Fast computation of absorption spectra for lidar data processing using principal component analysis

MATTHEW HAYMAN,1,* ROBERT A. STILLWELL,2 AND SCOTT M. SPULER1

1Earth Observing Laboratory, National Center for Atmospheric Research, Boulder, Colorado 80307, USA
2Advanced Studies Program, National Center for Atmospheric Research, Boulder, Colorado 80307, USA
*Corresponding author: mhayman@ucar.edu

Received 21 February 2019; revised 14 March 2019; accepted 16 March 2019; posted 18 March 2019 (Doc. ID 360784); published 2 April 2019

We describe a principal component-based technique for approximating absorption and scattering spectra commonly needed for lidar signal processing. Where previously these calculations had been bottlenecks in our lidar signal processing, the described approach has increased our spectrum calculation speed by over two orders of magnitude. The described approach also allows analytically calculated temperature and pressure derivatives, which is useful for propagating uncertainty and implementation of global optimization algorithms. © 2019 Optical Society of America

Knowledge of absorption and scattering spectra is commonly needed to perform quantitative lidar retrievals. For example, differential absorption lidar (DIAL) data processing makes use of a priori knowledge of the species of interest’s absorption features to obtain the species number density [1], and high spectral resolution lidar (HSRL) typically calculate the Cabannes backscatter spectra to estimate the lidar’s range dependent collection efficiency for molecular backscattered light [2].

At the National Center for Atmospheric Research (NCAR), we have historically computed absorption spectra using parameterizations from the high-resolution transmission (HITRAN) molecular absorption database [3]. To obtain an accurate estimate of water vapor absorption, we include lines over a broad range of wavelengths (tens of nm). Accurate calculation of how these features fold into a relatively narrow frequency band of interest (on the order of tens of GHz) is computationally expensive. The estimation of absorption spectra on a pixel-by-pixel basis can be a bottleneck in DIAL processing routines. For this reason, it is often not practical to estimate the spectrum for more than a few of the, potentially hundreds, of vertical profiles in a day.

Backscatter spectra from the Cabannes line (sometimes referred to as Rayleigh-Brillouin or Rayleigh-Doppler) are traditionally computed using a Tenti S6 model [4]. Similar to the absorption spectra, this calculation is computationally expensive though perhaps less critical for data accuracy than that needed for DIAL processing routines.

In a collaboration between NCAR and Montana State University, we have developed a series of lidar instruments based on diode-laser technology. These systems are, in effect, multiple lidars integrated into one package, consisting of a water vapor DIAL [5], an HSRL [6] and most recently, an oxygen DIAL (used to estimate temperature with retrieval details given in [7]). The calculation of temperature from oxygen relies on measuring the extinction coefficient of a temperature sensitive oxygen line. The analytical link between temperature and the observed extinction is quite complicated. Consequently, a numerical solution is typically employed and the spectrum calculated several times through the iteration process. This is problematic given the computational expense already posed by the calculation. In addition, any new development in global optimal estimation across all three lidars will require the spectrum derivatives as a function of temperature.

Binietoglou et al. [8] recently published a work showing that the Cabannes spectra from molecular scattering can be accurately approximated using a principal component based method. In this approach, they perform a two step analysis; they first perform principal component analysis on a training set of Cabannes spectra—computed using the Tenti S6 model [4] for a variety of temperatures and pressures spanning the expected pressure and temperature space—then, they estimate the weights for each component as a function of a single parameter that is dependent on pressure and temperature and fit a polynomial to those weights. The result is a product of two matrices; the principal components and the polynomial fit terms which allow the Cabannes spectra to be estimated for an arbitrary temperature and pressure within the training region. This principal component technique can be thought of as an interpolation technique for data tables, which are approximated by polynomials. In the case of Cabannes spectra described by Binietoglou et al., this is a 1D table.

