Estimating and suppressing background in Raman spectra with an artificial neural network.

Sigurdur Sigurdsson[1], Jan Larsen[1], Peter Alshede Philipsen[2],
Monika Gniadecka[2], Hans Christian Wulf[2] and Lars Kai Hansen[1]

November 26, 2003

[1] Informatics and Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark.

[2] Department of Pathology, Bispebjerg Hospital, University of Copenhagen, DK-2400 Copenhagen, Denmark.

Abstract

In this report we address the problem of skin fluorescence in feature extraction from Raman spectra of skin lesions. We apply a highly automated neural network method for suppressing skin fluorescence from Raman spectrum of skin lesions before dimension reduction with principal components analysis. By applying the background suppression, the effect of outlier spectrum in the principal components analysis was greatly reduced.

Introduction

Studies have shown that Raman spectra of skin cancer has potential for skin cancer diagnosis [2, 4, 5, 6, 3]. Raman spectra are obtained with equipment consisting of a radiation source and a spectrometer. The laser beam excites the molecules and two scattering processes are observed, elastic and inelastic. The elastic process is the so-called *Rayleigh scattering* where the reflected wavelength is the same as the exited wavelength. In contrast, the inelastic scattering changes the reflected wavelength, giving a frequency shift in the reflected Raman spectra. This is called *Raman scattering.* The Raman spectra of two molecules are different if they have different structure, thus specific substances can be identified by their Raman spectra.

The radiation source for the excitation used in this study is a neodymium doped yttrium-aluminium garnet laser emitting at 1064 nm. This frequency lies in the near infrared region which makes the spectra less vulnerable to sample fluorescence, which may severely corrupt the spectra. Nevertheless, the influence of sample fluorescence or

[0] S. Sigurdsson was supported by the Danish Research Councils through the project Signal and Image Processing for Telemedicine (SITE)

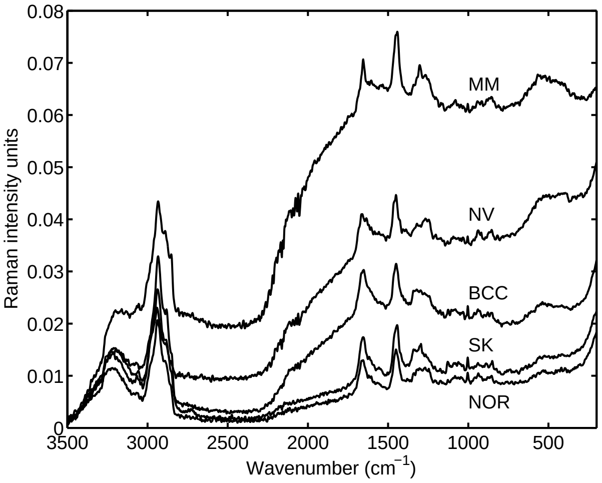

Figure 1 Raman spectra of skin samples from basal cell carcinoma (BCC), malignant melanoma (MM), normal skin (NOR), pigmented nevi (NV) and seborrhoeic keratosis (SK).

background may be observed. In Figure 1 we show examples of Raman spectra of 4 different types of skin tumors and also of normal skin. The background is the amplitude elevation in the region below 2800 cm^{-1}. The narrow peaks represent the vibration of chemical bonds carrying the information of molecular structure of different lesion types. It has been observed that one of the factors controlling the amplitude of the background in Raman spectrum of skin is the amount of pigmentation in the sample [7].

The dimension of the Raman spectra is usually high, e.g., for the data set used here 1711 dimension are used for representing each spectra. Using pattern recognition methods in high dimensional input spaces is problematic as it suffers from the *curse of dimensionality* [1]. This may be solved by using a dimension reduction method, where the most common method is the principal components analysis (PCA). Assuming that the presence of multiple signal classes in a skin lesion is the major source of variation, a natural choice for dimension reduction is PCA. The PCA is very sensitive to outliers in the data, i.e., data points far from the main density of the data. In the skin lesion data set there are a few spectra that have very high background amplitude, making them potential outliers for the PCA.

In this report we suggest a method for suppressing the background. We propose to to model the background with a neural network and subtracting from the original Raman spectra, thus generating a background suppressed spectra. The neural network model incorporates adaptive regularization to avoid overfitting.

