3
Mathematical Tools

3.1 STOCKWELL TRANSFORM

Stockwell Transform (or S Transform or ST)[Stockwell et al., 1996] is a phase corrected version of Continuous Wavelet Transform (CWT) which magnifies the information-specific frequency bands with proper dilation and contraction of the Gaussian mother wavelet. The ST has found a number of application in distribution systems with renewable energy sources [Mishra et al., 2008; Uyar et al., 2009; Behera et al., 2010; Mahela and Shaik, 2017b,a; Yang et al., 2014] and in transformer protection applications [Moravej et al., 2010].

It is a multi-resolution analysis based technique which provides frequency-dependent transform in direct relation to Fourier transform. It is also known to be an extended form of CWT with a Gaussian window as the mother wavelet. In addition to good resolutions at low frequencies (with wide window size) and high frequencies (with small window size) like CWT, it also provides referenced phase information. The local phase spectrum information delivers enhanced distinguishing features when used for classification cases. CWT can be used to derive S-transform with the addition of a phase term.

S-transform of signal $h(t)$ is defined as the CWT of $h(t)$ with a Gaussian mother wavelet multiplied by a phase factor, $e^{i2\pi f\tau}$,

$$S(\tau, f) = e^{i2\pi f\tau} W(\tau, d). \tag{3.1}$$

where, τ is the translation parameter, d is the dilation parameter which determines the width of the wavelet and $W(\tau, d)$ is the CWT of the signal $h(t)$ and is given as,

$$W(\tau, d) = \int_{-\infty}^{\infty} h(t) w(t - \tau, d) dt. \tag{3.2}$$

and the mother wavelet is defined as,

$$w(t, f) = \frac{|f|}{\sqrt{2\pi}} e^{-\frac{t^2 f^2}{2}} e^{i2\pi ft}. \tag{3.3}$$

The dilation factor d is the inverse of the frequency f. The S-transform $S(\tau, f)$ of the signal $h(t)$ is written as,

$$S(\tau, f) = \int_{-\infty}^{\infty} h(t) \frac{|f|}{\sqrt{2\pi}} e^{-\frac{(\tau - t)^2 f^2}{2}} e^{i2\pi ft} dt. \tag{3.4}$$

The S-transform can also be defined by shifting operations on Fourier Transform $H(f)$ of signal

$h(t)$ as,

$$S(\tau, f) = \int_{-\infty}^{\infty} H(\alpha + f) e^{-\frac{2\pi^2\alpha^2}{f^2}} e^{i2\pi\alpha\tau} dt. \tag{3.5}$$

where, $f(f \neq 0)$ is the shifting parameter in frequency domain and α is the frequency.

In correlation with Equation 3.5, its discrete analog is utilized to compute discrete S-transform which takes the advantage of high computational efficiency of the FFT.

The S-transform of the discrete-time signal $h[kT]$ for $k = 0, 1, 2, ...N-1$ of continuous counter-part $h(t)$ is defined as,

$$S\left[jT, \frac{n}{NT}\right] = \sum_{m=0}^{N-1} H\left[\frac{m+n}{NT}\right] e^{-\frac{2\pi^2 m^2}{n^2}} e^{\frac{i2\pi mj}{N}}, n \neq 0 \tag{3.6}$$

where, τ and f are replaced with jT and $\frac{n}{NT}$ respectively. $H[.]$ is the Discrete Fourier Transform of signal $h(t)$ given as,

$$H\left[\frac{n}{NT}\right] = \frac{1}{N} \sum_{k=0}^{N-1} h[kT] e^{-\frac{i2\pi nk}{N}} \tag{3.7}$$

where, j, m and $n = 0, 1, 2, ...N-1$, N and T is the period of the discrete-time and continuous-time signal; and for $n = 0$, it is defined by the constant as,

$$S[jT, 0] = \frac{1}{N} \sum_{m=0}^{N-1} h\left[\frac{m}{NT}\right] \tag{3.8}$$

The inverse of the Discrete S-transform is given as,

$$h[kT] = \sum_{n=0}^{N-1} \{\frac{1}{N} \sum_{j=0}^{N-1} S\left[jT, \frac{n}{NT}\right]\} e^{\frac{i2\pi nk}{N}} \tag{3.9}$$

The Discrete Stockwell Transform can be computed easily with the combination of the FFT algorithm and convolution. Thus, for computation purposes, shifted DFT of the signal and DFT of the Gaussian window are evaluated, then inverse DFT of their product is performed to obtain S-transform of the given signal. The resultant matrix provides the features of a signal in the time-frequency plane. The steps to determine ST are given below:
1. Calculate Fourier Transform of the signal, $x(t)$ using FFT.
2. Determine the shifted frequency response.
3. Multiply with the Gaussian window.
4. Take the inverse transform of the resultant signal.

