A Fully Automated, Faster Noise Rejection Approach Increasing the Analytical Capability of Chemical Imaging for Digital Histopathology

Soumyajit Gupta[a], **Shachi Mittal**[b], **Andre Kajdacsy-Balla**[c], **Rohit Bhargava**[b], **and Chandrajit Bajaj**[a,2]

[a]Computational Visualization Center, Department of Computer Science, University of Texas at Austin, Austin, TX 78705; [b]Department of Bioengineering, University of Illinois, Urbana Champaign, Urbana, IL 61801; [c]Department of Pathology, University of Illinois at Chicago, Chicago, IL 60612

This manuscript was compiled on January 19, 2018

High dimensional data, for example in chemical imaging, involves an inherent trade-off in the acquisition time and quality of data. Minimum Noise Fraction (MNF) developed by Green *et al.* (1) has been extensively studied as an algorithm for noise removal in HSI (HyperSpectral Image) data. However, there is a speed-accuracy trade-off in the process of manually deciding the relevant bands in the MNF space, which costs around a month's time for an entire TMA. We propose three approaches termed 'Fast MNF', 'Approx MNF' and 'Rand MNF' where the computational time of the algorithm is reduced, as well as the entire process of band selection is fully automated. This automated approach is shown to perform at the same level of reconstruction accuracy as MNF with large speedup factors, resulting in the same task to be accomplished in hours. The different approximations of the algorithm, show the reconstruction accuracy vs storage $(50\times)$ and runtime speed $(60\times)$ trade-off. We apply the approach for automating the denoising of different tissue histology samples, in which the accuracy of classification (differentiating between the different histologic and pathologic classes) strongly depends on the SNR (signal to noise ratio) of recovered data. Therefore, we also compare the effect of the proposed denoising algorithms on classification accuracy. Since denoising HSI data is done without any ground truth, we also use a metric that assesses the quality of denoising in the image domain between the noisy and denoised image in absence of ground truth.

Minimum Noise Fraction | De-noising | Computational Imaging

Chemical imaging is an emerging technology in which every pixel or voxel of an image contains hyperspectral data, often consisting of hundreds or thousands of data points. The spectrum at each pixel resolves the chemical components at that point and, thus, provides the molecular profile of the sample (2–4). Computer algorithms that can process the data to information useful for a particular problem often require a specific data quality, at that spectral resolution, that often determines scanning (signal averaging) time. In addition to the chemical signature of the data, another benefit of these technologies is that workflows can be automated with fully digital analysis of the data (5–7). For example, Fourier Transform Infrared (FT-IR) spectroscopic imaging is emerging as an automated alternative to human examination in studying disease development and progression by using statistical pattern recognition (8–13). For a practical protocol for tissue imaging, as demonstrated in at least one instance of tissue histopathology, the signal-to-noise ratio (SNR) of $4cm^{-1}$ resolution spectral data needs to be more than $1000:1$ (12). To achieve this SNR, especially for the emerging high definition IR imaging (14–16), extensive signal averaging is required. The need for signal averaging increases acquisition time (SNR $\sim \sqrt{t}$), in turn, increasing acquisition time (17) to the extent that clinical translation becomes impractical. Signal processing approaches to reduce noise has previously been suggested to mitigate this crippling increase in integration time by mathematical methods to utilize correlations in data to reduce noise but suffer from two major drawbacks. First, given the large size of the data, the mathematical operations require computer processing often comparable to the acquisition time itself (18). Second, such methods invariably try to separate data into informative and noisy components; subsequently, a manual selection step is required to identify the information-bearing components thus compromising the automation benefits of using spectroscopic imaging for tissue analysis (19).

One class of mathematical transform techniques for noise reduction utilize the property that noise is uncorrelated whereas spectra (signals) have a high degree of correlation. In a transform domain, hence, the signal becomes largely confined to a few eigenvalues whereas the noise is spread across all. Noise reduction can be achieved by retaining eigenvalue images that correspond to high signal content and computing the inverse transform. All the eigenvalue data contain signal and noise but the relative proportion of the signal to noise which forms a threshold criterion for inclusion of specific eigenimages in the inverse transform. Inclusion of too many will not allow for significant noise rejection, while inclusion of too few would result in loss of fine spectral features. Hence, identifying eigen-

Significance Statement

Hyperspectral Images (HSI) have important spectral features in specific combination of wavenumbers. Noise in these channels can easily overwhelm these relevant spectral features, leading to degradation in classification accuracy. Minimum Noise Fraction (MNF), a widely used algorithm for noise removal in HSI data, suffers from long processing time due to manual selection of bands and unnecessary computational and storage costs. We describe an approach to fully automate the process, that massively reduces the execution time $(30\times)$. We also provide three algorithmic variants by exploiting the geometric structure of MNF, with lower runtime $(2\times)$ and massively reduced storage space $(50\times)$. They scale linearly with increasing data size and have same classification accuracy as the classical MNF.

S.G. and C.B. designed research; S.G. and S.M. performed research implementation, A.K.B. provided pathological annotations. S.G. and C.B. contributed new analytic tools; S.G, S.M., R.B, and C.B. analyzed data; and S.G., S.M., R.B., and C.B. wrote the paper.

The authors declare no conflict of interest.

[2]To whom correspondence should be addressed. E-mail: bajaj @ cs.utexas.edu

values corresponding to high signal content is an important step in the noise reduction process.

