# Unsupervised Feature Selection for Pattern Search in Seismic Time Series

**Andreas Köhler** AKOEHLER@UNI-POTSDAM.DE
**Matthias Ohrnberger** MAO@GEO.UNI-POTSDAM.DE
**Carsten Riggelsen** RIGGELSEN@GEO.UNI-POTSDAM.DE
**Frank Scherbaum** FS@GEO.UNI-POTSDAM.DE
*Institut für Geowissenschaften, Universität Potsdam*
*Karl-Liebknecht-Str. 24, 14476 Golm, Germany.*

**Editor:** Saeys et al.

## Abstract

This study presents an unsupervised feature selection approach for the discovery of significant patterns in seismic wavefields. We iteratively reduce the number of features generated from seismic time series by first considering significance of individual features. Significance testing is done by assessing the randomness of the time series with the Wald-Wolfowitz runs test and by comparing observed and theoretical variability of features. In a second step the in-between feature dependencies are assessed based on correlation hunting in feature subsets using Self-Organizing Maps (SOMs). We show the improved discriminative power of our procedure compared to manually selected feature subsets by cross-validation applied to synthetic seismic wavefield data. Furthermore, we apply the method to real-world data with the aim to define suitable features for earthquake detection and seismic phase classification in seismic recordings.

## 1. Introduction

Our study is motivated by classification and detection problems in seismology. Due to the high number of receiver networks monitoring earthquakes worldwide, a large amount of data is produced consisting of time histories of ground motion in different spatial directions. Automatic detection and classification of earthquakes is required in order to prepare data for investigation of the subsurface earth structure and to develop automatic warning systems e.g. at volcanos or to monitor the compliance with the nuclear test band treaty (CTBTO) (Joswig, 1990; Dai and MacBeth, 1995; Ohrnberger, 2001; Riggelsen et al., 2007). For these purposes, features are generated from the raw recordings. Since there are a lot of different, common approaches in seismology, it is not easy to define an optimal, discriminative and significant feature set. Thus, automatic feature selection is mandatory. In this study we use 7 common seismic feature generation methods which are all listed in Table 1. All in all we have a set of 159 features. A feature is computed for a short time window of the seismogram. We employ unsupervised learning techniques since seismologists often deal with unknown, complexly composed data. As a first learning step unsupervised feature selection will aid further processing.

Table 1: Seismic feature generation methods, features and number features.

| |
|---|
| **1 Frequency-wavenumber analysis** (Kvaerna and Ringdahl, 1986)<br>Spatial coherency (3 frequency bands and 3 spatial components): 9 features |
| **2 Spatial averaged autocorrelation method** (Aki, 1957)<br>Real and imaginary autocorrelation coefficient (3 frequency bands and 3 spatial components): 18 features |
| **3 Complex 3c-covariance matrix** (Vidale, 1986; Park et al., 1987; Jurkevics, 1988)<br>Several degree of polarization measures, ellipticity, angle of incidence (3 frequency bands): 39 features |
| **4 Complex seismic trace analysis** (Taner et al., 1979)<br>Instantaneous attributes (polarization, frequency, polarization directions, 3 frequency bands and 3 spatial components): 42 features |
| **5 Spectral attributes**<br>Normalized horizontal and vertical spectra (10 frequency bands), dominant frequency, bandwidth: 25 features |
| **6 Spectra of polarization ellipsoid** (Pinnegar, 2006)<br>Normalized semi-mayor and semi-minor axis of polarization ellipsoid (10 frequency bands): 20 features |
| **7 Amplitude ratios**<br>Real over imaginary part of complex trace, horizontal over vertical component (3 frequency bands): 6 features |

In general, for many applications the number of all potential features can be very high. However, the information content or relevance of individual features e.g. for clustering or imaging of patterns in the data may vary considerably. Furthermore, strong correlations between features will occult important information which is encoded in less or non redundant components of the feature vector. Thus, the computation time may be unnecessarily increased and the quality of the final results may suffer. Moreover, the higher the dimension of the data, the more data is needed for learning, and the less suitable is the euclidian distance as a measure of similarity, due to the curse of dimensionality (Bellman, 1961; Bishop, 2006). Furthermore, interpretation of the results is much easier for low number of features.

