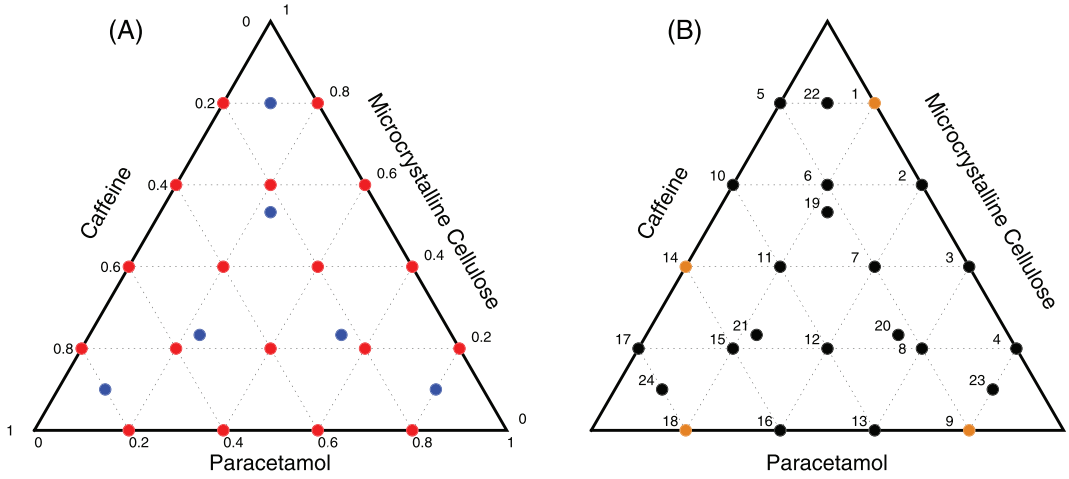


**FIGURE 1** Ternary mixture design of formulations used in the experiment. (A) Samples in red were included in the modelling for unpackaged and packaged subsets. Samples in blue were used only for the unpackaged subset. (B) Sample numbers from 1 to 24. Samples in orange were used for calibration in the models

modelling. A diffraction profile of the card and blister packaging together, without a sample, was recorded for 300 s for the purpose of obtaining an initial estimate of the background component for the model.

### 3.3 | Data preprocessing and selection

In order to recover interpretable solutions, the data was analysed without derivative pretreatment. Other pretreatments such as scatter corrections and the standard normal variate transformation were also avoided as they were shown in the previous study on this data, which used multivariate calibration, to not improve the models. Furthermore, and crucially for this study, they can lead to invalid interpretations as well as spoiling the bilinearity of the data.[18] The profiles for packaged samples were however corrected for the attenuation effect of the packaging material. The effect is a nonconstant function of the attenuating material and of photon energy and, as a result, the bilinear correspondence between unpackaged and packaged samples is not preserved in the measured data matrix. In order to restore the bilinear correspondence between subsets, an attenuation correction was applied for all packaged samples according to the relationship:

$$d^p_{0,j} = d^p_j \prod_{i=1}^{3} \exp\left(\mu_{i,j}(E_j)\rho_i x_i\right) \tag{17}$$

where $d^p_{0,j}$ is the corrected data point with corresponding channel $j, d^p_j$ is the measured data point and $E_j$ is the energy of the photon at channel $j$. The three materials corresponding to $i$ are aluminium, PVC and cellulose (card) and their linear attenuation coefficients $\mu$, measured in $cm^2/g$, were obtained from the NIST X-COM database.[19] The density and thicknesses of the packaging materials are denoted by $\rho$ and $x$ respectively.

All data analysis was performed in MATLAB version R2018a (The MathWorks, Natick, MA, USA).

## 4 | RESULTS AND DISCUSSION

### 4.1 | Exploratory analysis

Binary and ternary mixtures of paracetamol, caffeine and microcrystalline cellulose in tablet form were investigated to quantify their chemical composition and resolve pure component signals. A momentum transfer window of 0.50 to 2.04 $nm^{-1}$ (11.3 and 46.1 keV) was used for the recorded EDXRD profiles for multivariate analysis. In this window, characteristic diffraction peaks for paracetamol, caffeine and microcrystalline were observed. The EDXRD profiles of the pure samples of triplicate-averaged measurements are shown in Figure 2A for unpackaged samples and Figure 2D for the same pure samples when packaged. Peaks for paracetamol are located at 0.8 and 1.25 $nm^{-1}$. There are two main peaks for caffeine: one at 1.5 $nm^{-1}$, and a very high intensity peak located at a lower momentum transfer of 0.7 $nm^{-1}$. Microcrystalline cellulose has peaks at 0.8 and 1.3 $nm^{-1}$, and heavy overlap between the peaks of paracetamol and microcrystalline cellulose is evident. Momentum transfer values below 0.5 $nm^{-1}$ (11 keV) were not investigated as the attenuation effect of the tablets themselves is significant for the thicknesses of tablets used in this study (0.25–0.3 cm). In addition to detecting only a weak signal in this region, we also lose the bilinearity of the data. The attenuation effects of the packaging, when comparing Figures 2A and 2D, are very pronounced for low momentum transfer values in this range, as expected due to the higher scattering cross section of the packaging materials for lower energies.[19] The EDXRD profile of the packaging itself was recorded (Figure 2D) and the amorphous nature of the packaging materials results in a broad peak which is centred around 1.1 $nm^{-1}$.