One widely used algorithm was provided by Green *et al.*(1) that applies Minimum Noise Fraction (MNF) to order spectral components in terms of SNR in the transformed space. It assumes that the covariance matrix for the raw data Σ_Y and the noise Σ_δ can both be estimated. Similar to principal component analysis (PCA) that orders the components in terms of variance, after transformation in MNF space, the top components are chosen and filtered and rest are zeroed out. This reduced basis is then used for inverse transformation into the signal space. Noise Adjusted Principal Components (NAPC) transform by Lee *et al.* (20) is a reformulation of the MNF transform in terms of noise whitening process. While MNF and NAPC transforms are mathematically equivalent, the latter consists of a sequence of two principal component transforms: First to whiten the data (de-correlate noise from data); Second to perform eigen decomposition on the modified covariance matrix, to order the underlying data by SNR.

Our goals in this work are to address the major challenges in noise rejection using mathematical methods. Specifically, first, we aim to provide criteria for unsupervised band selection, thereby maintaining objectivity of the data analysis protocol and reducing analysis time within the data processing pipeline by dispensing with the need for manual intervention. Second, we aim to re-examine the mathematical formulation of the MNF approach to speed up the process of computation of the forward and inverse transform MNF vectors. Specifically, we examine two novel approaches: (a) use of a truncated singular value decomposition (SVD) variant that is computationally more efficient, and (b) the use of a randomized variant of the above that is also memory efficient and has a reconstruction accuracy-memory trade-off depending on the application. Finally, we seek to improve the signal processing methods to provide a higher confidence in the consistency and accuracy of the noise rejection pipeline. We propose a comparison between acquired and denoised biomedical images using a robust metric, thereby providing better denoising guarantees in terms of both root mean square error (RMSE) and structural similarity.

1. Methods

In practical situations, noiseless data d_j is recorded by instruments as a noisy signal estimate y_j, due to fluctuations in detector current, backgrounds and source/instrument factors, which is modeled as:

$$y_j = d_j + \delta_j \quad [1]$$

where $y_j \in \mathbb{R}^S$ is the actual data collected by the apparatus and $\delta_j \in \mathbb{R}^S$ is the noise in the same pixel. The goal is to estimate δ_j given y_j, so we can best estimate the true signal d_j such that the relevant features (peak positions, peak heights, relative peak spacing etc.) in the spectrum are preserved.

Given a HSI $Y_{orig} \in \mathbb{R}^{W\times H\times S}$, where W, H are the spatial dimensions (width and height respectively), it has been restructured into a $2D$ matrix $Y \in \mathbb{R}^{N\times S}$, where $N = (W \times H)$ is the number of pixels, S is the number of spectral channels. Each column $y_i, \forall i = [1 : S]$ represents the reshaped image for the i^{th} spectral band and each row $y_j, \forall j = [1 : N]$ represents the spectral signature for the j^{th} pixel. The i^{th} entry of the spectral vector $y_j \in \mathbb{R}^S$ of each pixel y_{ij} determines the absorbance value of the tissue at that wavenumber. Wavenumbers ν are defined as inverse of wavelength and have units cm^{-1}. The recorded value at each wavenumber has unit $absorbance/au$, where au is arbitrary unit.

Mathematical Background. Assuming additive noise only, raw data can be represented as $Y = D + \delta$, where $Y = \{Y_1, ..., Y_S\}$ and D and δ are the uncorrelated signal (actual spectral data with baseline included) and noise components of the raw data Y. $Cov\{Y\} = \Sigma_Y = \Sigma_D + \Sigma_\delta$, where Σ_D and Σ_δ are the covariance matrices of D and δ respectively. Noise Fraction (NF) for the i^{th} band is defined as the ratio of noise variance to the total variance for that band. Similarly Signal to Noise ratio (SNR) for the i^{th} band is defined as the ratio of signal variance to the noise variance for that band.

$$NF = Var\{\delta_i\}/Var\{Y_i\} \quad [2]$$

$$SNR = Var\{D_i\}/Var\{\delta_i\} \quad [3]$$

MNF is the set of linear transformation $(Y_{MNF})_i = Y\phi_i$, for $i = 1, ..., S$, such that the SNR for $(Y_{MNF})_i$ is maximum among all linear transformations orthogonal to $(Y_{MNF})_j$, for $j = i + 1, ..., S$. All the transformation vectors in MNF space follow $\phi_i^T \Sigma_Y \phi_i = 1, \forall i = [1 : S]$. Maximization of the noise fraction leads to a numbering of bands that gives decreasing image quality with increasing component number. The SNR for $(Y_{MNF})_i$ in MNF space can be formulated:

$$\frac{Var\{\phi_i^T D\}}{Var\{\phi_i^T \delta\}} = \frac{\phi_i^T \Sigma_D \phi_i}{\phi_i^T \Sigma_\delta \phi_i} = \frac{\phi_i^T \Sigma_Y \phi_i}{\phi_i^T \Sigma_\delta \phi_i} - 1 = \lambda_i - 1 \quad [4]$$

The Noise fraction itself can then be re-factored as follows:

$$\frac{Var\{\phi_i^T \delta(x)\}}{Var\{\phi_i^T Y(x)\}} = \frac{\phi_i^T \Sigma_\delta \phi_i}{\phi_i^T \Sigma_Y \phi_i} = \frac{1}{\lambda_i} \quad [5]$$

The vectors ϕ_i are thus the real, symmetric eigenvectors of the eigenvalue problem:

$$det\{\Sigma_Y \Sigma_\delta^{-1} - \lambda I\} = 0 \quad [6]$$

Hence ϕ_i are the eigenvectors of $\Sigma_Y \Sigma_\delta^{-1}$, and λ_i, eigenvalue corresponding to ϕ_i, equals to the noise fraction in $(Y_{MNF})_i$. Also, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_S$, so that the components show decreasing image quality. Thus SNR is given by $\lambda_i - 1$.