The work of Binietoglou et al. shows that principal component analysis with polynomial fitting is a computationally efficient means of approximating and interpolating a multi-dimensional table of spectral features, which underlies the approach described here. In [8], the spectral shape is determined by a single parameterized variable, the frequency axis is also rescaled depending on temperature; the method described here uses a fixed frequency
grid at the expense of an additional independent variable. The fixed frequency grid avoids the need to perform an additional interpolation step on the resultant spectra when combining with other analysis terms and simplifies the calculation of pressure and temperature derivatives. The trade-off is that Cabannes spectra cannot be generalized to an arbitrary wavelength and separate training data is needed for each wavelength of operation.

In this work, we focus on describing the implementation of this PCA based calculation of absorption spectra, and how this enables directly calculated derivatives. The results of this work can also be generalized to Cabannes spectra, but this will not be covered in detail.

The principal component analysis described here is a method of interpolating predefined spectra that are obtained using HITRAN parameterizations without storing individual training data points. The computation of spectra based on the HITRAN database is effectively moved offline when the estimation matrices are produced and written for future use. The result of the PCA analysis, combined with a polynomial fit, is a sparse data set that is stored as a matrix polynomial and a truncated set of principal components. The training data is not retained for the final spectrum calculations.

Consider a spectrum \( y_i(\nu_i, T, P) \) that is a function of temperature, \( T \), and pressure, \( P \), and the frequency components are discretely related to the optical frequency, \( \nu_i \), via the index \( i \). Using HITRAN parameters, we compute \( y_i(\nu_i, T, P) \) over a span of temperature, pressure and frequency, defined by the expected range needed for the application. This is the training data set.

A matrix of all simulated spectra \( Y \) is then generated such that a matrix element is given by \( y_i(\nu_i, T_j, P_j) \) where the frequency basis is predefined and the temperature/pressure combination comes from a previously defined mesh of the two variables. We assume that this matrix can be approximated by a set of principal components such that

\[
Y = \tilde{\mu}_Y + U\hat{W},
\]

where \( Y \) is the matrix of spectra, \( \tilde{\mu}_Y \) is a vector containing the mean of each row of \( Y \), \( U \) is a matrix containing the principal components of \( Y \) (obtained by preforming singular value decomposition on \( Y \)), and \( \hat{W} \) is a matrix containing the weights for each principal component. The weights can be determined for the training data set by solving Eq. (1) for \( \hat{W} \)

\[
\hat{W} = U^\dagger(Y - \tilde{\mu}_Y),
\]

where \( ^\dagger \) denotes a matrix transpose operation.

It was noted by [8] that the weights of the principal components tend to be smoothly varying and generally well approximated by polynomials. In that work, there was only one independent parameterized variable; in this case, we do not leverage any theory relating temperature and pressure, so we have to expand to a two variable polynomial. Nevertheless, we find that the weights are generally smoothly varying surfaces that are well approximated by two variable polynomials. Adding degrees of freedom increases the memory requirements for the analysis, but does not seem to present any significant challenge for a standard laptop computer. For a given temperature and pressure, we describe the polynomial expansion using a state vector

\[
\tilde{\theta}_j = [1 \ \tilde{T}_j \ \tilde{P}_j \ \tilde{T}_j\tilde{P}_j \ \tilde{T}_j^2\tilde{P}_j \ \ldots \ \tilde{T}_j^N\tilde{P}_j^N]^\dagger,
\]

where the \( \tilde{P} \) and \( \tilde{T} \) are a normalized temperature and pressure; respectively, to ensure the polynomial fit is well conditioned. In our routines, normalization consists of subtracting the mean value of the training set and dividing by its standard deviation. For example,

\[
\tilde{T}_j = \frac{T_j - \mu_T}{\sigma_T},
\]

where \( \mu_T \) is the average of all temperatures in the training data set and \( \sigma_T \) is the standard deviation of all temperatures in the training data set.

We use this description of the atmospheric state to approximate the principal component weights according to

\[
\hat{W} = C\theta,
\]

where \( C \) is a matrix containing the polynomial coefficients of the fit and \( \theta \) is a state matrix where each column is a state vector of the form in Eq. (5). In the training process, we obtain the polynomial weights from the state matrix and the principal component weights of the training data

\[
C = W\theta^{-1},
\]

where \( \theta^{-1} \) is the right inverse of \( \theta \) and \( W \) is obtained from Eq. (2). Ideally, this problem is overdefined so a pseudo-inverse is used to calculate \( C \).