Following these steps, a complex two-dimensional matrix (ST matrix) is obtained whose y-axis is the frequency and x-axis is the time. Thus, the ST matrix depicts the information in the time-frequency plane. With the help of contour plots, the significant differences can be observed between various cases. The S-transform localizes real and imaginary components separately, thus, providing a localized phase spectrum along with amplitude spectrum.

Applications of ST can is found in distribution systems, transmission line protection, power quality analysis, in some other fields such as in artefacts removal from fMRI, analysis of EEG and

brain signals, signal filtering, geophysical data analysis and image processing applications.

3.2 FISHER SCORE AND CORRELATION BASED FEATURE SELECTION

The feature set may contain members who are redundant or irrelevant features which may affect a classifier's performance. The feature selection aims at reducing the number of features based on irrelevancy and redundancy, thereby improving the performance of the classifiers [Guyon and Elisseeff, 2003]. The methods for feature selection are (1) Filter methods, (2) Wrapper methods. Filter methods select features on the basis of their scores obtained in various statistical tests or ranking algorithms for their relevance towards the output. These tests can be T-test, F-test, chi-square test, Pearson correlation, linear discriminant analysis, Fisher score. In wrapper methods, a subset of features is used to train and test the classifier which is formed by successively adding or removing features (forward or backward selection) [Chandrashekar and Sahin, 2014] and finally the subset which gives the best classifier performance is selected for classification. The wrapper methods are more efficient as compared to filter methods as they measure the worth of features concerning classifier's performance. However, they also suffer from computational time and complexity whereas the filter methods are fast, simpler and easier methods to implement.

In this study, the Fisher score criterion, a ranking based method is chosen for feature selection, owing to its simplicity. The Fisher score is a supervised feature selection method used to rank feature based on its score that quantifies its discriminating ability under Fisher criterion [He et al., 2006]. The Fisher score for the j^{th} feature is given as:

$$F_j = \frac{\sum_{i=1}^{c} n_i(\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2} \tag{3.10}$$

where, for c classes in which the samples are to be classified, and for the j^{th} feature, μ_i is the mean, σ_i is the standard deviation, and n_i is the number of samples for i^{th} class, μ is the mean of feature over the dataset. Thus, for a feature, higher the score, better is the discriminating ability it possess, higher is the rank among all features.

Further, as explained in [Guyon and Elisseeff, 2003], features should be highly correlated with the class and less correlated with each other for better classifier performance. Thus, the Pearson correlation coefficient is used to measure the feature correlation which is given as:

$$CCF(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} \tag{3.11}$$

where,

$$Cov(x,y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{3.12}$$

and, σ_x and σ_y are the standard deviations and $\bar{x}$ and $\bar{y}$ are means of x and y respectively.

3.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a popularly used mathematical tool for dimensionality reduction in modern data analysis. It helps to extract the most relevant information from a high-dimension dataset. It reveals the hidden structure of a complex dataset. By reducing the number of variables that represent a system or a process, first, sufficient information is achieved in fewer variables, second, lesser

relationships between these variables would reduce the computational burden and complexity. Thus, PCA is a technique for feature extraction, which does not involve removal of any feature directly, instead, extraction of the most informative combination of features. These features are independent of one another. This process transforms the original feature space into a space of lower dimensions, thus reducing the dimensions and also keeping the original information in each feature intact.