While a lot of approaches exist for supervised learning due to availability of labeled training data, for unsupervised learning feature selection is a more recent topic of research. Several approaches have been proposed to reduce the number of features, e.g. Principal Component Analysis (PCA). However, for PCA it is difficult to characterize the reduced data space since the (physical) meaning of the new features generated by linear combinations is unclear. Wrapper algorithms use a forward or backward selection procedure to search for the feature subset most relevant for clustering according to a particular evaluation criterion (Dy and Brodley, 2004). The computational complexity is very high for that approach, especially for high-dimensional data sets, since clustering has to be repeated for all potential subsets. In Basak et al. (1998) a fuzzy feature evaluation index for feature sets is used which does not require clustering. Feature selection is done by finding the feature subset with the smallest index. For a second method this evaluation index is minimized using a Neural Network approach in order to find the relative importance of individual features. For the first method still a search algorithm is necessary. A technique requiring no search is suggested by Mitra et al. (2002). This method reduces feature redundancy by grouping features based on a pairwise feature similarity measure called maximum information compression. Both approaches, Mitra et al. (2002) and Basak et al. (1998), are combined by Li et al. (2007) suggesting a two-level filter technique. Feature selection is done by first reducing redundancy and then assessing relevance for clustering of each feature using the fuzzy feature evaluation criterion.

Since an exhaustive wrapper search based on repeated clustering is not optimal for our real-world problem with up to 159 features, a filter approach for unsupervised feature selection is more promising. Furthermore, we also want to keep features that might show no clear cluster tendency but significant patterns in their time history, what is typical e.g. for earthquakes. Therefore, using a similar idea as Li et al. (2007) in this study, we introduce a multi-level feature selection procedure. We use significance testing using the Wald-Wolfowitz runs test (Wald and Wolfowitz, 1940) as a temporal context dependend feature relevance measure and Self-Organizing Maps (SOM) (Kohonen, 2001) for redundancy reduction.

Self-Organizing Maps is a popular and widespread unsupervised learning method. Especially for large data sets of high dimensions, SOMs allow an intuitive visualization of the data by vector quantization and dimension reduction. Based on the relatively simple SOM representation further processing like clustering or feature grouping can be done.

In Section 2 we give a more detailed introduction into the individual methods used. Section 3 presents our feature selection procedure in detail. We assess the reliability of our approach using synthetic and real-world data in Section 4.

## 2. Methods

In this section we introduce the techniques used for our feature selection procedure. We explain the Wald-Wolfowitz significance test and the Davies-Bouldin cluster validity index (Davies and Bouldin, 1979). Furthermore, we introduce the SOM learning algorithm.

### 2.1 Wald-Wolfowitz Runs Test

The Wald-Wolfowitz runs test can be used to assess the randomness of a two-valued time series by considering the distribution of runs. A “run” of a series is a maximal segment of adjacent equal elements (see background coloring in Fig. 1). In general, any time series can be transformed into a two-valued one by considering e.g. whether a data item is smaller or larger than the median of the series (Fig. 1). In order to find the features that show significant, non-random temporal patterns, we evaluate the test statistic of the runs test:

$$Z_{test} = \frac{r - \mathrm{E}[R]}{\sqrt{\mathrm{Var}[R]}} \ , \tag{1}$$

where $R$ is a random variable corresponding to the number of runs of a random time series which has the same length $N$ as the series of a particular feature under investigation. The variable $r$ is the number of observed runs for that feature. The mean:

