APPLICATION OF ENSEMBLE LEARNING ICA TO INFRA-RED IMAGING

James W. Miskin and David J. C. MacKay

Cavendish Laboratory, Cambridge
CB3 0HE, UK
jwm1003@mrao.cam.ac.uk

ABSTRACT

We apply Independent Component Analysis methods to reduce a large data set of 2160 spectra. The data were obtained by scanning an Infra-Red (THz) laser beam across a cross section of a tooth. We show that a Bilinear Decomposition can represent the 2160 spectra by a reduced basis of around 10 latent spectra. We also show that by selectively enforcing positivity of the variables in the Bilinear Decomposition we obtain more meaningful images. The method allows identification of different regions of the tooth.

1. INTRODUCTION

The data were produced by using an Infra-Red (THz) laser beam to obtain the spectral response of a cross section of tooth to a pulse of laser light. The large data set consists of 2160 such spectra obtained by moving the beam over a grid of points on the tooth and each spectrum consists of 801 data points. Figure 1 shows a set of three of the spectra from the data set.

To analyse the data it is necessary to reduce it. One method is an EM clustering algorithm, [2]. Figure 2 shows the results obtained if we fit the observed spectra with a Gaussian mixture model consisting of 9 clusters. The figure shows the most probable cluster for each pixel in the tooth cross section and clearly shows that the clusters are localised into different regions of the tooth. But there are several problems. First there is the problem of class boundaries: the pixels at the boundaries are mixtures of clusters, but the algorithm uses a single cluster to model them. Second, detailed information is lost and replaced by a single latent variable, the cluster from which the pixel was assumed to be generated. Third, it is difficult to identify the correct number of clusters to use.

2. BILINEAR DECOMPOSITION

Instead of using a cluster model, we could assume that the spectra are linear combinations of independent latent spectra. For this model we consider the Bilinear Decomposition

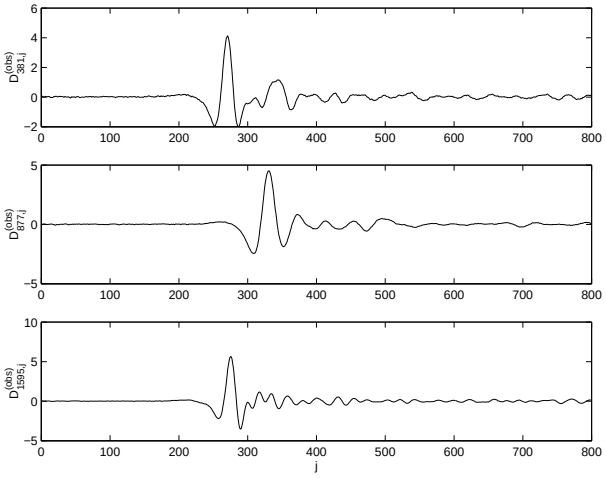

Figure 1: Sample spectra from the data set.

$$\mathbf{D} = \mathbf{A}^T\mathbf{B} \tag{1}$$

$\mathbf{D}$ is an $I \times J$ matrix whose rows are the observed spectra. $\mathbf{B}$ is a $K \times J$ matrix whose rows are the set of K latent spectra. $\mathbf{A}$ is a $K \times I$ weight matrix, each row of which contains the amount of each of the latent spectra in the observed spectra. If we were to constrain the columns of $\mathbf{A}$ such that each column contained one '1' with the remaining elements set to '0', we would have a clustering algorithm. Instead we will use a more general prior for the elements of the weight matrix, $\mathbf{A}$, which will allow mixtures of latent spectra. This model is the ICA model, [3], although we have the advantage that we have a large number of observed spectra and can assume $K \ll I$ and $K \ll J$.

Our prior for the weight matrix is a mixture of

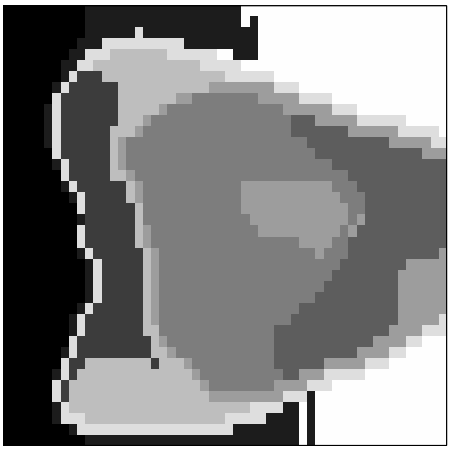

Figure 2: Cluster labels obtained by assuming that the spectra are drawn from a set of 9 clusters. The labels are obtained by finding the cluster with the maximum posterior probability for each spectrum. The figure shows that the clusters are grouped into localised regions, Enamel to left of image, Dentine to right of image, etc.