Thus, it solves the problem of high-dimensionality by finding a new basis, which is a linear combination of original features, that will transform the original space into new. Suppose $\mathbf{X}$ is a *mxn* matrix with *m* are the number of features or dimensions and *n* are the number of samples or observations, and *Y* is a projected matrix which has same dimensions, *mxn*, such that:

$$Y = PX \tag{3.13}$$

where, $p_i, x_i \in \Re^m$, and $p_i.x_i$ represent simple inner products. The *P* represents that matrix on the which the data is going to be projected. Thus, the data in *X* will be projected on the columns of *P*. Thus, rows of $P = p_1, p_2 ... p_m$ is the new basis and are termed as principal components. This basis is based on the variance of the data. The aim is to find the principal directions in the data such that variance is maximized and co-variance is minimized. The variance of each feature when the data is not zero mean is given as:

$$\sigma^2 = E[(X - X_{mean})^2] \tag{3.14}$$

Suppose, for the data in *X*, mean of *X* i.e. X_{mean} is subtracted to obtain again *X*, which has zero mean, then the co-variance between different features can be given as:

$$\mathbf{C_X} = \sigma^2 = \frac{1}{n}\mathbf{X}.\mathbf{X}^T \tag{3.15}$$

$\mathbf{C_X}$ is the covariance matrix has some properties like:
1. It is a square symmetric matrix of size *mxm*.
2. Diagonal elements are the variances of each feature.
2. Off-diagonal elements are the co-variances between each feature.
To optimize this matrix, two things are required:
1. Minimize the co-variance i.e. all off-diagonal elements are zero.
2. Maximize the variances of each feature.
For this purpose, the matrix $\mathbf{C_X}$ has to be diagonalized. This can be done by assuming that all the basis vectors of $\mathbf{P} = p_1, p_2 ... p_m$ are orthonormal i.e. $\mathbf{P}$ is a orthonormal matrix. To solve for this, co-variance matrix $\mathbf{C_Y}$ for $\mathbf{Y}$ is defined as:

$$\begin{aligned}
\mathbf{C_Y} &= \frac{1}{n}\mathbf{Y}\mathbf{Y}^T \\
&= \frac{1}{n}(\mathbf{PX})(\mathbf{PX})^T \\
&= \frac{1}{n}\mathbf{PX}\mathbf{X}^T\mathbf{P}^T \\
&= \mathbf{P}(\frac{1}{n}\mathbf{X}\mathbf{X}^T)\mathbf{P}^T \\
\mathbf{C_Y} &= \mathbf{P}\mathbf{C_X}\mathbf{P}^T
\end{aligned} \tag{3.16}$$

With the theorems of Linear Algebra, any symmetric matrix $\mathbf{A}$ can be diagonalized with the help of an orthonormal matrix of its eigenvectors. Thus, $\mathbf{A} = \mathbf{S}\mathbf{D}\mathbf{S}^T$, where *S* is the matrix of eigenvectors arranged as columns and *D* is the diagonal matrix. In the present case, *P* can be assumed as a orthonormal matrix whose rows are the eigenvectors of $\mathbf{C_X} = \frac{1}{n}\mathbf{X}.\mathbf{X}^T$. Thus, *P* can be written as

$P = E^T$, and $P^{-1} = p^T$ and hence, we can re-write C_Y as:

$$\begin{aligned}\mathbf{C_Y} &= \mathbf{PC_XP}^T \\ &= \mathbf{P}(\mathbf{EDE}^T)\mathbf{P}^T \\ &= \mathbf{P}(\mathbf{P}^T\mathbf{DP})\mathbf{P}^T \\ &= (\mathbf{PP}^T)\mathbf{D}(\mathbf{PP}^T) \\ &= (\mathbf{PP}^{-1})\mathbf{D}(\mathbf{PP}^{-1}) \\ &= \mathbf{D}\end{aligned} \tag{3.17}$$

Thus, with the assumption of **P** as orthonormal, **P** can diagonalize $\mathbf{C_Y}$ and we can summarize PCA as:
1. The eigenvectors of $\mathbf{C_X} = \frac{1}{n}\mathbf{X}.\mathbf{X}^T$ are the principal components of **X** and the rows of **P**.
2. The diagonal value of $\mathbf{C_Y}$ corresponds to the variances of **X** along **P**.
3. First principal component indicates the largest variance and second to the second largest and so on.

Thus, with the help of the co-variance matrix $\mathbf{C_X}$, their eigenvalues and eigenvectors, one can compute the principal components of **X**.

3.4 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a machine learning technique based on statistical learning theory given in [Vapnik, 1995]. It maps the input data into high dimensional space called the feature space, and then it finds an optimal hyperplane to separate the two classes while maximizing generalization. The chosen hyperplane is such that it maximizes the distance (margin) between the plane and the nearest data points of the two classes called the support vectors. Unlike ANN, SVM does not suffer from local convergence as due to the quadratic optimization problem, it always finds a global minimum. The performance of SVM is better than conventional pattern recognition techniques such as ANN which works on the principle of empirical risk minimization (ERM) (reducing the training error). At the same time, SVM is based on structural risk minimization (SRM) (derived from statistical learning theory).