Two examples of packaged and unpackaged sample profiles are shown in Figure 2B,C corresponding to a ternary mixture of paracetamol, caffeine and microcrystalline cellulose, and a binary mixture of paracetamol and caffeine, respectively. The packaging has a significant attenuation effect on the packaged tablet profiles for low momentum transfer, as exemplified in Figure 2C by the attenuation of the caffeine peak at 0.7 $nm^{-1}$. The attenuation-corrected profiles (green) show close correspondence in intensity of the first caffeine peak with the profiles, in a region where there is little scattering contribution from the packaging material. This shows initial promise that the method for correcting for packaging attenuation and the corresponding parameters for material thickness may have been selected

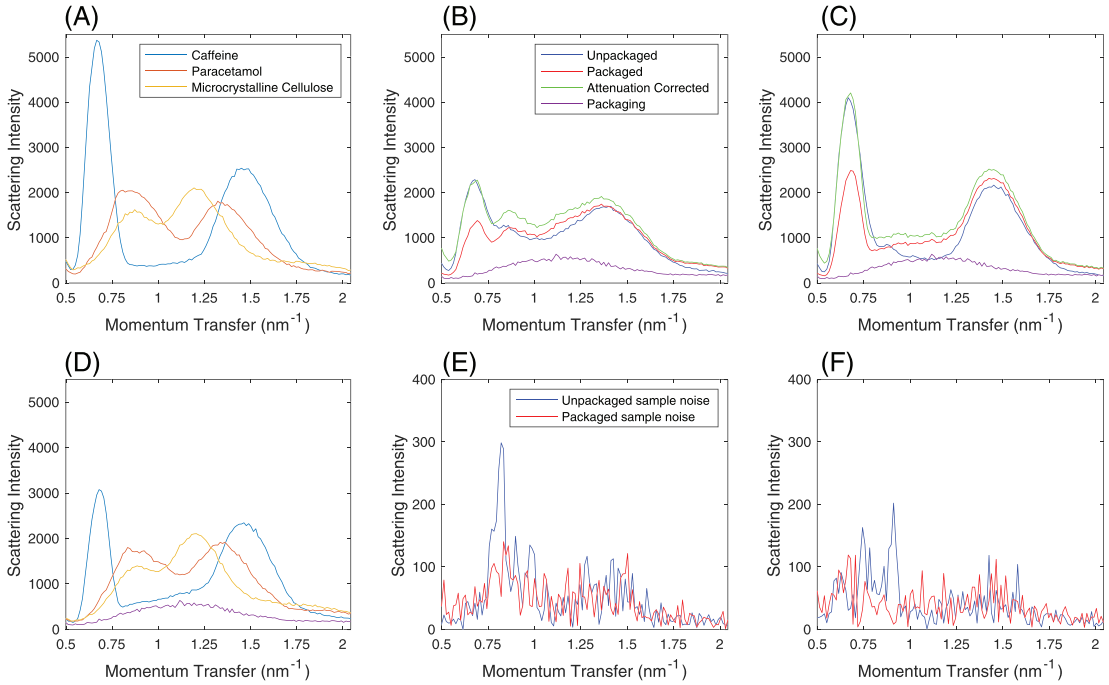


**FIGURE 2** Pure profiles for the three compounds while unpackaged (A) and packaged (D). Ternary mixture of 20% paracetamol, 40% caffeine and 40% microcrystaline cellulose (B) and the corresponding noise levels (E). Binary mixture of 80% caffeine and 20% paracetamol (C) and noise levels (F)

appropriately. For higher momentum transfer, around 1.1 nm$^{-1}$, the attenuation-corrected packaged profiles have greater intensity than the unpackaged profiles, where the scattering contribution of the packaging is more significant than the attenuation. It is hypothesised that by including the scattering contribution from the packaging as a fourth component in the MCR-ALS modelling, the tablets' constituent compounds can be modelled and resolved in the unpackaged and packaged samples simultaneously. Note that in this EDXRD setup, the scattering volume extends several centimetres to include the packaging material and thus the exact positioning of the packaging does not significantly affect the recorded profile, unlike with ADXRD where the positioning is critical. Any residual matrix effect which differentiates the unpackaged and packaged subsets will be accounted for through the application of internal calibration models for each subset using the correlation constraint. The results of the proposed model are reported in Section 4.2.3.

The corresponding noise levels (Figure 2E,F), calculated as the sample standard deviation of triplicate measurements on each sample, for the ternary and binary samples, and indeed all samples, are not constant across the profiles. The percentage noise level is above 15% for some channels, in particular those corresponding to the first peak of paracetamol between 0.8 and 1.0 nm$^{-1}$. This is due to the preferred orientation effect of paracetamol crystals resulting in the overrepresentation of some lattice planes and underrepresentation of others, depending on the alignment of crystals in the tablet region being scanned at the time. This propagates as higher levels of variation in repeated measurements, despite employing raster scanning to mitigate this to a large degree.[1] MLPCA is a method which has been reported to improve the results of MCR-ALS when applied to data sets with high levels of noise and where the noise is not constant between and within samples. MLPCA was applied as a pretreatment step in the reproduction of the data matrix prior to initialising the alternating least squares procedure for the analysis that follows.

## 4.2 | Quantitative analysis of paracetamol

### 4.2.1 | Quantitative analysis of paracetamol in unpackaged samples

EDXRD profiles from 24 binary and ternary mixtures of unpackaged tablets (Figure 1) were resolved into concentration and pure momentum transfer profiles by using MCR-ALS.