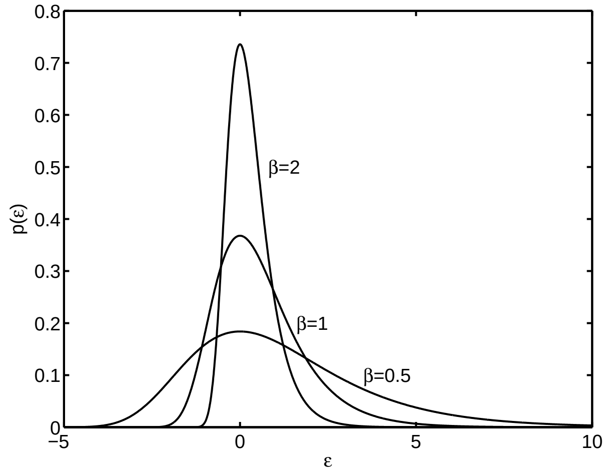

Figure 2 The Gumbel distribution for three different values of β. Note the skewness of the distribution giving higher probability of positive values.

Background modeling

As the background is little in the frequency region over 2800 cm^{-1}, we fit linearly the Raman spectra between the frequency points at 3500 cm^{-1} and 2800 cm^{-1}. The background in the spectra below 2800 cm^{-1} is modeled with a flexible artificial neural network due to the variability of the background. The remainder of this section describes the neural network in details.

Noise model

The noise model for a Raman spectrum may be defined

$$d(x) = y(x) + \varepsilon, \tag{1}$$

where the spectrum $d(x)$ is generated from the background $y(x)$ at a wave number x and an additive noise ε, in this case the Raman peaks. Visual inspection of the spectra shows that the noise, in this case the Raman peaks, is far from symmetric. This can be seen in Figure 1, where the noise is mostly positive. Thus, assuming the usual Gaussian distributed noise is inappropriate. Instead, we suggest to model the noise as a zero location Gumbel distribution [10] denoted $\text{Gum}(0, \beta^{-1})$, where β is the inverse scale parameter. The probability density function is written as

$$p(\varepsilon) = \beta \exp(-\beta\varepsilon) \exp\left(-\exp\left(-\beta\varepsilon\right)\right). \tag{2}$$

In Figure 2 we show the probability density function for three different values of β.

Neural network architecture

The neural network architecture used for the modeling is a two-layer feed-forward neural network. The input-to-hidden layer is given by

$$h_j(x) = \tanh\left(w_{j1}x + w_{j0}\right), \tag{3}$$

where w_{j1} are the input to hidden weights, w_{j0} is the input to hidden bias and $h_j(f)$ is the output of the jth sigmoidal activation function of the hidden layer. The neural network output, an estimate of the background, is given by

$$\hat{y}(x) = \sum_{j=1}^{H} w_j h_j(x) + w_0, \tag{4}$$

where w_j are the hidden to output weights, w_0 is the input to hidden bias and H is the number of units in the hidden layer. To simplify notation we define the neural network weight vector as $\mathbf{w} = [w_1, w_2, \ldots, w_W]^\top$ holding all weights, where W is the number of weights.

Inferring the weights

To infer the weights we use a set of data, consisting of input-output pairs $\mathcal{D} = \{d^{(n)}, x^{(n)}\}$ where $n = 1, 2, \ldots, N$, $d^{(n)}$ is the value of the Raman spectrum and $x^{(n)}$ is the corresponding wave number.

The neural network weights are inferred with a *maximum a posteriori* (MAP) estimate, where we maximize the *posterior*

$$p(\mathbf{w}|\mathcal{D}, \beta, \alpha) \propto p(\mathcal{D}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha), \tag{5}$$

where $p(\mathcal{D}|\mathbf{w}, \beta)$ is the *likelihood* and $p(\mathbf{w}|\alpha)$ is the *prior.* We assume that the hyperparameters α and β are known when inferring the weights. For numerical convenience we minimize a cost function which is the negative logarithm of the posterior,

$$\begin{aligned} C(\mathbf{w}, \alpha, \beta) &\propto -\ln p(\mathbf{w}|\mathcal{D}, \beta, \alpha) & (6) \\ &= E(\mathbf{w}, \beta) + R(\mathbf{w}, \alpha), & (7) \end{aligned}$$

where $E(\mathbf{w}, \beta) \propto -\ln p(\mathcal{D}|\mathbf{w}, \beta)$ and $R(\mathbf{w}, \alpha) \propto -\ln p(\mathbf{w}|\alpha)$. The term $R(\mathbf{w}, \alpha)$ may be interpreted as a regularization term. The negative log-likelihood may be derived directly from the noise distribution in equation 2

$$E(\mathbf{w}, \beta) = \sum_{n=1}^{N} \left[\beta\left(\hat{y}(\mathbf{w}, x^{(n)}) - d^{(n)}\right) + \exp\left(-\beta\left(\hat{y}(\mathbf{w}, x^{(n)}) - d^{(n)}\right)\right)\right], \tag{8}$$

where we have exchanged $y(x)$ with the estimate $\hat{y}(\mathbf{w}, x^{(n)})$. We assume a weight decay regularization, given by

$$R(\mathbf{w}, \alpha) = \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}. \tag{9}$$