$$\mathrm{E}[R] = \frac{2N^{-}N^{+}}{N} + 1 \ , \tag{2}$$

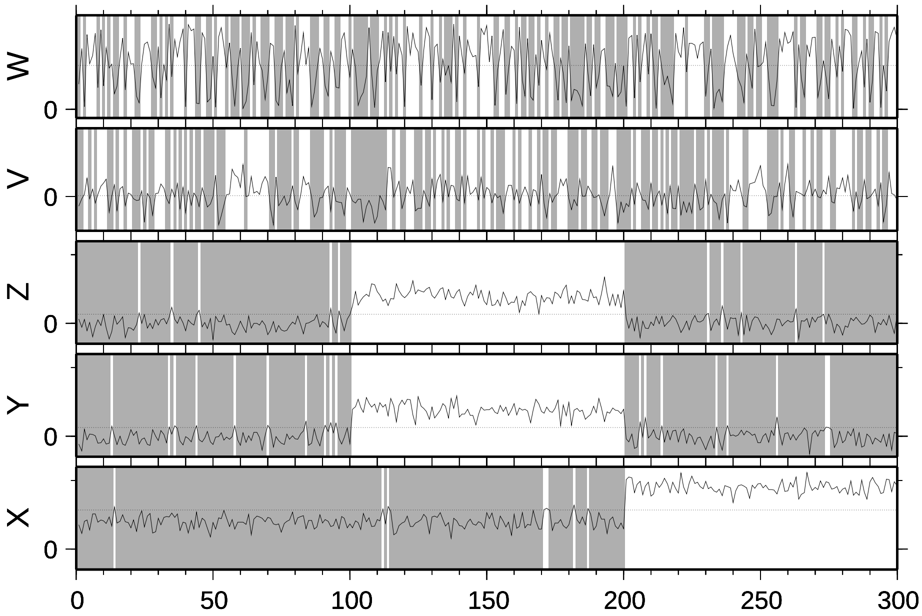

Figure 1: Demonstration of runs test for 5 time histories. Horizontal lines correspond to median. Background colorings highlight values above and below median. Whenever coloring changes with time, a new "run" is starting.

and the variance

$$\mathrm{Var}[R] = \frac{2N^-N^+(2N^-N^+ - N)}{N^2(N-1)} \ , \tag{3}$$

of $R$ is computed given the number of data items larger and smaller than the median ($N^+$ and $N^-$) considering the observed time series. Whenever the hypothesis of randomness is not rejected ($Z_{test} < 1.96$ for a significance level of 5%), the corresponding feature shows no significant patterns and therefore has no information content.

### 2.2 Cluster Validity

In order to validate that a particular clustering fits the natural grouping of the data, several quality criteria have been proposed (Halkidi et al., 2002). In the following we use the Davies-Bouldin (DB) index (Davies and Bouldin, 1979):

$$DB = \frac{1}{C}\sum_{k=1}^{C} \max_{l \neq k} \left\{ \frac{D_k + D_l}{d_{kl}} \right\} \ , \tag{4}$$

where $d_{kl}$ is the distance between cluster centroids $k$ and $l$, $D$ the average distance to the cluster centroid within a cluster and $C$ the number of clusters.

### 2.3 Self-Organizing Maps

The SOM learning algorithm combines vector quantization (generation of prototype vectors, black symbols in Fig. 2a) and an ordered, topology preserving mapping into a space of lower dimension