Gaussians.

$$P\left(A_{ki}\right)=\sum_{n=1}^{N_A}\pi_{kn}^{(A)}\mathcal{G}\left(A_{ki}\left|0,\beta_{kn}^{(A)}\right.\right) \tag{2}$$

where $\mathcal{G}\left(x\left|a,b\right.\right)$ is a Gaussian distribution over x with mean a and inverse variance b. If we choose $N_A=1$ then the model is equivalent to Variational Principal Components, [1]. We put a hyper-prior on the scale parameters

$$P\left(\beta_{kn}^{(A)}\right)=\Gamma\left(\beta_{kn}^{(A)}\left|a^{(A)},b^{(A)}\right.\right) \tag{3}$$

where

$$\Gamma\left(x\left|a,b\right.\right)=\frac{1}{\Gamma\left(b\right)}a^{b}x^{(b-1)}e^{-ax}. \tag{4}$$

If we choose the hyper-parameters to be much less than 1, $a^{(A)}=b^{(A)}=10^{-3}$, then we have an approximately scale invariant hyper-prior for the mixture of Gaussians. We use a Dirichlet prior for the mixture weights,

$$P\left(\left\{\pi_{kn}^{(A)}\right\}_{n=1}^{N_A}\right)=\mathcal{D}\left(\left\{\pi_{kn}^{(A)}\right\}_{n=1}^{N_A}\left|\left\{c_{n}^{(A)}\right\}_{n=1}^{N_A}\right.\right) \tag{5}$$

where

$$\mathcal{D}\left(\left\{x\right\}_i\left|\left\{c\right\}_i\right.\right)=\frac{\Gamma\left(\sum_n c_n\right)}{\prod_n\Gamma\left(c_n\right)}\delta\left(\sum_n x_n-1\right)\prod_n x_n^{c_n-1}. \tag{6}$$

If we choose $c_n^{(A)}=1$ then the prior is uniform.

Our prior for the latent spectra is a mixture of Gaussians of the form

$$P\left(B_{kj}\right)=\sum_{n=1}^{N_B}\pi_{kn}^{(B)}\mathcal{G}\left(B_{kj}\left|0,\beta_{kn}^{(B)}\right.\right) \tag{7}$$

with similar hyper-priors on the parameters.

We assume that the observed spectra are noisy and our noise model is Gaussian

$$P\left(D_{ij}\right)=\mathcal{G}\left(D_{ij}\left|\sum_{k=1}^{K}A_{ki}B_{kj},\gamma\right.\right) \tag{8}$$

and we use the conjugate hyper-prior

$$P\left(\gamma\right)=\Gamma\left(\gamma\left|a^{(\gamma)},b^{(\gamma)}\right.\right) \tag{9}$$

where we choose $a^{(\gamma)}=b^{(\gamma)}=10^{-3}$.

These priors are the conjugate priors for the posterior distributions, so they form a natural set of priors.

We wish to infer the set of parameters

$$\begin{aligned}\Theta &= \{\mathbf{A},\mathbf{B},\gamma,\\ &\quad\left\{\pi^{(A)}\right\},\left\{\beta^{(A)}\right\},\left\{\pi^{(B)}\right\},\left\{\beta^{(B)}\right\}\}\end{aligned} \tag{10}$$

given the observed data matrix, $\mathbf{D}$.

3. ENSEMBLE LEARNING

One solution is to find the MAP parameters. But this method can over-fit because it is sensitive to probability density rather than probability mass. The correct way to perform the inference would be to average over all possible parameter values by drawing samples from the posterior probability density.

Rather than performing a Markov Chain Monte Carlo (MCMC) approach to sample from the true posterior we use the Ensemble Learning approximation [4]. We approximate the posterior probability density, $P\left(\Theta\left|D\right.\right)$, by a more tractable distribution, $Q\left(\Theta\right)$, for which it is easy to perform inferences by integration rather than by sampling.