To optimize the neural network weights we use a state of the art BFGS quasi-Newton optimizer [9]. Thus, only the gradient of the cost function with respect to the weights needs to be computed. The gradient is given by

$$\frac{\partial C(\mathbf{w},\alpha,\beta)}{\partial \mathbf{w}} = \beta \sum_{n=1}^{N} \left[\left(1 - \exp\left(-\beta\left(\hat{y}(\mathbf{w},x^{(n)}) - d^{(n)}\right)\right)\right) \frac{\partial \hat{y}(\mathbf{w},x^{(n)})}{\partial \mathbf{w}} \right] + \alpha \mathbf{w}. \tag{10}$$

The derivative for the output of the neural network for the architecture given in equations (3) and (4) needs to be computed. For a single input-to-hidden and input bias weight

$$\frac{\partial \hat{y}}{\partial w_{j1}} = w_j(1-(h_j(x))^2)x \qquad \frac{\partial \hat{y}}{\partial w_{j0}} = w_j(1-(h_j(x))^2), \tag{11}$$

where we use differential rule $\tanh'(x) = 1 - \tanh^2(x)$. For a single hidden-to-output and output bias weight

$$\frac{\partial \hat{y}}{\partial w_j} = h_j(x) \qquad \frac{\partial \hat{y}}{\partial w_0} = 1. \tag{12}$$

Adapting hyperparameters

After the weights have been optimized the hyperparameters α and β have to be adapted.

The update rule for β is simply the maximum likelihood estimate [10], given by

$$\beta^{\text{new}} = \left(\frac{1}{N}\sum_{n=1}^{N} \epsilon^{(n)} - \frac{\sum_{n=1}^{N} \varepsilon^{(n)} \exp\left(-\beta\varepsilon^{(n)}\right)}{\sum_{n=1}^{N} \exp\left(-\beta\varepsilon^{(n)}\right)} \right)^{-1} \tag{13}$$

where $\varepsilon^{(n)} = \hat{y}(\mathbf{w},x^{(n)}) - d^{(n)}$.

The update rule for the α is computed as in [8], where $\ln p(\mathcal{D}|\alpha,\beta)$ is maximized. By evaluating $\partial \ln p(\mathcal{D}|\alpha,\beta)/\partial\alpha$ we acquire the following update formula

$$\alpha^{new} = \frac{\gamma}{2E_W(\mathbf{w}_{\text{MP}})}, \tag{14}$$

where $\gamma = W - \alpha \text{Trace}\mathbf{A}^{-1}(\mathbf{w})$ is the effective number of weights in the network and $\mathbf{A}(\mathbf{w}) = \frac{\partial^2 C(\mathbf{w},\alpha,\beta)}{\partial\mathbf{w}\partial\mathbf{w}^\top}$ is the Hessian. The full Hessian may be computed as

$$\begin{aligned}\mathbf{A}(\mathbf{w}) \quad = \quad & \beta \sum_{n=1}^{N} \left[\left(1 - \exp\left(-\beta\left(\hat{y}(\mathbf{w},x^{(n)}) - d^{(n)}\right)\right)\right) \frac{\partial^2 \hat{y}(\mathbf{w},x^{(n)})}{\partial\mathbf{w}\partial\mathbf{w}^\top} \right.\\ & \left. -\beta \exp\left(-\beta\left(\hat{y}(\mathbf{w},x^{(n)}) - d^{(n)}\right)\right) \frac{\partial \hat{y}(\mathbf{w},x^{(n)})}{\partial \mathbf{w}} \frac{\partial \hat{y}(\mathbf{w},x^{(n)})}{\partial \mathbf{w}^\top} \right] + \alpha\mathbf{I}. \end{aligned} \tag{15}$$

Computing the full Hessian is demanding, thus an outer product approximation is applied, given by

$$\mathbf{A}(\mathbf{w}) \approx -\beta^2 \sum_{n=1}^{N} \left[\exp\left(-\beta\left(\hat{y}(\mathbf{w},x^{(n)}) - d^{(n)}\right)\right) \frac{\partial \hat{y}(\mathbf{w},x^{(n)})}{\partial \mathbf{w}} \frac{\partial \hat{y}(\mathbf{w},x^{(n)})}{\partial \mathbf{w}^\top} \right] + \alpha\mathbf{I}. \tag{16}$$

The approximation also ensures a symmetric Hessian.

