Preprocessing: Smoothing and Derivatives

Vladimir Bochko    Jarmo Alander

University of Vaasa

November 1, 2010
Contents

- Preprocessing
- Baseline correction.
- Autoscaling.
- Savitzky-Golay Smoothing Filters
- Savitzky-Golay Filters for Derivatives
- Smoothing Splines
- Splines for Derivatives
Preprocessing

- **Normalizing.** It is done by dividing each feature related to a sample vector by a constant. The constant value may be defined as the 1-norm or 2-norm of the vector.
- **Weighting.** For weighting any values are used. The purpose to give importance only to some samples.
- **Smoothing.** Reduce noise and keep useful variation. Smoothing filters.
- **Baseline correction.** Reduce unwilling slow variation. Estimation of baseline and subtraction from spectrum, Derivatives, Multiplicative Scatter Correction MSC.
We consider the measured spectrum \( s \). The baseline is approximated using polynomial [3], for example up to linear term.

\[
    s = \tilde{s} + \alpha + \beta x
\]

where \( x \) is a wavelength number, \( \alpha \) defines a shift (offset) in spectrum and \( \beta \) describes a linear change in spectrum.

- We are interested in \( \tilde{s} \) and the rest terms describe unwilling variation.
- If the baseline model has only shift then after estimating the shift we may subtract it from each spectrum.
Derivatives

- If we have only shift \( s = \tilde{s} + \alpha \) we can use derivative with respect \( x \) to remove the shift as follows: \( s' = (\tilde{s} + \alpha)' = \tilde{s}' \).
- Computing the successive derivatives we may remove the baseline defined by the higher order model.
- Numerically the first derivative for \( s = (s_1, s_2, \ldots, s_k) \) can be computed using a moving window where \( s_i \) is replaced by \( s_i - s_{i-1} \).
- For the second derivate we use the moving window:
  \[
  (s_i - s_{i-1}) - (s_{i-1} - s_{i-2}) = s_i - 2s_{i-1} + s_{i-2}
  \]
- We can also use the derivative using the moving window:
  \[
  \frac{(s_i + s_{i-1})}{2} - \frac{(s_{i-1} + s_{i-2})}{2}
  \]
  This computing is robust to noise because we average spectrum values.
The autoscaling procedure uses both mean centering and variance scaling and helps for removing unwilling variation.

- Mean centering. For each feature we compute mean and subtract it from the other feature values.
- Variance scaling is implemented by division each feature value by standard deviation. Standard deviation is defined using values of this feature.
Savitzky-Golay Smoothing Filters

- The measured spectra are usually slowly varying and corrupted by random noise. The beginning and the end of the spectra contain mostly noise. The purpose of smoothing is to reduce the level of noise keeping spectrum details.
- Savitzky-Golay filters were designed to analyze the absorption peaks in noisy spectra. The absorption peaks suggest which chemical elements are presented in the tested materials.
- Filters have several names Savitzky-Golay, least squares and Digital Smoothing Polynomial (DISPO) filters.
- The filters operate in the time domain and relate to low-pass filters.
Suppose we have a spectrum with equally spaced values \( s_i = s_0 + i\Delta \), where the spectral values are obtained at intervals \( \Delta \) and \( i = \ldots, -2, -1, 0, 1, 2, \ldots \). The finit impulse response filter replaces \( s_i \) by a linear combination of its neighbor values

\[
f_i = \sum_{n=n_{left}}^{n=n_{right}} c_n s_{i+n}
\]

where \( n_{left} \) and \( n_{right} \) are the number of points to the left and right of a current point, respectively. \( c_n \) are weight coefficients.
Savitzky-Golay Smoothing Filters

• For an averaging filter all coefficients are constant $c_n = 1/(n_{left} + n_{right} + 1)$. The coefficients are computed in a moving window.

• The Savitzky-Golay smoothing filter uses a polynomial of higher order. Least-squares fitting is used to fit the polynomial to the data inside the moving window. The use of the least square fit in this way may be computationally demanding. However it is possible to find the coefficients $c_n$ for which least square fitting inside a moving window is automatically implemented.
Next the Savitzky-Golay smoothing filter coefficients are given for the polynomial order $M$ [1].

<table>
<thead>
<tr>
<th>$M$</th>
<th>$c_{n\text{left}}$</th>
<th>$c_{n0}$</th>
<th>$c_{n\text{right}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.086 0.343</td>
<td>0.486</td>
<td>0.343 -0.086</td>
</tr>
<tr>
<td>4</td>
<td>0.035 -0.128 0.070 0.315</td>
<td>0.417</td>
<td>0.315 0.070 -0.128 0.035</td>
</tr>
</tbody>
</table>

In practice, however, the larger values $n_{\text{left}}$ and $n_{\text{right}}$ are used, e.g. $n_{\text{left}} = n_{\text{right}} = 16, 32$. 
Savitzky-Golay Smoothing Filters

- a) The generated synthetic signal \( a(x) \) which is used as an underlying function. b) The same signal with additive Gaussian random noise.
• a) Moving average, $n_{left} = n_{right} = 15$. b) Savitzky-Golay filter, $n_{left} = n_{right} = 15$, polynomial order 2. The smoothed result is shown by green dots.
Savitzky-Golay Smoothing Filters

- a) Savitzky-Golay filter, \( n_{left} = n_{right} = 15 \), polynomial order 2.
- a) Savitzky-Golay filter, \( n_{left} = n_{right} = 30 \), polynomial order 2. The smoothed result is shown by green dots.
This technique may be used with noisy data.

The polynomial and least square fitting are used at each point of spectra.

The measured derivative is the derivative of the fitted polynomial.

The Savitzky-Golay filters from libraries can be used in the derivative mode.

Computation is made in the moving window.
Smoothing Splines

- Splines and wavelets are the other popular techniques used for smoothing. For example, the smoothing spline may fit spectral data. For the next illustration the smoothing splines were used from the functional data analysis library [2]. a) The original spectrum of the gastrointestinal tract of the dog. b) The smoothed spectrum.
The computation of derivatives may be unstable numerically. To achieve robust computation a B-spline is used. After obtaining the analytical spline expansion of spectrum the first and second derivatives can be exactly computed. 

a) The smoothed spectrum of the gastrointestinal tract of the dog. 
b) The first derivative.


SOLO toolbox:
Questions