We optimise the approximate distribution by minimising the Kullback–Leibler divergence between the

approximate and true posterior distributions.

$$\begin{aligned} C_{\mathrm{KL}} &= \left\langle \log \frac{Q(\Theta)}{P(\Theta|D)} \right\rangle_Q - \log P(D) \\ &= \left\langle \log \frac{Q(\Theta)}{P(D|\Theta)P(\Theta)} \right\rangle_Q \\ &\geq -\log P(D) \end{aligned} \tag{11}$$

where $\langle . \rangle_Q$ denotes the expectation under the approximate distribution. Minimising C_{KL} with respect to $Q(\Theta)$ gives an upper bound on the negative evidence, $-\log P(D)$, for the model. If we choose a separable distribution for $Q(\Theta)$, C_{KL} will split into a sum of simpler terms.

Before minimising C_{KL} we make one further approximation. The priors for the elements of $\mathbf{A}$ and $\mathbf{B}$ are mixtures of Gaussians and so we derive the following bound using Jensen's inequality

$$\begin{aligned} \log P(A_{ki}) &= \log \sum_{n=1}^{N_A} \pi_{kn}^{(A)} \mathcal{G}\left(A_{ki} \middle| 0, \beta_{kn}^{(A)}\right) \\ &\geq \sum_{n=1}^{N_A} \lambda_{kin}^{(A)} \log \left(\frac{\pi_{kn}^{(A)} \mathcal{G}\left(A_{ki} \middle| 0, \beta_{kn}^{(A)}\right)}{\lambda_{kin}^{(A)}} \right) \end{aligned} \tag{12}$$

where $\left\{\lambda_{kin}^{(A)}\right\}$ are a set of weights to be found; they satisfy the constraints

$$\sum_{n=1}^{N_A} \lambda_{kin}^{(A)} = 1. \tag{13}$$

We minimise C_{KL} by performing a functional minimisation subject to the approximation that the approximate posterior is separable.

$$\begin{aligned} Q(\Theta) &= \prod_{ki} Q(A_{ki}) \times \prod_{kj} Q(B_{kj}) \times Q(\gamma) \\ &\quad \times Q\left(\left\{\pi^{(A)}\right\}, \left\{\beta^{(A)}\right\}, \left\{\pi^{(B)}\right\}, \left\{\beta^{(B)}\right\}\right). \end{aligned} \tag{14}$$

Optimising C_{KL} with respect to $Q(A_{ki})$, subject to the constraint that it is normalised, leads to

$$Q(A_{ki}) = \mathcal{G}\left(A_{ki} \middle| A_{ki}^{(1)}, A_{ki}^{(2)}\right) \tag{15}$$

where

$$\begin{aligned} A_{ki}^{(2)} &= \sum_{n=1}^{N_A} \lambda_{kin}^{(A)} \left\langle \beta_{kn}^{(A)} \right\rangle_Q + \sum_j \langle \gamma \rangle_Q \left\langle B_{kj}^2 \right\rangle_Q \\ A_{ki}^{(1)} A_{ki}^{(2)} &= \sum_j \left(D_{ij} - \sum_{k' \neq k} \langle A_{k'i} B_{k'j} \rangle_Q \right) \langle B_{kj} \rangle_Q \end{aligned} \tag{16}$$

The optimal approximate posterior distributions for parameters of the prior distributions are

$$\begin{aligned} Q\left(\left\{\pi_{kn}^{(A)}\right\}_{n=1}^{N_A}\right) &= \mathcal{D}\left(\left\{\pi_{kn}^{(A)}\right\}_{n=1}^{N_A} \middle| \left\{\bar{c}_{kn}^{(A)}\right\}_{n=1}^{N_A}\right) \\ Q\left(\beta_{kn}^{(A)}\right) &= \Gamma\left(\beta_{kn}^{(A)} \middle| \bar{a}_{kn}^{(A)}, \bar{b}_{kn}^{(A)}\right) \end{aligned} \tag{17}$$

where

$$\begin{aligned} \bar{a}_{kn}^{(A)} &= a^{(A)} + \frac{1}{2} \sum_i \lambda_{kin}^{(A)} \left\langle A_{ki}^2 \right\rangle_Q \\ \bar{b}_{kn}^{(A)} &= b^{(A)} + \frac{1}{2} \sum_i \lambda_{kin}^{(A)} \\ \bar{c}_{kn}^{(A)} &= c^{(A)} + \sum_i \lambda_{kin}^{(A)} \end{aligned} \tag{18}$$