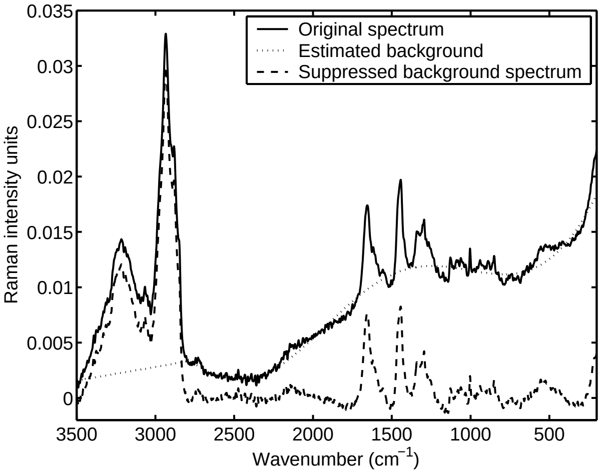

Figure 3 An example of a background suppression for a single Raman spectrum.

Experiments

The data set with Raman spectra of skin lesions consisted of 222 spectra. The background was estimated and suppressed with a neural network. The neural network was initialized with $H = 20$ hidden units. In Figure 3 we illustrate the background suppression of a single Raman spectrum. The background is fairly well estimated and smooth, giving a suppressed background spectra with emphasized Raman peaks.

The PCA may be evaluated using singular value decomposition. The data matrix $\mathbf{D}$ of size $F \times N$, where N is the number of examples and F is the number of frequency components, is decomposed as

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^\top, \tag{17}$$

where $\mathbf{U}$ is a $F \times N$ orthonormal matrix, $\mathbf{S}$ is a $N \times N$ diagonal matrix and $\mathbf{V}$ is $N \times N$ orthonormal matrix by using an economy size decomposition. The diagonal of matrix $\mathbf{S}$ has nonnegative elements in descending order. These diagonal elements are the singular values that correspond to standard deviations of the input data projected onto the given basis vectors represented by matrix $\mathbf{U}$. The reduced input space is obtained by using only some fixed $K \leq N$ number of the largest principal components giving the transformation matrix $\widetilde{\mathbf{U}}$ of size $K \times N$. A set of dimension reduced $K \times N$ feature vectors $\mathbf{X}$ may be computed with

$$\mathbf{X} = \widetilde{\mathbf{U}}^\top \mathbf{D}. \tag{18}$$

In the following experiment we use $K = 25$ principal components.

The evaluation of the difference for a single feature vector $\mathbf{x}^{(n)}$ when used to evaluate the PCA or not, may be computed using the squared two norm, given by

$$e(n) = \left\| \left(\widetilde{\mathbf{U}}^\top - \widetilde{\mathbf{U}}_{\setminus n}^\top \right) \mathbf{d}^{(n)} \right\|^2, \tag{19}$$

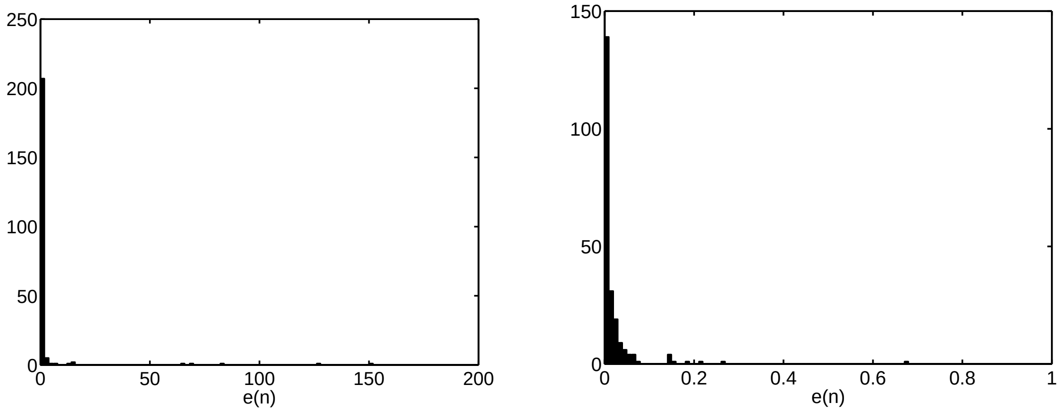

Figure 4 Histograms over error distribution from equation 19 for the 222 examples (left) without background suppression and (right) with background suppression.