The first step undertaken was to reproduce the data matrix $\boldsymbol{D}$ by singular value decomposition with the first three singular values chosen in accordance to the known chemical rank of the data. A new matrix was subsequently formed, $\boldsymbol{D}_{PCA}$, to be factorised into concentrations and pure momentum transfer profiles by MCR-ALS. The same procedure was applied when using MLPCA as a pretreatment to the data resulting in a matrix $\boldsymbol{D}_{MLPCA}$. As the noise is sample as well as momentum transfer dependent, the sample standard deviation of the triplicate measurements for each momentum transfer value per sample was calculated and was used for the standard deviation matrix, required to compute MLPCA. Another potentially more robust approach is to account for anticipated structure in the noise, such as modelling proportional noise for counting data, or modelling the covariance structure between samples. The optimal model potentially involves both proportional noise and covariance, as we observe that the variance is derived from preferred orientation effects, varying from sample to sample, and counting statistics, directly related to the intensity of variable. We believe this represents an interesting area of further investigation.

Initial solutions for MCR-ALS were obtained by the purest variable selection method SIMPLISMA, which obtained the three purest samples from $\boldsymbol{D}_{PCA}$or $\boldsymbol{D}_{MLPCA}$ to form the initial pure profile matrix $\boldsymbol{S}$. An initial concentration matrix $\boldsymbol{C}$ was obtained from an unconstrained least squares step using $\boldsymbol{D}$ and the initial estimates of $\boldsymbol{S}$.

The alternating least squares procedure was then initiated using the initial solutions as inputs. Constraints were applied at each iteration in both the concentration and spectral directions, and the routine proceeded with improved lack of fit at each iteration until convergence—a change in LOF of less than 0.01%—was achieved. The sole constraint applied in the spectral direction was nonnegativity. In the concentration direction, both nonnegativity and a correlation constraint, as described in Section 2.3, were applied to the paracetamol concentration vector. Four samples were included in the calibration set and 20 samples included in the test set (Figure 1). Furthermore, an equality constraint for concentration was applied to the caffeine and microcrystalline samples in the calibration set.

The results for the models are presented in Table 1. RMSE in all results reported here refers to root-mean-square error of the concentrations resolved for samples that were not included in the calibration set. An apparent improved accuracy of concentration resolution for paracetamol was achieved with MLPCA-MCR-ALS compared to standard MCR-ALS, on the basis of RMSE, with both methods providing comparable performance to PLSR. Furthermore, a pure profile for each compound was resolved and there was strong agreement between the resolved pure and measured pure profiles (Figure 3). The low RMSE, high $R^2$ value and well-recovered pure profiles support the hypothesis that the signals of the mixtures from EDXRD have an underlying bilinear model and MCR-ALS with correlation and equality constraints were capable of obtaining accurate resolution for $\boldsymbol{C}$ and $\boldsymbol{S}$. The correlation constraint in this model had the effect of ensuring the correct scale of paracetamol concentrations. A PLSR model with two factors and with the same four calibration samples was assessed in order to compare the performance. A PLSR model with three factors was also assessed but did not present improved results and is therefore not used for comparison or reported. MLPCA-MCR-ALS appears to perform better than PLSR, on the basis of RMSE, for paracetamol quantification, furthering the claim by a previous study that first order calibration with MCR-ALS can achieve successful quantification with only very few samples in the calibration set.[20] The two other compounds, caffeine and microcrystalline cellulose, which had equality-constrained calibration samples, were also recovered with prediction errors and performance comparable to PLSR.

### 4.2.2 | Quantitative analysis of paracetamol in packaged samples

The same procedure as described for unpackaged samples in the previous section was applied to 18 binary and ternary samples of packaged tablets with three components modelled. The same four samples as for the unpackaged sample model were used as the calibration set for the correlation and equality constraints were applied (Figure 1).

Figure 4 shows the resolved pure profiles from the procedure along with the measured pure profiles, which were not included in the resolution procedure, for comparison. The resolved profiles are in strong agreement with the measured profiles. Once again, MLPCA-MCR-ALS outperforms standard MCR-ALS and both methods compare well in the performance to PLSR (Table 1).

### 4.2.3 | Simultaneous analysis of paracetamol in unpackaged and packaged samples

Unpackaged and packaged samples were modelled together in a multiset structure, forming a set of 42 samples of which 24 belonged to the unpackaged subset and 18 to the packaged subset. The packaged data had been transformed

**TABLE 1** Model results of PLSR, MCR-ALS and MLPCA-MCR-ALS with correlation constraint for concentrations of paracetamol

| | | Method | No. samples (*N*) | No. calib. Samples | Test samples | No. factors | Total LOF (%) | $R^2$ | RMSE (%w/w) |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Unpackaged** | | | | | | | | |
| 1 | Single set | PLSR | 24 | 4 | 20 | 3 | — | — | 3.55 |
| 2 | | MCR-ALS | 24 | 4 | 20 | 3 | 2.96 | 0.990 | 2.67 |
| 3 | | MLPCA-MCR-ALS | 24 | 4 | 20 | 3 | 3.63 | 0.993 | 2.23 |
| 4 | Multiset | MCR-ALS | 24 | 4 | 20 | 4 | 3.23 | 0.991 | 2.75 |
| 5 | | MLPCA-MCR-ALS | 24 | 4 | 20 | 4 | 3.69 | 0.993 | 2.27 |
| | **Packaged** | | | | | | | | |
| 6 | Single set | PLSR | 18 | 4 | 14 | 3 | — | — | 2.98 |
| 7 | | MCR-ALS | 18 | 4 | 14 | 3 | 2.16 | 0.988 | 3.08 |
| 8 | | MLPCA-MCR-ALS | 18 | 4 | 14 | 3 | 2.45 | 0.995 | 2.15 |
| 4 | Multiset | MCR-ALS | 18 | 4 | 14 | 4 | 3.23 | 0.988 | 3.78 |
| 5 | | MLPCA-MCR-ALS | 18 | 4 | 14 | 4 | 3.69 | 0.996 | 1.93 |

Abbreviations: LOF, lack-of-fit; MCR-ALS, multivariate curve resolution-alternating least squares; MLPCA-MCR-ALS, maximum likelihood principal component analysis multivariate curve resolution-alternating least squares; PLSR, partial least squares regression; RMSE, root-mean-square error.