The weights that were used to approximate C_{KL} can be found by using the update rule

$$\lambda_{kin}^{(A)} \propto \exp\left(\left\langle \log\left(\pi_{kn}^{(A)} \mathcal{G}\left(A_{ki} \middle| 0, \beta_{kn}^{(A)}\right)\right)\right\rangle_Q \right) \tag{19}$$

subject to constraint 13.

We can follow the same procedure to find the optimal distributions for $\mathbf{B}$.

The noise model can be optimised by

$$Q(\gamma) = \Gamma\left(\gamma \middle| \bar{a}_{kn}^{(\gamma)}, \bar{b}_{kn}^{(\gamma)}\right) \tag{20}$$

where

$$\begin{aligned} \bar{a}^{(\gamma)} &= a^{(\gamma)} + \frac{1}{2} \sum_{ij} \left\langle \left(D_{ij} - \sum_k A_{ki} B_{kj} \right)^2 \right\rangle_Q \\ \bar{b}^{(\gamma)} &= b^{(\gamma)} + \frac{IJ}{2}. \end{aligned} \tag{21}$$

The algorithm solves equations 16, 18, 19 and 21 iteratively until convergence is achieved.

Figure 3 shows the spectra weights ($\mathbf{A}$), figure 4 shows the latent spectra ($\mathbf{B}$) and figure 5 shows the reconstructed observed spectra when a basis of $K = 9$ latent spectra were used. The Bilinear Decomposition represents the data with a basis consisting of far fewer latent spectra than observed spectra. The spectra weights distinguish distinct regions of the tooth.

4. ENFORCING POSITIVITY

The previous section assumed that the elements of $\mathbf{A}$ and $\mathbf{B}$ had a mixture-of-Gaussians prior; the elements could be both positive or negative. The amounts of the latent spectra in each observed spectrum must be related to the thickness of the tooth at each point; we

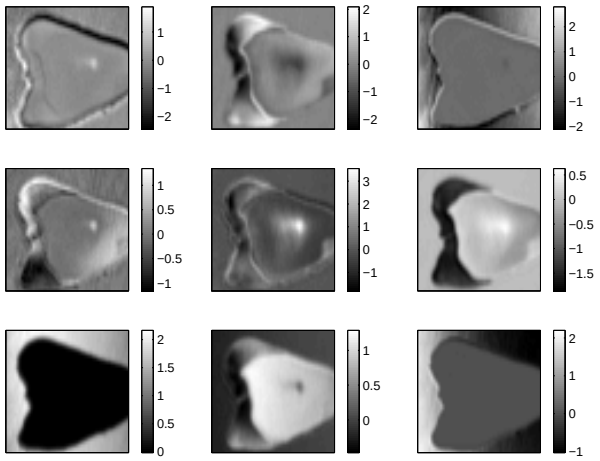

Figure 3: The rows of $\mathbf{A}$ rearranged to the image size. Each image shows the amount of one of the nine latent spectra in the spectrum for the corresponding pixel position. You can clearly see the Enamel (to left of tooth), Dentine (to right of tooth) and background (surrounding tooth). In addition a cavity is visible in the middle of the Enamel. The method could be used to generate images of sections of teeth.

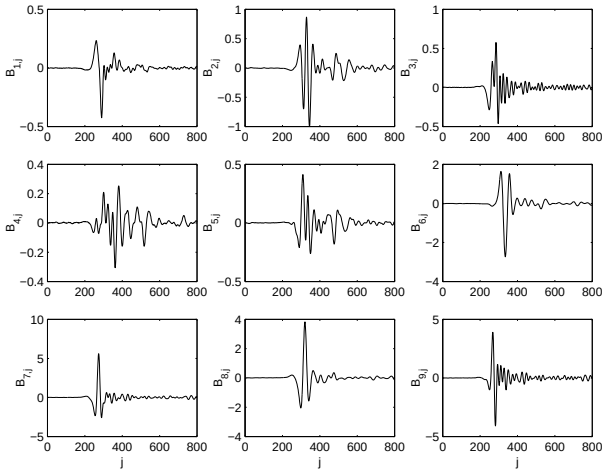

Figure 4: The set of latent spectra (rows of $\mathbf{B}$) learnt from the data.