where $\widetilde{\mathbf{U}}^{\top}_{\setminus n}$ is evaluated with all examples except spectrum $\mathbf{d}^{(n)}$. This is done for all examples in a leave-one-out manner, with and without background suppression. Figure 4 shows the histogram over the error with and without background suppression. Without the suppression of background, a few spectra have extremely high error compared to background suppressed spectra.

To evaluate the influence of outliers we examine the density in the tail of the distribution of the error. This is done by computing the ratio of standard deviations of the error, given by $\sigma_e/\sigma_{e,0.95}$, where σ_e is the standard deviation using all examples, and $\sigma_{e,0.95}$ is the standard deviation where 5% of the examples with the largest error are removed. The results gave the ratio 5.5 and 27.2, with and without background suppression, respectively. Thus, by removing the background the 'tail to body' ratio $\sigma_e/\sigma_{e,0.95}$ is reduced by a factor 5.

Conclusion

We have applied a highly automated method for removing skin fluorescence from Raman spectrum of skin lesions prior to dimension reduction with PCA. This was done by modelling the skin fluorescence with a regression type neural network and subtracting from the original spectra.

The influence of outlier spectrum, having high background amplitude, on the PCA estimation was investigated. The influence was evaluated by computing the difference in the PCA transformation of each Raman spectrum when used for estimating the PCA and not. The results showed that spectrum with high background amplitude clearly influenced the PCA, for some cases of Raman spectrum the difference was considerable. By applying the background suppression, the effect of outlier spectrum was reduced in the PCA.

References

[1] R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princton University Press, Princeton, New Jersey, first edition, 1961.

[2] S. Fendel and B. Schrader. Investigation of Skin and Skin Lesions by NIR-FT-Raman Spectroscopy. *Fresenius Journal of Analytical Chemistry*, 360:609–613, 1998.

[3] M. Gniadecka, P.A. Philipsen, S. Sigurdsson, S. Wessel, O.F. Nielsen, D.H. Christensen, J. Hercogova, K. Rossen, H.K. Thomsen, R. Gniadecki, L.K. Hansen, and H.C. Wulf. Malignant Melanoma Diagnosis by Raman Spectroscopy and Neural Network: Structure Alterations in Proteins and Lipids in Intact Cancer Tissue. *Journal of Investigative Dermatology, in press*, 2003.

[4] M. Gniadecka, S. Wessel, O.F. Nielsen, D.H. Christensen, J. Hercogova, K. Rossen, and H.C. Wulf. Potential of Raman Spectroscopy for in vitro and in vivo Diagnosis of Malignant Melanoma. In A.M. Heyns, editor, *Proceedings of the sixteenth International Conference on Raman Spectroscopy*, volume 16, pages 764–765, Chichester, 1998. John Wiley and Sons.

[5] M. Gniadecka, H.C. Wulf, N.N. Mortensen, O.F. Nielsen, and D.H. Christensen. Diagnosis of Basal Cell Carcinoma by Raman spectra. *Journal of Raman Spectroscopy*, 28:125–129, 1997.

[6] M. Gniadecka, H.C. Wulf, O.F. Nielsen, D.H. Christensen, and J. Hercogova. Distinctive Molecular Abnormalities in Benign and Malignant Skin Lesions: Studies by Raman Spectroscopy. *Photochemistry and Photobiology*, 66:418–423, 1997.

[7] L. Knudsen, C.K. Johansson, P.A. Philipsen, M. Gniadecka, and H.C. Wulf. Natural Variations and Reproducibility of in vivo near-infrared Fourier Transform Raman Spectroscopy of Normal Human Skin. *Journal of Raman Spectroscopy*, 33:574–579, 2002.

[8] D.J.C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 1992.

[9] H.B. Nielsen. UCMINF - An Algorithm for Unconstrained Nonlinear Optimization. Technical Report IMM-REP-2000-19, Department of Mathematical Modelling, Technical University of Denmark, 2000.

[10] S. Kotz and N.L. Johnson and C.B. Read, editor. *Encyclopedia of Statistical Scienes*, volume 3. John Wiley & Sons, New York, 1982.