Geometric Interpretation. As the data D and noise δ are additive and uncorrelated, we can write:

$$\begin{aligned} Y &= D + \delta \\ \Sigma_Y &= \Sigma_D + \Sigma_\delta \end{aligned} \quad [7]$$

Let the spectral decomposition of Σ_δ be $\Sigma_\delta = E\Lambda_\delta E^T$. Rotating Σ_Y in Eq.7 with eigenvector matrix E and rescaling it using the inverse-square root of the noise singular values Λ_δ, results in new covariance matrix where the contribution of noise component has been turned into an identity matrix.

$$\begin{aligned} \Sigma_W &= (\Lambda_\delta^{-1/2})^T E^T \Sigma_Y E(\Lambda_\delta^{-1/2}) \\ &= (\Lambda_\delta^{-1/2})^T E^T \Sigma_D E(\Lambda_\delta^{-1/2}) + (\Lambda_\delta^{-1/2})^T E^T \Sigma_\delta E(\Lambda_\delta^{-1/2}) \\ &= \Sigma_{W(D)} + I_\delta \end{aligned} \quad [8]$$

Let the eigen decomposition of Σ_W be $\Sigma_W = G\Lambda_{MNF}G^T$. Rotating Σ_W in Eq.8 with eigenvector matrix G results in:

$$\begin{aligned} \Lambda_{MNF} &= G^T \Sigma_W G \\ &= G^T \Sigma_{W(D)} G + G^T I_\delta G \\ &= \Lambda_{W(D)} + I_\delta \end{aligned} \quad [9]$$

The series of transforms that we used to change the original data covariance matrix Σ_Y into Λ_{MNF} is given by the transformation vector $\Phi = E\Lambda_\delta^{-1/2}G$ which we call the MNF projection vectors and Λ_{MNF} estimates of SNR of the data.

Optimizing MNF. Expanding the entire MNF transform, we can infer the following:

$$\begin{aligned}
D &= Y * \Phi * R * \Phi^{-1} \\
&= Y * (E * \Lambda_\delta^{-1/2} * G) * R * (E * \Lambda_\delta^{-1/2} * G)^{-1} \\
&= Y * E * \Lambda_\delta^{-1/2} * G * R * G^{-1} * \Lambda_\delta^{1/2} * E^{-1} \\
&= Y * E * \Lambda_\delta^{-1/2} * G * R * G^{T} * \Lambda_\delta^{1/2} * E^{T} \\
&= Y * E * \Lambda_\delta^{-1/2} * G * R * R^{T} * G^{T} * \Lambda_\delta^{1/2} * E^{T} \\
&= Y * (E * \Lambda_\delta^{-1/2} * G * R) * (E * \Lambda_\delta^{1/2} * G * R)^{T} \\
&= Y * \hat{\Phi} * \tilde{\Phi}^{T} \\
\Rightarrow \hat{\Phi} &= E * \Lambda_\delta^{-1/2} * G * R \qquad \text{//forward MNF transform} \\
\Rightarrow \tilde{\Phi} &= E * \Lambda_\delta^{1/2} * G * R \qquad \text{//inverse MNF transform}
\end{aligned}$$

Since R is a block identity matrix $R = \begin{bmatrix} I_K & 0 \\ 0 & 0 \end{bmatrix}$, introducing an extra R^T term keeps the value of the expression unaltered as $R * R^T = R$. Writing the MNF transform in this way, we ensure that we skip the costly matrix inversions of the MNF vectors. Also $G * R$ is effectively choosing the top K eigenvalues of G. Therefore we can reduce the computation cost by finding the reduced rank-K SVD of Σ_Y. The Λ_δ matrix inversion can also be replaced by more efficient versions due to its diagonal structure.

Automatic Band Selection. The optimal value of K can be determined by inspecting the entries of $\Lambda_{MNF} = SNR + 1$ which is a diagonal matrix. The Rose criteria (21) states that an SNR of at least 5.0 is needed to be able to distinguish image features at 100% certainty. We select the top K bands in the MNF space for which $SNR = \Lambda_{MNF} - 1 \geq 5.0$. Automating this process is the main computational speed factor that brings down the processing time from days down to few hours.

Fast MNF. By exploiting the MNF formulation, we can avoid all inverse operations and replace them with transpose, thereby making computations faster. Owing to the symmetric structure of covariance matrices, we also compute the singular value decomposition using eigen decomposition which is faster. Also, the transformation matrices are of size $(S \times K)$ instead of $(S \times S)$ where $K \ll S$. This is the main factor responsible for the algorithmic speedup.

Approx MNF. Since $K \ll S$, it is inefficient to compute the full spectral decomposition of the covariance matrix. Empirically, it was observed over different datasets, that the optimal value of the automatically selected K is $2-3\%$ of the total number of bands S. Hence we compute only a rank $\hat{K}$ truncated SVD of the whitened covariance matrix. This results in reduced computation time as well as memory. The standard solutions to truncated SVD include the power iteration algorithm and the Krylov subspace methods. Since power iteration is unstable at times due to the structure of the singular values, we use a version of the Block Lanczos method (22). We set $\hat{K} = 0.03 \times S$ and compute the rank $\hat{K}$-SVD, then let the band selection criteria to decide the optimal K.