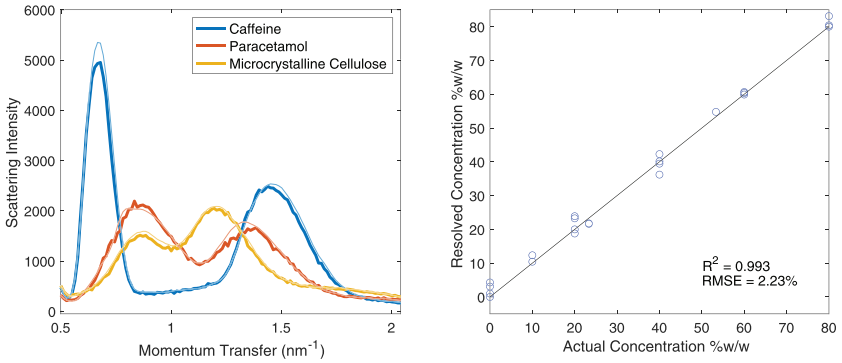


**FIGURE 3** Maximum likelihood principal component analysis multivariate curve resolution-alternating least squares (MLPCA-MCR-ALS) with correlation constraint for unpackaged samples. Resolved pure profiles in bold along with measured pure profiles (left). Resolved concentrations against reference concentrations for paracetamol (right). RMSE, root-mean-square error

according to their nominal thicknesses to account for the attenuation effect of the packaging. Local internal calibration models were formed for the paracetamol component of each subset using the correlation constraint with the same calibration samples as for the separately modelled subsets. The unpackaged subset was used as the reference subset for rescaling of the resolved profiles in order to preserve the underlying relationship between the data matrix and the pure momentum transfer profiles. Concentrations were equality constrained for caffeine and microcrystalline cellulose concentration vectors for those calibration samples in the unpackaged subset only. This is due to the difference in the scaling of concentrations between the two subsets required to preserve the approximation to bilinearity in the matrix factors.

Four components were used to reproduce the data using MLPCA, using the sample standard deviation matrix calculated from triplicate measurements for all data points of all sample profiles. SIMPLISMA was used to determine the three purest sample profiles, which were used as initial momentum transfer profile for the three pharmaceutical compounds. The recorded packaging profile was assigned to the fourth momentum transfer profile vector as an initial

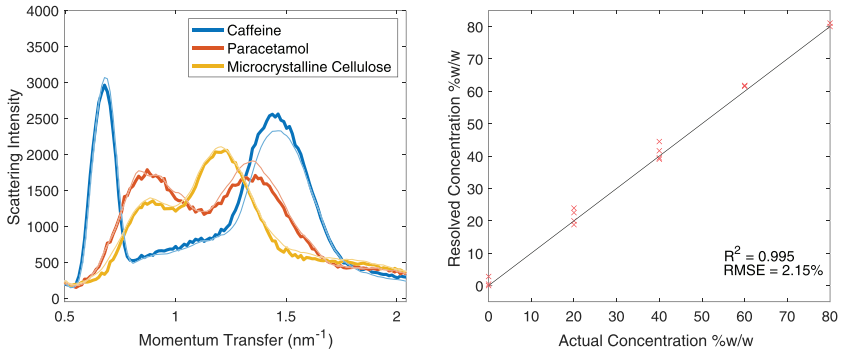


**FIGURE 4** Maximum likelihood principal component analysis multivariate curve resolution-alternating least squares (MLPCA-MCR-ALS) with correlation constraint for packaged samples. Resolved pure profiles in bold along with measured pure profiles (left). Resolved concentrations against reference concentrations for paracetamol (right). RMSE, root-mean-square error

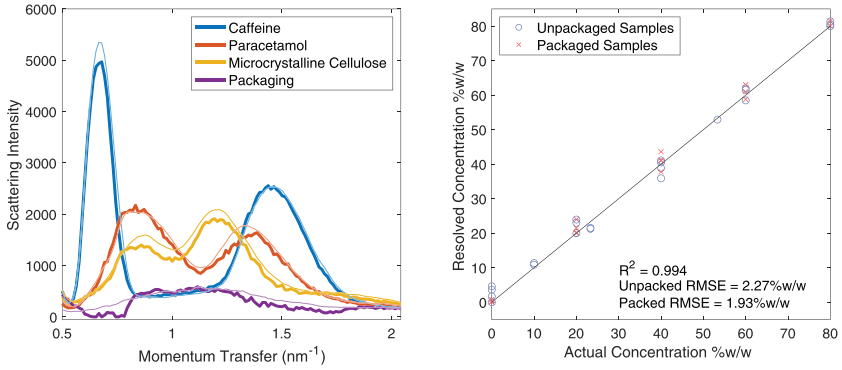


**FIGURE 5** Maximum likelihood principal component analysis multivariate curve resolution-alternating least squares (MLPCA-MCR-ALS) with correlation constraint for unpackaged and packaged samples modelled simultaneously. Resolved pure profiles in bold along with measured profiles (left). Resolved concentrations against reference concentrations for paracetamol (right). RMSE, root-mean-square error