Thus, for future calculations of a particular temperature and pressure pair, we can approximate the spectrum using the matrix equation

\[
\tilde{y}_j = \tilde{\mu}_Y + M\tilde{\theta}_j,
\]

where

\[
M = UC.
\]

The process of defining the principal component matrices, needed to later calculate a spectrum with Eq. (7), proceeds as follows:

1. Calculate a data set, \( Y \), (e.g., using HITRAN parameterizations) and corresponding \( \theta \) for an expected range of temperature and pressure.
2. Obtain \( \tilde{\mu}_Y \) and the principal components, \( U \), from \( Y \). Generally, \( U \) will consist of a truncated set of all principal components based on the number of principal components/accuracy desired.
3. Calculate the weights, \( \hat{W} \), of the principal components for each temperature/pressure pair using Eq. (2).
4. Calculate the polynomial fit matrix, \( C \), using Eq. (6).
5. Calculate \( M \) from Eq. (8).

After completion of these steps, we store \( \tilde{\mu}_Y \) and \( M \) along with the frequency grid of the columns of \( Y \) (the training data \( Y \) is not stored). At any future point, we can estimate the spectrum using Eq. (7) for any temperature/pressure pair as long as it is inside the bounds of the training data. Thus, in some sense, this method is an efficient way of interpolating a table of spectra with temperature and pressure dependence. The computation time is also accelerated considerably (hundreds of times faster than our HITRAN calculations).

There is an additional benefit to the PCA approach—the approximation of the spectra has analytically defined derivatives with pressure and temperature. For example, from Eq. (7), we find the temperature derivative is...
where the derivative of the state variable is easily obtained using the power rule on Eq. (3)

$$\frac{\partial \tilde{\theta}_j}{\partial T_j} = [0 \ 1 \ 0 \ \hat{T}_j \ 0 \ N\hat{T}_j^{N-1}N_j^N] \frac{\partial \hat{T}_j}{\partial T_j},$$

where from Eq. (4) we obtain

$$\frac{\partial \hat{T}_j}{\partial T} = \frac{1}{\sigma_T},$$

Since numerical calculation of gradients can be computationally expensive (requiring at least one additional function call for each independent parameter)—particularly in cases where lidar data consists of hundreds of data points with potentially independent temperature and pressure terms—this can have a significant impact on the speed of numerical solvers. It also provides a simple way to propagate temperature and pressure uncertainty into the final derived data product.

As a demonstration of the PCA parameterization, we consider the water vapor absorption line near 828.2 nm used by the Micro-Pulse DIAL (MPD) [5].

The training dataset is a 50 × 50 point mesh of temperature and pressure pairs, loosely bounded by a hydrostatic assumption to avoid performing the analysis on unlikely combinations of temperature and pressure. For these pairs, we calculate \(Y\) over a 10 GHz range at 10 MHz frequency resolution using HITRAN. The corresponding polynomial state variable \(q\) is obtained for 10 orders in both temperature and pressure. From \(Y\) we obtain the mean of each row \(\tilde{\mu}_Y\) and the principal components, \(U\).

The weights of the principal components are determined according to Eq. (2). Figure 1 shows the surfaces describing the weights for the first four principal components. Note that they are smoothly varying and generally good candidates for polynomial approximation. It is important that the data points are sufficiently dense to sample surface variations. The 50 × 50 grid appears to be sufficient for most of the cases, but variations increase near the surface edge at low pressure. Using these weights, and \(q\) of the training data, we obtain the polynomial coefficients, \(C\), from Eq. (6).