Rand MNF. Although the block Lanczos algorithm can attain machine precision, it inevitably goes many passes through Σ_W, and it is thus slow when Σ_W is large or does not fit in memory. To circumvent this scenario, we use a faster randomized and memory efficient version which computes the $\hat{K}$-SVD of Σ_W up to $1+\epsilon$ Frobenius norm relative error (23).

Error Metric. In the absence of ground truth images of denosied data, we use a non-reference image quality metric this is simple and easy to use. The Method Noise Image (MNI) (24) metric aims at maximizing the structure similarity between the input noisy image and the estimated image noise around homogeneous regions and the structure similarity between the input noisy image and the denoised image around highly-structured regions, and is computed as the linear correlation coefficient of the two corresponding structure similarity maps.

2. Materials

Sample Preparation and Data Collection. A paraffin embedded breast tissue microarray ($BR1003$) consisting of 101 cores were obtained from US Biomax, Inc. The unstained sections of the TMA were placed on a BaF_2 salt plate for IR imaging. The sections were deparaffinized using a $24h$ hexane bath. High Definition data was acquired using the Agilent Stingray imaging system with 0.62 numerical aperture and a 128×128 focal plane array. A spectral resolution of $4cm^{-1}$ along with a pixel size of $1.1\mu m$ was obtained at the sample plane. The final FTIR sample has a spatial dimension of 11620×11620 pixels and spectral dimension of 1506 channels.

Classification. A random forest classifier was used to differentiate between the different histologic classes of a tissue sample. Labeled pixels for each class were obtained by the cases annotated by a pathologist. In this study, we have used a four class model separating benign epithelium from malignant epithelium. Finally, to assess the performance of the classifier, sensitivity and specificity is calculated for all the classes to generate the receiver operating characteristic curve. The area under this curve signifies the diagnostic potential of the model.

3. Results and Discussion

Setup. For all the experimentation, a standalone machine with Intel Xeon $E5-1660$@3.20GHz CPU and 64GB of RAM was used. Software for the simulations, results and plots include Matlab and ENVI.

Algorithm	Time	Space
1. MNF	$\mathcal{O}(S^3 + NS^2)$	$\mathcal{O}(NS + S^2)$
2. Fast MNF	$\mathcal{O}(S^3 + NSK)$	$\mathcal{O}(NK + S^2)$
3. Approx MNF	$\mathcal{O}(S^2K + NSK)$	$\mathcal{O}(NK + SK)$
4. Rand MNF	$\mathcal{O}(nnz(\Sigma_W)K + NSK)$	$\mathcal{O}(NK + SK)$

Table 1. Comparison of Time and Space complexity of MNF versions. $\mathcal{O}$: big-O complexity. $nnz(X)$: #non-zero elements in X. N: # pixels, S: # spectral bands and K: # chosen bands.

Complexity. Table 1 shows the algorithmic time and space complexity in terms of memory usage for the different MNF versions. The best algorithm in terms of time-space-accuracy

is Approx MNF. This is because it computes the best rank-K SVD with little loss in its approximation or memory usage. Depending on how much one wants a memory-accuracy trade-off, one may choose to switch to use Rand MNF, as it is a randomized version with approximation error guided by its parameters. For a typical FTIR spectrum with $S \sim 1500$ bands, the algorithm estimated the optimal number of bands to be $K \sim 30$, resulting in a efficiency factor of $50\times$.

Fig. 1. Runtime comparison of different implementations of MNF. The figure illustrates how time scales with varying size of the input data for proposed versions and optimizations of the standard MNF. The size is varied from $10K$ pixels (100×100) to $1000K$ pixels (1000×1000). Compared to the standard MNF, all the other versions have a speedup factor of $\sim 1.8 - 2.0$. This speedup is obtained by utilizing the fact that we do not need all the forward MNF vectors, but only the top K ones that are digitally calculated, avoiding manual selection. Also, a set of optimized matrix operations have been implemented, for instance replacing all inverse operations with transpose for improved computational performance. Due to BLAS optimized modules in Matlab, matrix multiplications are efficiently distributed across the cores automatically.

Timing Analysis. Standard MNF is worst in terms of time complexity as it computes the full eigen decomposition of Σ_W ($\mathcal{O}(S^3)$) and uses all the S eigenvectors to transform the input data into MNF space ($\mathcal{O}(NS^2)$). In comparison, Fast MNF still computes the full decomposition of Σ_W ($\mathcal{O}(S^3)$), but uses the Band selection criteria to retain only the top K eigenvalues. Hence only the top K eigenvectors are used to transform the data into MNF space ($\mathcal{O}(NSK)$). For Approx MNF, we only compute a truncated rank $\hat{K}$-SVD of Σ_W, since it was empirically observed that the top K chosen bands lie within $\hat{K} = 2 - 3\%$ of S, thereby highly reducing the computational cost ($\mathcal{O}(S^2K)$). Again, only the top K eigenvectors are used to transform the data into MNF space ($\mathcal{O}(NSK)$). Rand MNF uses a randomized algorithm to compute a truncated rank $\hat{K}$-SVD of Σ_W, hence it is much more efficient in terms of speed ($\mathcal{O}(nnz(\Sigma_W))K$), as the values in Σ_W will only be non-zero for channels which are correlated in terms of signal. Only the top K eigenvectors are used to transform the data into MNF space ($\mathcal{O}(NSK)$). Refer to Fig.1 for runtime. In all the three variants with $S \sim 1500$ and $K \sim 30$, the algorithmic scaling of SK instead of S^2 reduces runtime by $\sim 50\times$. For the BR1003 data, the entire MNF denoising process was reduced from ~ 1 month to ~ 6 hours. This clearly shows that the larger the data size, the better is the scaling of the proposed algorithms, thereby massively reducing the computational time of denoising for large TMA and associated datasets.