estimate of the packaging profile. The paracetamol component was correlation-constrained using local correlation constraint models for both subsets. The resolved pure profiles for the MLPCA-MCR-ALS procedure are shown in Figure 5, including the resolved packaging component. It can be seen that the resolved profiles are in good agreement with the measured pure profiles for all three compounds as well as the packaging component. Furthermore, paracetamol concentration has been resolved to high accuracy when compared to the reference concentrations (Figure 5). For unpackaged samples and packaged samples, the RMSEs are found to be 2.27 %w/w and 1.93 %w/w, respectively. Both of these results appear to represent minor improvements over the PLSR method previously reported, which modelled the subsets separately, and the packaged subset results could be described as statistically significant, as discussed in the appendix. Furthermore, the current method recovers the pure momentum transfer profiles, improving interpretability of results. The accuracy of quantification is also comparable to the resolution of concentrations when applying MCR-ALS and MLPCA-MCR-ALS to the subsets separately. MLPCA-MCR-ALS appears to perform considerably better than MCR-ALS using the same constraints for multiset modelling for both unpackaged and packaged subsets, somewhat vindicating the use of the MLPCA reconstruction of the data matrix prior to MCR-ALS. In the case of the packaged subset the former method could be described as statistically significant, which is discussed in the appendix with results presented in Tables 3 and 4.

Figure 6 shows the resolved concentrations for these 42 samples across both sets. Of particular interest is the packaging concentration vector which has lower values for unpackaged samples (0 %w/w–0.15 %w/w) and higher values (0.8 %w/w–1 %w/w) for packaged samples, which is in reasonable agreement with the absence and presence of packaging in the first and second subset, respectively. The packaging concentration vector was constrained to a maximum value of unity.

**TABLE 3** The bias, standard deviation and RMSE values for modelling paracetamol concentration in unpackaged and packaged subsets using MCR-ALS (multiset), MLPCA-MCR-ALS (multiset) and PLSR

| Subset | Model | Bias (%w/w) | Standard deviation (%w/w) | RMSE (%w/w) |
|---|---|---|---|---|
| Unpackaged | PLSR | 2.41 | 2.68 | 3.55 |
| | MCR-ALS | 1.43 | 2.42 | 2.75 |
| | MLPCA-MCR-ALS | −0.77 | 2.19 | 2.27 |
| Packaged | PLSR | −0.70 | 3.01 | 2.98 |
| | MCR-ALS | −2.56 | 2.88 | 3.78 |
| | MLPCA-MCR-ALS | −1.04 | 1.70 | 1.93 |

Abbreviations: MCR-ALS, multivariate curve resolution-alternating least squares; MLPCA-MCR-ALS, maximum likelihood principal component analysis multivariate curve resolution-alternating least squares; PLSR, partial least squares regression; RMSE, root-mean-square error.

**TABLE 4** Tests for significant differences in bias and standard deviations between models for paracetamol concentrations comparing MLPCA-MCR-ALS (multiset) with MCR-ALS (multiset) and PLSR

| Set | Models | Metric | *t* | Degrees of freedom | *p* |
|---|---|---|---|---|---|
| Unpackaged | MCR-ALS versus MLPCA-MCR-ALS | Bias | 5.6159 | 19 | 0.00002* |
| | | Standard deviation | 0.6055 | 18 | 0.55240 |
| | MLPCA-MCR-ALS versus PLSR | Bias | −9.1235 | 19 | 0.00000* |
| | | Standard deviation | −1.4872 | 18 | 0.15430 |
| Packaged | MCR-ALS versus MLPCA-MCR-ALS | Bias | −2.5769 | 13 | 0.02300* |
| | | Standard deviation | 2.5096 | 12 | 0.02740* |
| | MLPCA-MCR-ALS versus PLSR | Bias | −0.4911 | 13 | 0.63160 |
| | | Standard deviation | −2.4464 | 12 | 0.03080* |

*Significant at 5% level.
Abbreviations: MCR-ALS, multivariate curve resolution-alternating least squares; MLPCA-MCR-ALS, maximum likelihood principal component analysis multivariate curve resolution-alternating least squares; PLSR, partial least squares regression.

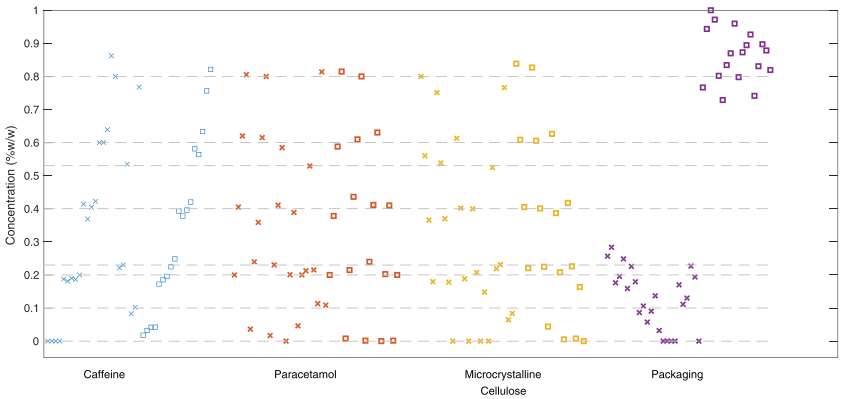


**FIGURE 6** Maximum likelihood principal component analysis multivariate curve resolution-alternating least squares (MLPCA-MCR-ALS) resolved concentrations for caffeine, paracetamol, micrcrystalline cellulose and packaging. Unpackaged samples (crosses) and packaged samples (squares) are shown in order of sample number for a given component (see Figure 1B). The concentration levels by experimental design are shown as grey horizontal lines