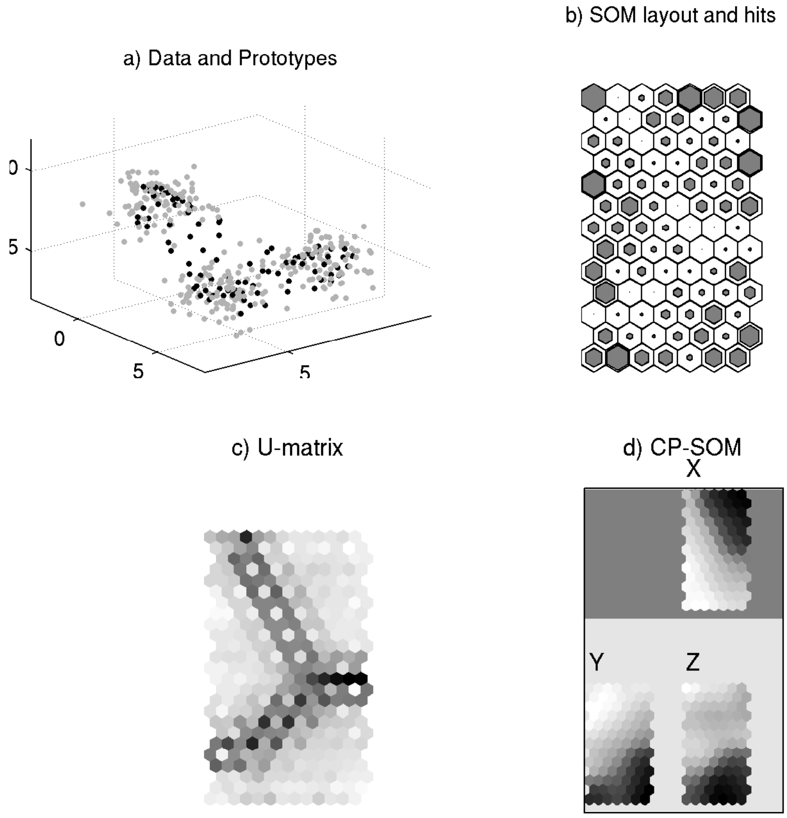

Figure 2: Example for Self-Organizing Maps applied to a simple 3D-data set.

(Fig. 2b). Usually, SOMs are built on a regular, hexagonal grid. Each grid unit $n$ is represented by a prototype vector $\vec{m}_n$. For each data sample $\vec{x}_t$ (gray symbols in Fig. 2a and 2b) the closest prototype vector $\vec{m}_c$ can be found, where $c$ is called the best matching unit (BMU). At each learning step $t$, the prototype vectors in the neighborhood of unit $c$ are moved towards the selected vector $\vec{x}_t$:

$$\vec{m}_n = \vec{m}_n + \alpha(t)h_{cn}(t)(\vec{x}_t - \vec{m}_n)\ , \tag{5}$$

where $h_{cn}(t)$ defines the Gaussian neighborhood around unit $c$ and $\alpha(t)$ is the learning rate, both decreasing with time. For more details see the SOM-Toolbox implemented in MATLAB® by Vesanto et al. (2000).

The SOM can be used to visualize high-dimensional data and therefore to identify and manually define clusters e.g. by showing the prototype distance between neighborhood SOM units (U-Matrix in Fig. 2c, black stands for high distances). Furthermore, since each SOM prototype vector itself can already be regarded as a cluster centroid, standard clustering algorithms can directly be applied on the set of all prototype vectors. In order to find the number of clusters, often the clustering algorithm is applied for different numbers of clusters. The best clustering is chosen according to the lowest Davies-Bouldin index (Davies and Bouldin, 1979; Vesanto and Alhoniemi, 2000).

In order to reduce redundancy in the data space (correlation hunting), SOMs can be used by considering the so-called component planes (CPs, overlaying panels in Fig. 2d, black stands for high values). A CP is built on the trained SOM ($N$ units) where each unit $n$ is represented by a particular component $i$ of the corresponding prototype vector $\vec{m}_n$. The components of the absolute correlation matrix $\mathbf{A}$ between all CPs is defined as:

$$a_{ij} = \frac{1}{N} \sum_{n=1}^{N} \|m_{ni} \cdot m_{nj}\|. \tag{6}$$

As proposed by Vesanto and Ahola (1999) the correlation matrix can be used as input data for the training of a second SOM on a rectangular grid. The data vector $\vec{x}_t$ is then defined as:

$$\vec{x}_t \overset{def}{=} \mathbf{a}_{\cdot j}, \tag{7}$$

where $\mathbf{a}_{\cdot j}$ is a column of $\mathbf{A}$. The so-called component plane SOM (CP-SOM) can be used to visualize intuitively correlation or similarity between components on a 2D-map (base map of Fig. 2d).