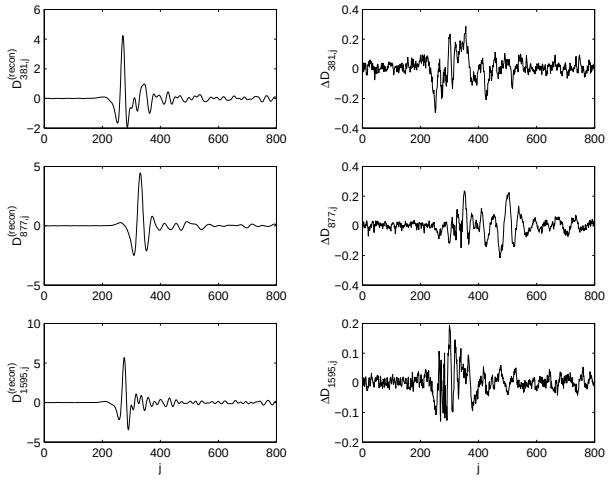

Figure 5: The plots on the left show the reconstructions of the observed spectra shown in figure 1 and the plots on the right show the errors in the reconstructions. We can see that the errors are small compared to the signal strength and so the reduced basis of nine spectra can be used to represent the much larger number of observed spectra.

shall assume that the weighting of each latent spectrum is linear in the thickness of the corresponding material. Negative weights are forbidden since they correspond to a negative thickness.

To force positive values of the weight matrix, we use a mixture of Laplacians prior

$$P\left(A_{ki}\right) = \begin{cases} \sum_{n=1}^{N_A} \pi_{kn}^{(A)} \beta_{kn}^{(A)} e^{-\beta_{kn}^{(A)} A_{ki}} & A_{ki} \geq 0 \\ 0 & A_{ki} < 0 \end{cases} \tag{22}$$

Following the derivation of the previous section, we obtain the following approximate posterior distributions

$$Q\left(A_{ki}\right) = \mathcal{G}^*\left(A_{ki} \middle| A_{ki}^{(1)}, A_{ki}^{(2)}\right) \tag{23}$$

where

$$\begin{aligned} A_{ki}^{(2)} &= \sum_j \langle \gamma \rangle_Q \left\langle B_{kj}^2 \right\rangle_Q \\ A_{ki}^{(1)} A_{ki}^{(2)} &= \sum_j \left(D_{ij} - \sum_{k' \neq k} \langle A_{k'i} B_{k'j} \rangle_Q \right) \langle B_{kj} \rangle_Q \\ &\quad - \sum_{n=1}^{N_A} \lambda_{kin}^{(A)} \left\langle \beta_{kn}^{(A)} \right\rangle_Q \end{aligned} \tag{24}$$

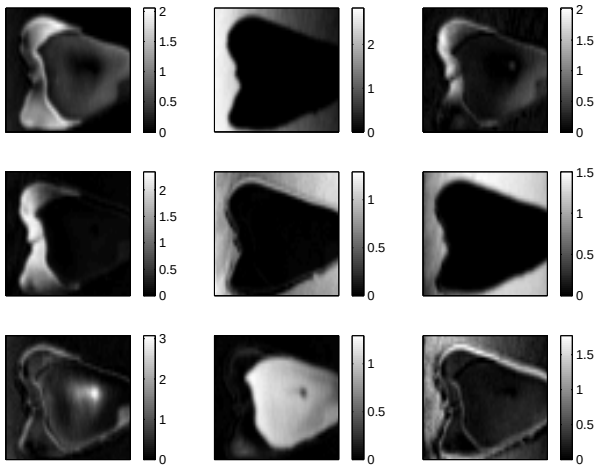

Figure 6: The elements of the weight matrix corresponding to each of the latent spectra. When compared to figure 3 we can see that there is much better separation of the tooth into distinct regions.

and

$$\mathcal{G}^*(x\,|a,b) \propto \begin{cases} e^{-\frac{b}{2}(x-a)^2} & x \geq 0 \\ 0 & x < 0 \end{cases} . \qquad (25)$$