### 4.2.4 | Multiset MCR-ALS for prediction on a test set

A potentially more useful modelling and prediction scenario may be envisaged after pure reference profiles have been resolved from a data set of unpackaged and packaged subsets, such as that described in the previous sections. Upon

measuring a profile for a new packaged sample, $\boldsymbol{d}_{p,new}$, its paracetamol concentration can be estimated using a nonnegativity constrained classical least squares calculation:

$$\boldsymbol{c}_{p,new} = \boldsymbol{d}^{*}_{p,new}\left(\boldsymbol{S}^{T}\right)^{+} \tag{18}$$

where $\boldsymbol{d}_{p,new}$ has first been projected using MLPCA to form $\boldsymbol{d}^{*}_{p,new}$, again to account for the nonconstant error weights in the model. However, we must again consider that the concentration of paracetamol for packaged samples does not have the same scale as unpackaged profiles due to the matrix effect introduced by the packaging. Using the local calibration regression parameters, defined in Equation 13 and obtained from the final iteration of the MLPCA-MCR-ALS procedure, the concentrations of paracetamol in the new sample can be estimated according to the following rescaling:

$$\boldsymbol{c}^{*}_{p,new} = \left(\boldsymbol{c}_{p,new} - b_{p,0}\right)/b_{p,1} \tag{19}$$

To demonstrate this method, a leave-one out cross validation was carried out, in which one sample was not included in the MCR-ALS procedure and the paracetamol concentration of this sample was calculated according to Equations 18 and 19. The RMSECV for applying Equation 18 without the correction was 7.97 %w/w and with the correction was 3.42 %w/w (Figure 7). This simple method demonstrates the utility of MCR-ALS in the prediction of newly obtained samples using previously obtained pure profiles and calibration parameters.

## 4.3 | Quantitative analysis of caffeine and microcrystalline cellulose

Concentrations of caffeine and microcrystalline cellulose in the mixture set were also resolved using the same MCR-ALS models described for paracetamol, using both the separate unpackaged and packaged scenarios as well as the simultaneous modelling of subsets with the multiset structure. Equality constraints were applied to these two compounds for those samples in the calibration set for all MCR-ALS models. In the case of the multiset modelling, however, equality constraints were applied only to those samples in the unpackaged scenario, as seen previously. Results of resolved concentrations for these compounds using the various implementations are provided in Table 2 along with the results of PLSR modelling. It reveals that PLSR gives a higher performance than MLPCA-MCR-ALS for caffeine, which has less peak-overlap with the other compounds and higher signal-to-noise ratio. However, MLPCA-MCR-ALS appears to perform better than PLSR, on the basis of RMSE, for microcrystalline cellulose which suffers from lower signal-to-noise ratio and greater peak overlap.

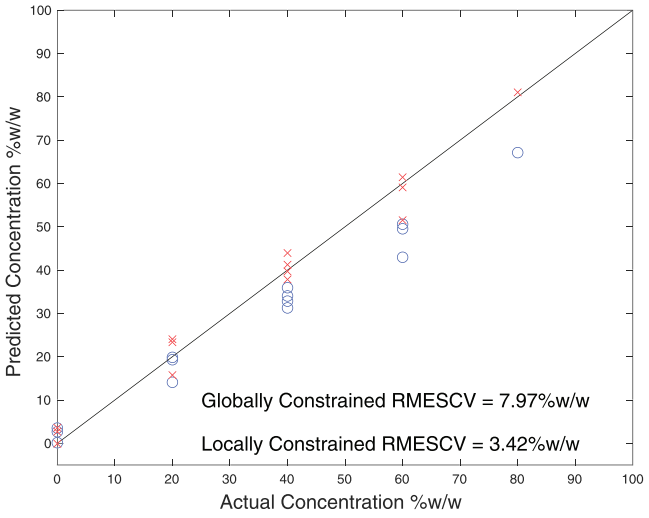


**FIGURE 7** Predicted paracetamol concentrations of test samples from multiset maximum likelihood principal component analysis multivariate curve resolution-alternating least squares (MLPCA-MCR-ALS) with correlation constraint using a leave-one-out cross validation with local calibration (red) and global calibration (blue)

**TABLE 2** Model results of PLSR, MCR-ALS and MLPCA-MCR-ALS for concentrations of caffeine and microcrystalline cellulose

| | | Method | No. samples (*N*) | No. calib. Samples | Test samples | Caffeine—RMSE (%w/w) | Mic. Cellulose—RMSE (%w/w) |
|---|---|---|---|---|---|---|---|
| **Model** | **Unpackaged** | | | | | | |
| 1 | Single set | PLSR | 24 | 4 | 20 | 1.76 | 3.77 |
| 2 | | MCR-ALS | 24 | 4 | 20 | 1.92 | 2.49 |
| 3 | | MLPCA-MCR-ALS | 24 | 4 | 20 | 2.16 | 2.16 |
| 4 | Multiset | MCR-ALS | 24 | 4 | 20 | 2.04 | 3.11 |
| 5 | | MLPCA-MCR-ALS | 24 | 4 | 20 | 2.16 | 2.86 |
| | **Packaged** | | | | | | |
| 6 | Single set | PLSR | 18 | 4 | 14 | 2.15 | 2.84 |
| 7 | | MCR-ALS | 18 | 4 | 14 | 2.30 | 2.24 |
| 8 | | MLPCA-MCR-ALS | 18 | 4 | 14 | 2.41 | 2.22 |
| 4 | Multiset | MCR-ALS | 18 | 4 | 14 | 5.24 | 8.00 |
| 5 | | MLPCA-MCR-ALS | 18 | 4 | 14 | 2.87 | 1.83 |

Abbreviations: MCR-ALS, multivariate curve resolution-alternating least squares; MLPCA-MCR-ALS, maximum likelihood principal component analysis multivariate curve resolution-alternating least squares; PLSR, partial least squares regression; RMSE, root-mean-square error.