Note: Since the structure of the noise is unknown, we do not instinctively perform the rank-K approximation of Σ_δ, because depending on experiment and instrument there is no guarantee on the strength of noise present in the raw data, hence any prior assumptions cannot be made on the singular values of the noise. If we have some knowledge about the noise, then same reduced SVD approximations can be made while whitening the data. This would further reduce the computational time of the process ($\sim 3\times$ extra speedup).

Space Analysis. Standard MNF computes the full SVD decomposition ($\mathcal{O}(S^2)$). The transformed data in MNF space contains all S bands ($\mathcal{O}(NS)$). Fast MNF again does the full SVD decomposition ($\mathcal{O}(S^2)$). However the data in transformed domain only contain K bands ($\mathcal{O}(NK)$). Both Approx MNF and Rand MNF compute the truncated rank K decomposition, hence there are K eigenvectors each of S dimension ($\mathcal{O}(SK)$). The data in the transformed domain contains only K bands ($\mathcal{O}(NK)$). For the FTIR data with $S \sim 1500$ bands and $K \sim 30$, we achieve a RAM space saving of $\sim 50\times$, allowing us to process more data simultaneously in one go.

Denoising Profiles. The improvements offered by the different versions of the MNF presented in this study are illustrated in Fig. 2.The extent of denoising both in the spectral and spatial domain is approximately the same for all the different MNF algorithms. Fig. 2.A. depicts the spatial detail offered by different MNF versions with zoomed in sections in Fig. 2.C. and Fig. 2.B. shows the horizontal signal profile across the sample. Next, spectral profiles are compared across the different algorithms with reference spectrum (without MNF) to illustrate the extent of noise removal in each case (Fig. 2.D.). It can be seen that even with a speed up factor of $\sim 1.8 - 2.0$ there is no significant reduction in the spectral and spatial image quality.

Error Metric. Along with evaluating the performance of the presented MNF versions by examining the tissue profile, we utilize the MNI (method noise image) metric (24) aiming to maximize the structural similarity between the input noisy image and the denoised image around highly-structured regions, in the absence of ground truth. Fig. 3 shows the metric values for a core, over all the 1506 bands. A lower value of MNI indicate better denoising and structure preservation. This is evident from the fingerprint region ($900 - 1800 cm^{-1}$) which has very low values of MNI, while it is higher for the IR silent region ($1800 - 2700 cm^{-1}$). Since Standard MNF, Fast MNF and Approx MNF are almost the same in terms of reconstruction accuracy, all the plots from those three methods are same. The Rand MNF method which has a space-accuracy trade-off, produces similar results but with slight variations in most bands (notice the wavering in the plot).

Visualizing MNF bands. Fig. 4 visualizes the top K eigenimages (where K is automatically determined by the selection criteria) for two different patient cases (core1 and core 2). The bands in MNF space are arranged in decreasing order of SNR (increasing order of noise fraction), resulting in decreasing image quality with increase in band number. This suggests that