We must also change the optimal posteriors for the prior parameters to obtain

$$\begin{aligned} \bar{a}_{kn}^{(A)} &= a^{(A)} + \sum_i \lambda_{kin}^{(A)} \langle A_{ki} \rangle_Q \\ \bar{b}_{kn}^{(A)} &= b^{(A)} + \sum_i \lambda_{kin}^{(A)} \\ \lambda_{kin}^{(A)} &\propto \exp\left(\left\langle \log\left(\pi_{kn}^{(A)} \beta_{kn}^{(A)} e^{-\beta_{kn}^{(A)} A_{ki}} \right) \right\rangle_Q \right) \end{aligned} \qquad (26)$$

Figure 6 shows the spectra weights ($\mathbf{A}$). The different regions of the tooth (Enamel, Dentine, etc) are separated. Figure 7 shows the latent spectra ($\mathbf{B}$) and figure 8 shows the reconstructed observed spectra. These results show that the observed spectra are still well represented while the regions of the tooth are now separated.

5. MODEL SELECTION

Equation 11 proves that C_{KL} is an upper bound on the negative evidence for a model. We can identify which model best represents the observed data if we assume that the bound is approximately equal to the negative evidence. Figure 9 shows how C_{KL} varies as a function of K for three different models; no positivity, positivity

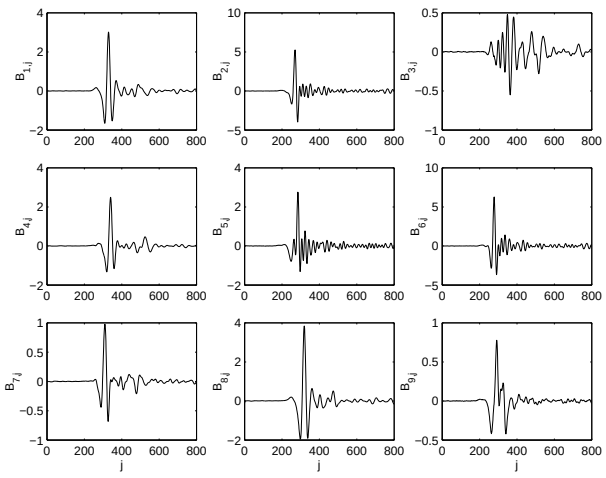

Figure 7: The latent spectra obtained when positivity of the weight matrix is enforced.

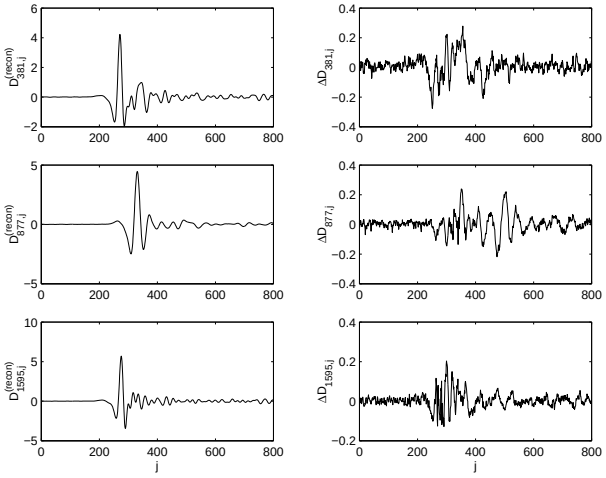

Figure 8: The plots on the left show the reconstructions of the observed spectra shown in figure 1 and the plots on the right show the errors in the reconstructions. These errors are comparable to those shown in figure 5.