Concentration resolution of caffeine performs less well for all scenarios when using MLPCA-MCR-ALS compared to standard MCR-ALS or PLSR. Hence, further investigation of the error structure estimated for MLPCA should be carried out. A high level of noise is observed in momentum transfer space variables corresponding to paracetamol due to the aforementioned preferred orientation effect. The authors propose the investigation of an error structure which accounts for both the preferred orientation phenomenon, and the variance in Poisson count data, the latter of which was investigated by Keenan et al.[21] However, in the present study, the resolution of the active ingredient paracetamol was prioritised.

A further interesting result was that MLPCA-MCR-ALS appeared to show a better performance than that of MCR-ALS, on the basis of RMSE, for the resolution of microcrystalline cellulose. The peaks of microcrystalline cellulose are located close to the peaks of paracetamol, and therefore a better fit to these variables may result from the application of MLPCA to account for the additional noise in the data points corresponding to paracetamol.

## 5 | CONCLUSIONS

This study demonstrates that EDXRD can be used with MCR-ALS as a methodology to accurately determine the concentrations of constituent compounds within packaged pharmaceutical formulations of simple mixtures. Unpackaged and packaged samples have been resolved simultaneously using a multiset structure. Using MLPCA as a pretreatment step and a local correlation constraint during the alternating least squares procedure, paracetamol concentrations were resolved for both subsets simultaneously to a greater degree of accuracy than PLSR achieves, even when PLSR models the subsets separately. Furthermore, MLPCA-MCRALS resolves both the other two compounds in the mixture set to a comparable accuracy to PLSR.

An advantage of EDXRD over other characterisation and quantification methods is the clear correspondence of physical phenomena with the measured features. The packaging material has been modelled freely as a fourth component to recover an accurate estimate of the reference scattering profile of the packaging materials. There is a potential practical application of this to the nondestructive testing of packaged pharmaceuticals, since unpackaged samples with known reference information can be modelled together with screened packaged samples to provide an interpretable

and accurate characterisation of the former. MLPCA-MCR-ALS provides a soft-modelling framework to model noisy and nonhomoscedastic noise found in EDXRD data across different packaging contexts simultaneously. Advantages in the approach taken in this study include the simultaneous resolution of all compounds in the ternary mixtures as well as its ability to accurately resolve the EDXRD profile of the packaging material while preserving the bilinearity of the data. A combination of parameterised data transformations made according to understood physical phenomena, and the use of soft modelling of the effects of packaging enable EDXRD and MCR-ALS to be a powerful and flexible approach to the resolution of the composition of pharmaceutical formulations in nondestructive testing.

## ACKNOWLEDGEMENTS

Peter Kenny acknowledges funding received from EPSRC Grant No. EP/M506448/1 and the support of Chiaki Crews who set up and calibrated the EDXRD system and was responsible for the design of the experiment, sample preparation, system setup and data acquisition. Chiaki Crews acknowledges funding received from EPSRC Grant No. EP/G037264/1 (Security Science Doctoral Training Centre).

## PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/CEM.3329.

## ORCID

*Peter S. Kenny* https://orcid.org/0000-0003-2992-5157
*Chiaki Crews* https://orcid.org/0000-0002-1439-3074
*Tom Fearn* https://orcid.org/0000-0003-2222-6601
*Robert D. Speller* https://orcid.org/0000-0001-7080-6622

## REFERENCES

1. Crews C, Kenny PS, O'Flynn D, Speller RD. Multivariate calibration of energy-dispersive X-ray diffraction data for predicting the composition of pharmaceutical tablets in packaging. *J Pharm Biomed Anal*. 2018;151:186-193.
2. Cook E, Fong R, Horrocks J, Wilkinson D, Speller R. Energy dispersive X-ray diffraction as a means to identify illicit materials: a preliminary optimisation study. *Appl Radiat Isot*. 2007;65(8):959-967.
3. Drakos I, Kenny P, Fearn T, Speller R. Multivariate analysis of energy dispersive X-ray diffraction data for the detection of illicit drugs in border control. *Crime Sci*. 2017;6(1):1-10.
4. Luggar RD, Horrocks JA, Speller RD, Royle GJ, Lacey RJ. Optimization of a low-angle X-ray scatter system for explosive detection. *Law Enforc Technol Identif Technol Traffic Saf*. 1995;2511:46-55.
5. Moss RM, Amin AS, Crews C, et al. Correlation of X-ray diffraction signatures of breast tissue and their histopathological classification. *Sci Rep*. 2017;7:1-9.
6. Soulez F, Crespy C, Kaftandjian V, Duvauchelle P. Diffraction peaks restoration and extraction in energy dispersive X-ray diffraction. *Nucl Instruments Methods Phys Res Sect a Accel Spectrometers, Detect Assoc Equip*. 2011;654:441-449.
7. Luggar RD, Horrocks JA, Speller RD, Lacey RJ. Determination of the geometric blurring of an energy dispersive X-ray diffraction (EDXRD) system and its use in the simulation of experimentally derived diffraction profiles. *Nucl Instruments Methods Phys Res Sect a Accel Spectrometers, Detect Assoc Equip*. 1996;383:610-618.
8. Cook EJ et al. Multivariate data analysis for drug identification using energy-dispersive X-ray diffraction. *IEEE Trans Nucl Sci*. 2009;56(3):1459-1464.
9. De Juan A, Jaumot J, Tauler R. Multivariate curve resolution (MCR). Solving the mixture analysis problem. *Anal Methods*. 2014;6:4964-4976.
10. de Juan A, Tauler R. Multivariate curve resolution (MCR) from 2000: progress in concepts and applications. *Crit Rev Anal Chem*. 2006;36(3-4):163-176.
11. Lyndgaard LB, Van den Berg F, De Juan A. Quantification of paracetamol through tablet blister packages by Raman spectroscopy and multivariate curve resolution-alternating least squares. *Chemom Intel Lab Syst*. 2013;125:58-66.
12. Wentzell PD, Andrews DT, Hamilton DC, Faber K, Kowalski BR. Maximum likelihood principal component analysis. *J Chemometr*. 1997;11(4):339-366.
13. Windig W, Guilment J. Interactive Self-Modeling Mixture Analysis. *Anal Chem*. 1991;63:1425-1432.
14. Abdollahi H, Tauler R. Uniqueness and rotation ambiguities in multivariate curve resolution methods. *Chemom Intel Lab Syst*. 2011;108(2):100-111.
15. Ahmadi G, Tauler R, Abdollahi H. Multivariate calibration of first-order data with the correlation constrained MCR-ALS method. *Chemom Intel Lab Syst*. 2015;142:143-150.
16. Wentzell PD, Karakach TK, Roy S, et al. Multivariate curve resolution of time course microarray data. *BMC Bioinformatics*. 2006;7:1-19.