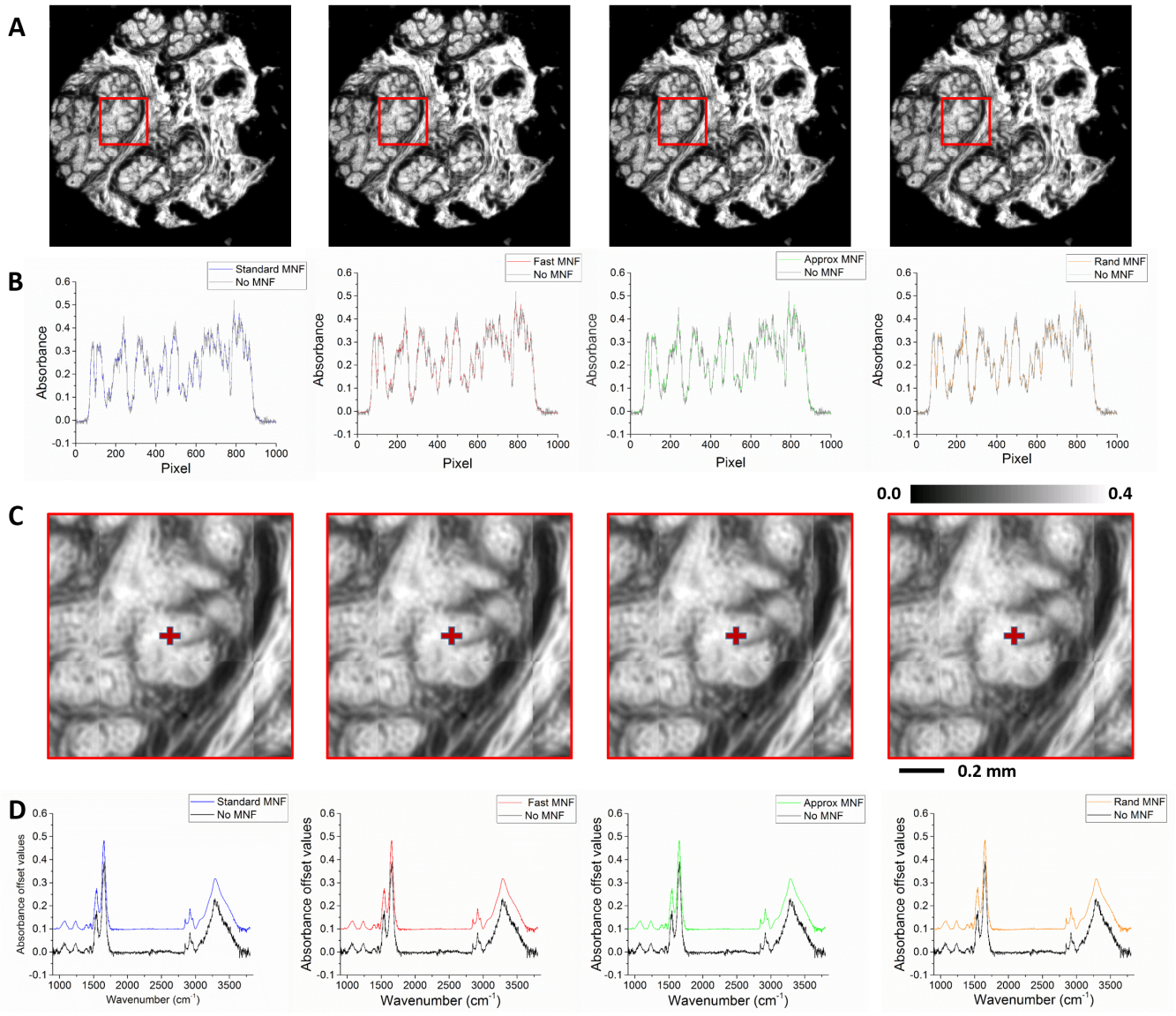

Fig. 2. Spectral and SNR comparisons: A. IR image of a patient case at amide 1 band. B: Intensity profile along the horizontal line in the cases shown in A spanning the entire core for both the noisy and the denoised (in gray) version. C. Zoomed in view of the area marked with a red box in top row. D. Comparison of the spectral profile of the noisy and the MNF version at the pixel marked in red in C.

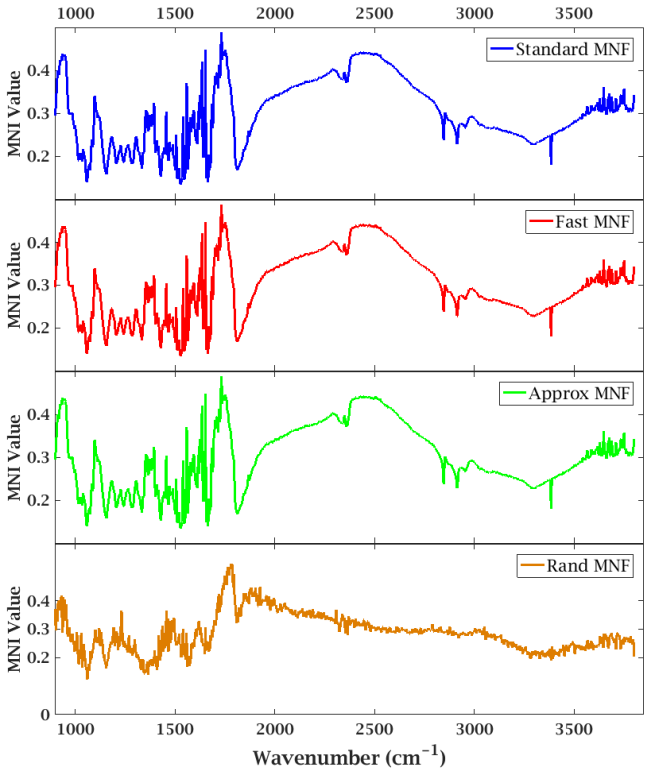

Fig. 3. MNI metric plot for a breast tissue biopsy. Lower values of MNI indicate better denoising and perseverance of structure. Across all the four variations we notice low values of MNI in the fingerprint region $(900 - 1800cm^{-1})$, which is mainly responsible for tissue classification. This shows that even in the presence of noise, the subtle structural features of the tissues are preserved after denoising. High MNI values in the IR silent region $(1800 - 2700cm^{-1})$ are expected as they mainly contain noise and no relevant signal components. The functional region $(2700 - 3600cm^{-1})$ has low MNI again followed by high MNI values in the water vapor absorption bands$(3600 - 3800cm^{-1})$.

the eigenimages in MNF space have decreasing image quality in terms of both the noise and structural detail. So,a few top bands in the MNF space should be able to capture most of information represented in the data along with denoizing. This concept can further be utilized to develop automated selection criteria for the number of bands to be kept after the transform. This can help eliminate user based subjectivity, make the process faster and easier to implement.

Impact on Tissue Classification. Furthermore, we studied the effect of MNF based data processing on tissue classification models. In particular we have investigated the performance of different MNF algorithms against raw data for distinguishing cancer from benign breast cells. It can be seen in Fig. 5, that for all the MNF versions, the Area Under Curve (AUC) value of the malignant epithelium (cancer) class is the same and there is a 10% drop in accuracy without the use of MNF. This suggests that for the development of highly accurate and efficient diagnostic models with HSI data, MNF gives better performance. Also,the MNF techniques presented in this paper (Fast MNF, Rand MNF and Approx MNF), offers the same performance as compared to standard MNF with a factor of 60× reduction in the processing time and 50× reduction in memory space required for computation.