Correlated features can be grouped e.g. by clustering the CP-SOM using hierarchical clustering based on the distance matrix of the CP-SOM prototypes (Vesanto and Sulkava, 2002; Barreto, 2007) (coloring of base map in Fig. 2d). Guerif et al. (2005) propose a related method, where the features are weighted during SOM-training based on a simultaneous generated CP-SOM.

## 3. An Unsupervised Feature Selection Procedure

In the previous section we introduced different techniques which we will now combine for an unsupervised feature selection procedure. I order to keep significant features and reduce redundant information for a feature set generated by different approaches, we propose a three-level feature selection approach which iteratively reduces the number of features. The processing flow is illustrated in Fig. 3. In the first level we chose potential feature candidates by assessing the information content of each feature individually, while in the second and third level dependencies between features are considered. In the next sections we discuss each level in more detail.

### Level 1: Within Individual Features

In this level we first compute three criteria for each feature:

- Ratio $R_{exp}/R_{obs}$ between reasonably expected range $R_{exp}$ of a feature $f$ derived theoretically from physical or data processing parameters and observed variability $R_{obs} = \max(f) - \min(f)$.
- Wald-Wolfowitz test statistic $Z_{test}$ (equation 1).
- Lowest DB index (equation 4) computed from 1D-K-Means clusterings allowing 2 to $N_{clus}$ clusters (e.g. $N_{clus}$=5) .

The first two criteria are used to exclude features. We reject those features providing no significant discrimination between time windows due to small observed ranges ($R_{exp}/R_{obs} < r_{limit}$, e.g. $r_{limit} = 0.1$) and which show no significant temporal patterns ($Z_{test} < Z_{limit}$). For the amplitude features from generation methods 5, 6 and 7 no physical limits can be given. Therefore, $r_{limit} = 0$

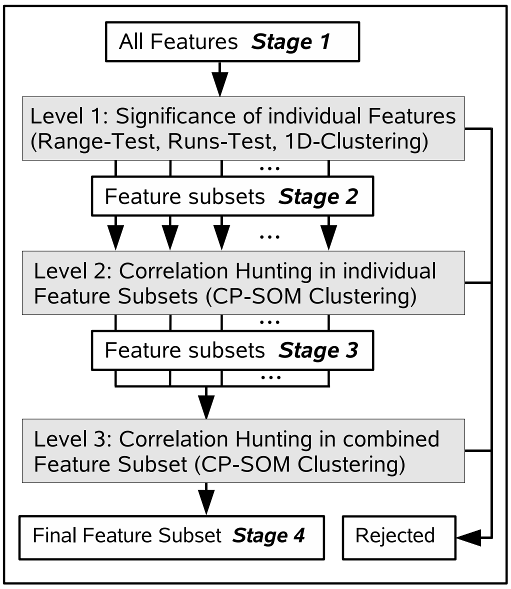
All Features Stage 1
Level 1: Significance of individual Features (Range-Test, Runs-Test, 1D-Clustering)
Feature subsets Stage 2
Level 2: Correlation Hunting in individual Feature Subsets (CP-SOM Clustering)
Feature subsets Stage 3
Level 3: Correlation Hunting in combined Feature Subset (CP-SOM Clustering)
Final Feature Subset Stage 4
Rejected

Figure 3: Three-level feature selection procedure. Stages 1-4 correspond to different feature subsets after or before particular processing steps. Feature subsets at stage 2 and 3 correspond to different feature generation methods.

(accepting all features) is used. As mentioned in Section 2.2, $Z_{limit} = 1.96$ is an appropriate threshold for the runs test. However, if the duration of expected temporal patterns is longer, increasing this value may improve the performance.

The DB index is used to assess the cluster tendency of the feature. This criterion is used together with $Z_{test}$ in the next level to rank features.

For more discussion on the sensitivity of parameters $N_{clus}$, $r_{limit}$ and $Z_{limit}$ see Section 4.2.