17. Dadashi M, Abdollahi H, Tauler R. Application of maximum likelihood multivariate curve resolution to noisy data sets. *J Chemometr*. 2013;27(1-2):34-41.
18. Fearn T. The effect of spectral pre-treatments on interpretation. *NIR News*. 2009;20:15-16.
19. Berger MJ, Hubbell JH, Seltzer SM et al. NIST Standard Reference Database 8 (XGAM). https://dx.doi.org/10.18434/T48G6X
20. Debus B, Kirsanov DO, Panchuk VV, Semenov VG, Legin A. Three-point multivariate calibration models by correlation constrained MCR-ALS: a feasibility study for quantitative analysis of complex mixtures. *Talanta*. 2017;163:39-47.
21. Keenan MR. Multivariate analysis of count data. In: *Techniques and Applications of Hyperspectral Image Analysis*. Chichester: John Wiley & Sons Ltd; 2007.
22. Fearn T. Comparing standard deviations. *NIR News*. 1996;7(5):5-6.
23. Fearn T. Comparing standard deviations (continued). *NIR News*. 2009;20(7):24-25.
24. Snedecor GW, Cochran WG. *Statistical Methods*. Ames, Iowa: Iowa State University Press; 1967.

**How to cite this article:** Kenny PS, Crews C, Fearn T, Speller RD. Determination of ingredients in packaged pharmaceutical tablets by energy dispersive X-ray diffraction and maximum likelihood principal component analysis multivariate curve resolution-alternating least squares with correlation constraint. *Journal of Chemometrics*. 2021;e3329. https://doi.org/10.1002/cem.3329

## APPENDIX A: Significance tests for model comparison

As seen in Section 4.2.3, it would seem from inspection that MLPCA-MCR-ALS using local correlation constraint achieves lower RMSE values than PLSR on separately modelled subsets and MCR-ALS with local correlation constraint, for both the unpackaged and packaged subsets. It is more difficult, however, to establish statistical significance when comparing the RMSE values of concentration vectors derived from the same samples but from different models.

The RMSE can be decomposed into the bias and variance of the resulting concentration vector. Testing for statistically significant differences in bias and variance (or standard deviation) between models can help determine statistical significance in the difference in RMSE values.[22,23] It is worth noting that testing for statistical significance in the difference of standard deviations is complicated further when you are testing results from the same samples for different models as these concentrations are correlated by definition.[24]

The results of biases, standard deviations and RMSEs for multiset MLPCA-MCR-ALS, multiset MCR-ALS and PLSR for the unpackaged and packaged subsets are given in Table 3. Table 4 tests for significant differences in the biases and standard deviations when comparing MLPCA-MCR-ALS against MCR-ALS and PLSR.

In the case where both the bias and variance are statistically significantly different between two models then it is reasonable to conclude that the RMSEs are different, assuming they are both lower in one model compared to the other. This is the case for comparing the packaged results for MCR-ALS and MLPCA-MCR-ALS for paracetamol. We may therefore conclude that the MLPCA-MCR-ALS method performs better than the MCR-ALS method in predicting the paracetamol concentration of packaged samples.

In the case of comparing MLPCA-MCR-ALS against MCR-ALS for the unpackaged subsets for paracetamol, both the bias and the standard deviations are significantly different for the two models. A complication arises because the biases have different signs, therefore it is not possible to say that the square of the biases are significantly differently. Consequently, we cannot conclude that the RMSEs are statistically different. We can only say therefore that the RMSE for paracetamol for MLPCA-MCR-ALS appears lower than for MCR-ALS. The same problem arises for comparing the unpackaged results for paracetamol for MLP-MCR-ALS and PLSR.

Lastly, for the packaged concentration vectors for paracetamol for the models MLPCA-MCR-ALS and PLSR, the biases are not significantly different, but the standard deviations are significantly different. As we do not need to consider a difference in sign for biases as their difference is insignificant, we can tentatively conclude that the RMSE values for paracetamol for packaged samples are significantly lower for MLPCA-MCR-ALS than for PLSR.