Suppose there is a set of training samples x which belongs to a class y, $\{(x_1,y_1), (x_2,y_2),...\}$ where $x \in R^n$ and $y \in \{-1,1\}$, the hyperplane is given as:

$$w^T x + b = \sum_{i=1}^{n} w^T x_i + b = 0 \tag{3.18}$$

where, w is a n-dimensional column vector and b is a scalar. The samples are said to be separated by this optimal decision boundary when the distance between the boundary and the nearest points or support vectors is maximum such that the following constraint is satisfied (shown in Figure 3.1):

$$y_i(w^T x_i + b) \geq 1 \tag{3.19}$$

This is achieved by minimizing the following parameter,

$$\arg\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.20}$$

which becomes a convex quadratic optimization problem that is solved using Lagrangian

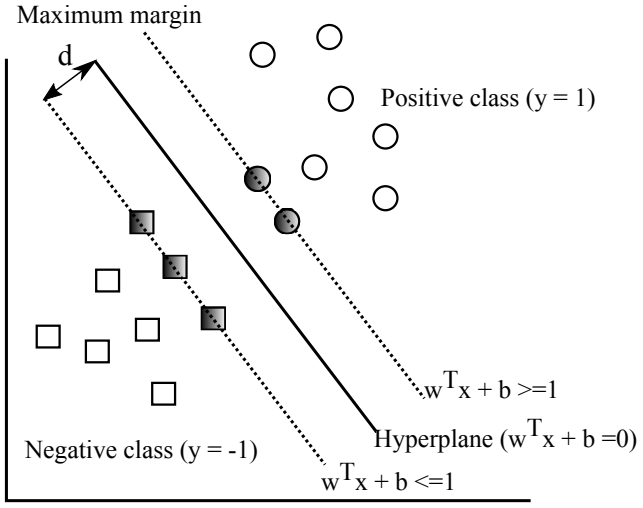

Figure 3.1 : Optimal hyperplane separating two classes with maximum margin from support vectors

multiplier method. This will lead to:

$$\min_{w} L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i}^{M} \alpha_i y_i (\mathbf{w}^T x_i + b) + \sum_{i}^{M} \alpha_i \tag{3.21}$$

such that $\forall \alpha_i \geq 0$ and α_i are Lagrangian multipliers. This problem is converted from minimizing **w** to maximizing the slack variable α_i by differentiating L w.r.t **w** and b and then replacing them in the equation to achieve the following equation:

$$\max_{\alpha} L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=0}^{n} \alpha_i \alpha_j y_i y_j x_i x_j \tag{3.22}$$

$$\text{s.t} \quad \alpha_i \geq 0, i = 1, 2, \ldots n, \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.23}$$

This gives the following non-linear decision function,

$$f(x) = \text{sign}\left(\sum_{i,j=0}^{M} \alpha_i y_i (\mathbf{x_i}.\mathbf{x_j}) \right) + b \tag{3.24}$$

In practice, for overlapping class distributions, a soft margin is provided to allow misclassification of some points with softly penalizing those points. To obtain this, slack variables $\xi_i \geq 0$ are introduced for each training sample s.t. $\xi_n = 0$ for data points on the boundary and $\xi = |y - (w^T \phi(x) + b)|$ i.e., $\xi_n \leq 1$ for points within the margin, but $\xi_n \geq 1$, when they are misclassified (i.e., on the wrong side of the boundary). This follows minimizing the following equation:

$$C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.25}$$

subject to, $y_i(w^T x_i + b) \geq 1 - \xi_i$. C is the cost parameter which controls the penalty for the points outside the margin.

In the case of non-linear classification, non-linear mapping is used and kernel functions are

employed to transform the input data into high-dimensional space. For a non-linear classifications, the decision function is written as,

$$f(x) = \text{sign}\left(\sum_{i,j=0}^{M} \alpha_i y_i \kappa(\mathbf{x}, \mathbf{x}') \right) + b \tag{3.26}$$

where, $\kappa(x, x')$ is the kernel function. The kernel functions can be of the form radial basis function, polynomial, sigmoid functions etc.

For multi-class SVM, One-Against-All (OAA) method can be been employed which considers one class at a time while combining all other classes.