Figure 2 shows an example of the water vapor spectrum near 828.2 nm at arbitrarily selected 0.58 atm and 263.1 K. The spectrum calculated from HITRAN is shown in blue. The spectrum resulting from the PCA parameterization is shown in orange. A linearly interpolated spectrum from the training data is shown in green and the nearest neighbor in the training data (at 0.57 atm and 262.4 K) is shown in red. Below the spectrum is a plot of the residual absolute value errors in the approximations relative to the HITRAN calculation. In this particular example, the maximum error using the PCA approximation is about five orders of magnitude less than the amplitude of the absorption feature. The comparatively high error in the linear interpolation and nearest neighbor spectra indicate that these methods would require substantially higher point density than the PCA training set to achieve comparable accuracy.

The method for evaluating the merit of a PCA approximation might take a few different approaches. Here we consider the absolute largest percent error in the spectrum, but in Section 4 we will also consider percent RMS error. Figure 3 shows the maximum error in the polynomial computed spectrum for each temperature and pressure point used in the PCA training dataset. Most of the cross section errors are on the order of 0.01% to 0.1%, but errors increase substantially in edge regions of very low temperature and pressure. Figure 1 suggests these regions are also where the surfaces describing the PCA weights vary rapidly. If higher accuracy is needed in these areas, we would consider expanding the training data to ensure that

![Fig. 1. Weights of the first four principal components as a function of temperature and pressure.](image1)

![Fig. 2. Example of water vapor spectrum (top) near 828 nm for a pressure of 0.58 atm at 263.1 K computed using HITRAN (blue), the PCA method described here (orange) linear interpolation of training data (green) and the nearest neighbor in the training data (red). The bottom shows the difference between the HITRAN calculations and the approximation approaches.](image2)

![Fig. 3. Maximum percent error of each training data point in the example case.](image3)
important points are away from edges, or might also increase the polynomial order. In the case of MPD, these points are well outside the region where we expect to perform water vapor retrievals.

The primary benefit of a PCA spectroscopy approach is the speed it provides. For calculations requiring derivatives, it saves considerably on the number of computations by providing an analytically defined derivative. In addition, the PCA calculation itself tends to be much faster than our original HITRAN function.

We compared the calculation time of our HITRAN based code to the PCA approach for a range of 1–10,000 temperature/pressure pairs for the water vapor absorption line near 828.2 nm. When only a few spectra are computed at a time, the PCA approximation is roughly ten times faster than our HITRAN calculations. As the number of spectra calculations increases to hundreds and thousands, the PCA approximation approaches 500 times faster than our HITRAN method. It warrants consideration that our HITRAN code may not be optimized for speed, but it is unlikely that we can speed up its calculations by a factor of 500 to compete with the PCA approximation.

The PCA calculation for spectra is an approximation, and therefore the usefulness of this technique will be determined by the accuracy needed for a particular application. To some extent, the accuracy can be improved by adding principal components and/or polynomial orders. An analysis calculating the error in the fit was conducted for two to fourteen principal components and polynomial orders from one to twenty. Figure 4 shows error contours of the fit. The error is estimated four different ways. The overall largest maximum error of each spectrum represents the absolute worst point in the entire training set and is shown in (a) The training set average of the maximum error in each spectra is shown in (b) The overall maximum RMS error of each spectrum is shown in (c) Finally, the training set average of the RMS spectrum error is shown in (d).

In Fig. 4, we see that the worst case error drops below 1% for more than 10 principal components and polynomial orders greater than 9. The same region is comfortably below 0.1% by the other error metrics. In general, contours run along the axes of the plots, so adding more of one term (principal components or polynomial orders) may not improve the spectrum accuracy if it is not the limiting factor. This can be further extended to also consider the polynomial order of pressure and temperature separately if one wishes to optimize speed and memory usage.

Principal component derived spectra for lidar signal processing has the ability to substantially improve data processing efficiency. Where previously we limited our spectroscopy calculations to only a few profiles in a day, the PCA approach has enabled calculations for each pixel while still accelerating the processing. In addition, this PCA parameterization allows us to calculate spectrum derivatives as a function of temperature and pressure. This is key to enabling forward inversions of temperature from oxygen DIAL data, where estimation of a temperature dependent absorption cross section is used to obtain a profile.

**Funding.** National Center for Atmospheric Research (NCAR) is funded by the National Science Foundation.

**REFERENCES**