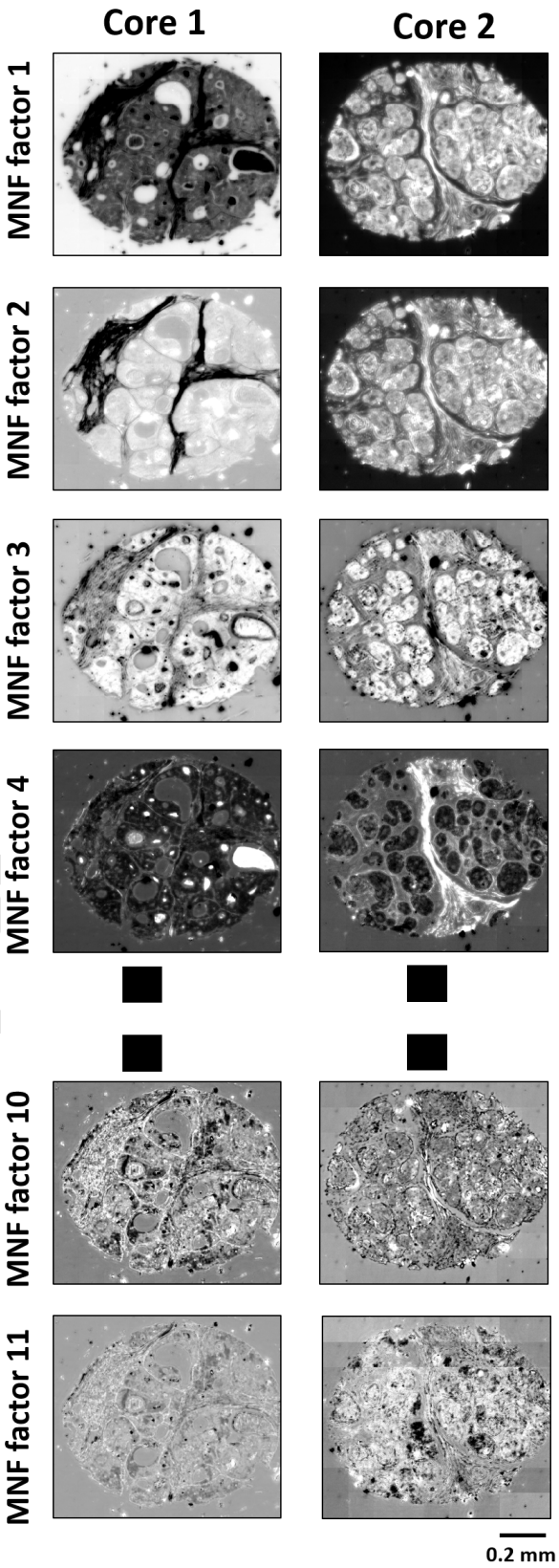

Fig. 4. Eigenimages of the top K bands in the MNF space for two given TMA cores. There is an evident decrease in structural features and SNR with increase in band number. Manually inspecting these eigenimages or defining some measurement metric (Reddy *et al.*(19)) on them, increases the processing time and computation cost for MNF. Our approach automatically determines the optimal value of K from the MNF eigenvalues in a computationally efficient manner.

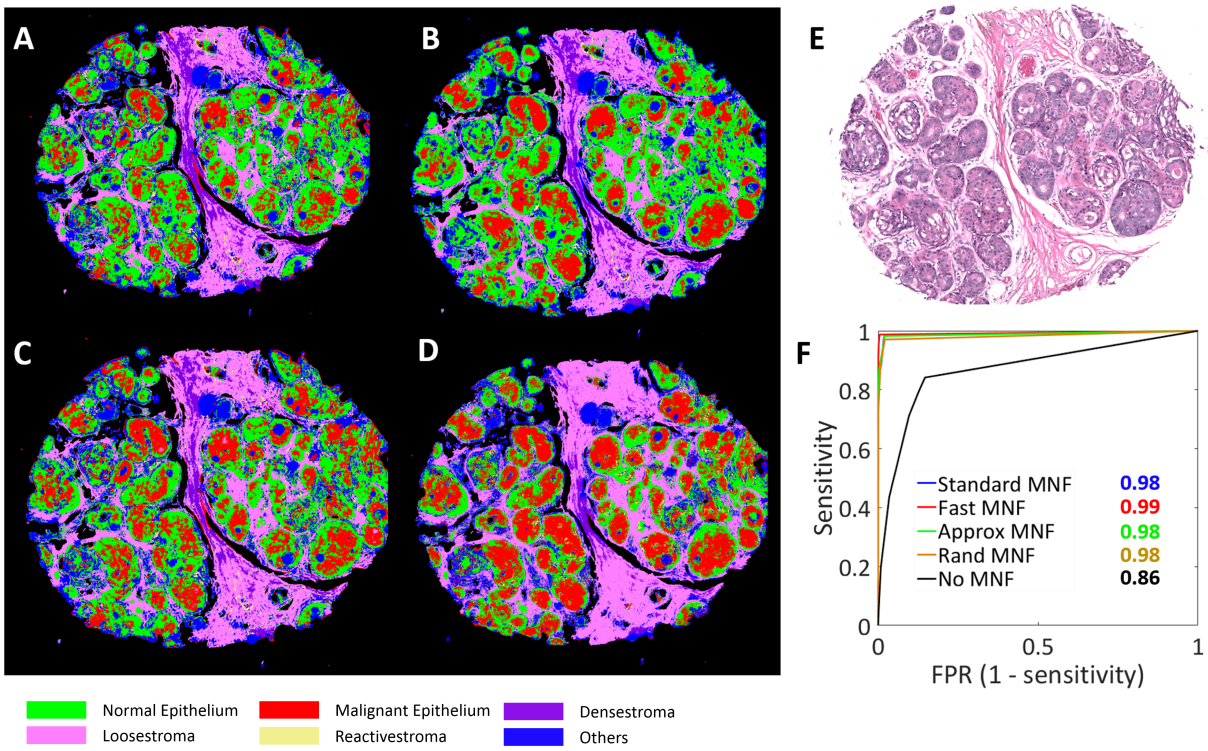

Fig. 5. Effect of MNF variants on the classification accuracy of a breast tissue biopsy. A-D. Classified images after standard MNF, fast MNF, Approx MNF and Rand MNF in a clockwise order. E. H&E (Hematoxylin and Eosin) stained image of an adjacent slice of the same tissue. F. Receiver Operating Characteristic (ROC) curve with area under the curve (AUC) values for the malignant epithelium class.