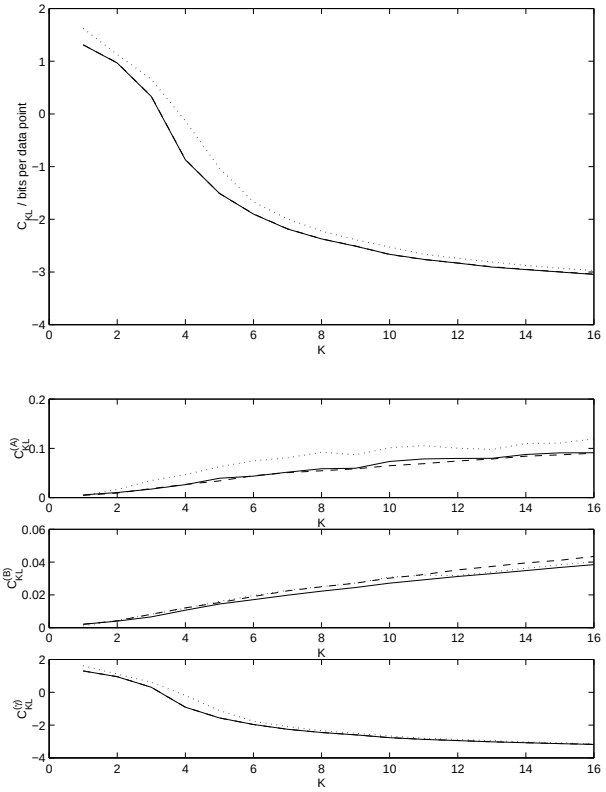

Figure 9: The top plot shows the variation of C_{KL} for three experiments, no positivity constraint (solid), enforcing $\mathbf{A}$ to be positive (dashed - under solid line) and enforcing $\mathbf{B}$ to be positive (dotted). C_{KL} is normalised to bits per data point (there are $I \times J$ data points). Enforcing $\mathbf{B}$ to be positive results in a worse model. There is no significant penalty for enforcing $\mathbf{A}$ to be positive. The bottom three plots show the variation of the component parts of C_{KL} for the three experiments.

of $\mathbf{A}$ and positivity of $\mathbf{B}$. The figure shows that as K is increased, C_{KL} decreases for all three models, since increasing K allows for a better fit to the observed spectra. The figure also shows that there is no significant penalty when $\mathbf{A}$ is forced to be positive whereas there is a penalty when $\mathbf{B}$ is forced to be positive.

We can consider C_{KL} to be a sum of three parts

$$\begin{aligned} C_{\mathrm{KL}}^{(A)} &= \left\langle \log \left(\frac{Q\left(\mathbf{A}, \left\{\pi^{(A)}\right\}, \left\{\beta^{(A)}\right\}\right)}{P\left(\mathbf{A}, \left\{\pi^{(A)}\right\}, \left\{\beta^{(A)}\right\}\right)} \right) \right\rangle_Q \\ C_{\mathrm{KL}}^{(B)} &= \left\langle \log \left(\frac{Q\left(\mathbf{B}, \left\{\pi^{(B)}\right\}, \left\{\beta^{(B)}\right\}\right)}{P\left(\mathbf{B}, \left\{\pi^{(B)}\right\}, \left\{\beta^{(B)}\right\}\right)} \right) \right\rangle_Q \\ C_{\mathrm{KL}}^{(\gamma)} &= \left\langle \log \left(\frac{Q(\gamma)}{P(D, \gamma | \mathbf{A}, \mathbf{B})} \right) \right\rangle_Q \end{aligned} \quad (27)$$

where the first two are the cost of encoding elements of $\mathbf{A}$ and $\mathbf{B}$ and the third is the cost associated with encoding the observed data. Figure 9 shows how these three cost functions vary as a function of K for the three models. We can see that the dominant term is $C_{\mathrm{KL}}^{(\gamma)}$ which drops as K increases. The other terms are much smaller, but increase approximately linearly as K increases.

6. CONCLUSION

An Ensemble Learning model can approximate the full posterior of a Bilinear Decomposition model by a more tractable separable distribution. The model can be applied to spectra and yield a reduced basis of spectra. By enforcing positivity of the weight matrix, one can obtain an intuitive separation into latent spectra which results in a separation of the data into localised regions.

The localised separation results in images corresponding to a clustering model but does not suffer from problems of classifying region boundaries and preserves detailed structure within the clusters.

7. ACKNOWLEDGEMENTS

We thank Toshiba Research Europe Ltd for the use of the THz imaging data.

8. REFERENCES

[1] Christopher M. Bishop. Variational principal components. In *Proceedings of the Ninth International Conference on Artifical Neural Networks*, pages 509–514, 1999.

[2] A. P. Dempster, Laird N. M., and Rubin D. B. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[3] Harri Lappalainen. Ensemble learning for independent component analysis. In *Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation*, 1998.

[4] David J. C. Mackay. Developments in probabilistic modelling with neural networks - ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks*, pages 191–198. Springer, 1995.