ACKNOWLEDGMENTS. This research supported in part by NIH grants R01GM117594, R41GM116300 and Dell-Seton 201602388.

4. Conclusion

In this paper, We demonstrate how to automate the band selection process in the MNF space, which drastically reduces the workflow duration of MNF denoising of TMAs from almost a month down to a matter of hours. We introduced three different optimizations of the MNF algorithm depending on the speed-memory-accuracy trade-off, resulting in a 2× runtime improvement and 50× memory efficiency. A well established error metric is also used which helps us decide the quality of denoising, in the absence of ground truth images. Similar classification performance of the suggested approaches as compared to conventional techniques suggesting the potential of the developed methods for computationally efficient analysis of big datasets for diagnostic applications. As a future work, we would like to make better approximations of the noise model itself, so that we can apply approximations for the eigen decomposition of the noise covariance matrix, hence further reducing the computational time of the process.

1. Green AA, Berman M, Switzer P, Craig MD (1988) A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *Geoscience and Remote Sensing, IEEE Transactions on* 26(1):65–74.
2. Bhargava R (2012) Infrared spectroscopic imaging: the next generation. *Applied spectroscopy* 66(10):1091–1120.
3. Elkins KM (2011) Rapid presumptive "fingerprinting" of body fluids and materials by atr ft-ir spectroscopy. *Journal of forensic sciences* 56(6):1580–1587.
4. Diem M, et al. (2013) Molecular pathology via ir and raman spectral imaging. *Journal of biophotonics* 6(11-12):855–886.
5. Mayerich D, Walsh M, Schulmerich M, Bhargava R (2013) Real-time interactive data mining for chemical imaging information: application to automated histopathology. *BMC bioinformatics* 14(1):156.
6. Pilling M, Gardner P (2016) Fundamental developments in infrared spectroscopic imaging for biomedical applications. *Chemical Society reviews* 45(7):1935–1957.
7. Lasch P, Diem M, Hänsch W, Naumann D (2006) Artificial neural networks as supervised techniques for ft-ir microspectroscopic imaging. *Journal of chemometrics* 20(5):209–220.
8. Fernandez DC, Bhargava R, Hewitt SM, Levin IW (2005) Infrared spectroscopic imaging for histopathologic recognition. *Nature biotechnology* 23(4):469.
9. Krafft C, et al. (2006) Identification of primary tumors of brain metastases by simca classification of ir spectroscopic images. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1758(7):883–891.
10. Bassan P, et al. (2014) Automated high-throughput assessment of prostate biopsy tissue using infrared spectroscopic chemical imaging in *Proc SPIE*. Vol. 9041, pp. 90410D–90416D.
11. Fabian H, et al. (2006) Diagnosing benign and malignant lesions in breast tissue sections by using ir-microspectroscopy. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1758(7):874–882.
12. Bhargava R (2007) Towards a practical fourier transform infrared chemical imaging protocol for cancer histopathology. *Analytical and bioanalytical chemistry* 389(4):1155–1169.
13. Pounder FN, Reddy RK, Bhargava R (2016) Development of a practical spatial-spectral analysis protocol for breast histopathology using fourier transform infrared spectroscopic imaging. *Faraday discussions* 187:43–68.
14. Leslie LS, et al. (2015) High definition infrared spectroscopic imaging for lymph node histopathology. *PloS one* 10(6):e0127238.
15. Reddy RK, Walsh MJ, Schulmerich MV, Carney PS, Bhargava R (2013) High-definition infrared spectroscopic imaging. *Applied spectroscopy* 67(1):93–105.
16. Sreedhar H, et al. (2015) High-definition fourier transform infrared (ft-ir) spectroscopic imaging of human tissue sections towards improving pathology. *Journal of visualized experiments: JoVE* (95).
17. Pilling MJ, et al. (2016) High-throughput quantum cascade laser (qcl) spectral histopathology: a practical approach towards clinical translation. *Faraday discussions* 187:135–154.
18. Anastasio MA, La Riviere P (2012) *Emerging imaging technologies in medicine.* (CRC Press).
19. Reddy RK, Bhargava R (2010) Accurate histopathology from low signal-to-noise ratio spectroscopic imaging data. *Analyst* 135(11):2818–2825.
20. Lee JB, Woodyatt AS, Berman M (1990) Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform. *Geoscience and Remote Sensing, IEEE Transactions on* 28(3):295–304.
21. Bushberg JT, Boone JM (2011) *The essential physics of medical imaging.* (Lippincott Williams & Wilkins).
22. Musco C, Musco C (2015) Stronger approximate singular value decomposition via the block lanczos and power methods. *Advances in Neural Information Processing Systems (NIPS).*
23. Halko N, Martinsson PG, Tropp JA (2011) Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* 53(2):217–288.
24. Kong X, Li K, Yang Q, Wenyin L, Yang MH (2013) A new image quality metric for image auto-denoising in *Proceedings of the IEEE International Conference on Computer Vision.* pp. 2888–2895.