### Level 2: In-between Features of Individual Subsets

In the second level, we consider 7 feature subsets corresponding to the different feature generation methods (Table 1). Only features accepted by Level 1 are used. We first learn a SOM and afterwards a CP-SOM for each subset and then apply the CP-SOM clustering. From each CP-SOM cluster the features with the lowest DB index and the highest test statistic $Z_{test}$ are chosen as representative features for the particular cluster. Thus, we keep features with both, best cluster tendency and most significant temporal patterns. In case both features have the same BMU on the CP-SOM, only the latter one is selected.

### Level 3: In-between all Remaining Features

From Level 2 we get a reduced subset for each feature generation method. In the third level, we learn a single SOM and CP-SOM combining all subsets together in order to assess correlations between methods. Finally, we chose the features like in Level 2. The final set of features can then be used for further processing i.e. to learn the final SOM and to cluster the data set.

Table 2: Results of our feature selection method applied on a simple data example.

| Feature | $X$ | $Y$ | $Z$ | $V$ | $W$ |
|---|---|---|---|---|---|
| Observed runs $r$ | 70 | 77 | 78 | 149 | 154 |
| Runs test statistic $Z_{test}$ | 9.37 | 8.56 | 8.44 | 0.23 | 0.35 |
| DB Index | 0.35 | 0.46 | 0.47 | 0.63 | 0.56 |
| Selected after Level 1 | yes | yes | yes | no | no |
| Index of CP-SOM cluster | 1 | 2 | 2 | - | - |
| Selected after Level 2 | yes | yes | no | - | - |

Table 3: Results of feature selection based on a wrapper algorithm for a simple data example. Best feature subset (bold) is found when normed scatter separability criterion $S$ for the next subset becomes smaller.

| | $N_{clus} = 10$ | | $N_{clus} = 3$ | |
|---|---|---|---|---|
| Search Step | Feature Subset | sign $(S - S_{prev})$ | Feature Subset | sign $(S - S_{prev})$ |
| 1 | **V** | | X | |
| 2 | V, W | -1 | X, Y | +1 |
| 3 | V, W, X | -1 | **X, Y, Z** | **+1** |
| 4 | V, W, X, Y | +1 | X, Y, Z, W | -1 |
| 5 | V, W, X, Y, Z | -1 | X, Y, Z, W, V | 0 |

### Simple Example

In Table 2 we demonstrate our feature selection procedure using a simple data set of 5 features ($N = 300$). Values for features $X$, $Y$ and $Z$, together defining three clusters, can be found in Fig. 2a. Features $Y$ and $Z$ are strongly correlated. The data for the remaining features $V$ and $W$ are drawn from a Normal and from a Uniform distribution, respectively. The temporal context of all features is given in Fig. 1. We omit the range test in Level 1 and only use a single subset (no Level 3) because the feature have no physical background.

Features $V$ and $W$ are correctly rejected by the runs test ($Z_{test} < 1.96$) because of their temporal randomness. The result of CP-SOM clustering is shown in Fig. 2d. Features $Y$ and $Z$ belong to the same CP-cluster. Thus, features $X$ and $Y$, the second one due to the higher runs test statistic $Z_{test}$ and DB index, are finally selected.

We also test a wrapper approach for feature selection using the same feature set (Table 3). The forward search based on a normalized cluster scatter separability criterion as proposed by Dy and Brodley (2004) results in a best feature subset including $V$ ($N_{clus} = 10$) or $X, Y$ and $Z$ ($N_{clus} = 3$), respectively. Thus, only for the second run the random features are correctly rejected. However, no redundancy reduction is obtained and maximum number of clusters has to be limited to avoid overfitting.

## 4. Experiments

We conduct experiments using both, synthetic seismic data and real earthquake recordings. Synthetic data is used to validate the feature selection procedure, while real-world data is employed to show the potentials of the method for seismic wave phase detection.

### 4.1 Synthetic Data

In order to assess the validity and performance of the feature selection procedure, we apply a 10-fold cross-validation technique (Dy and Brodley, 2004) on synthetic seismic network data. The validation is based on hierarchical clusterings of SOM prototype vectors. The data set consists of 4 classes corresponding to 3 different types of seismic waves (Rayleigh waves, Love waves, mixture of both: class 1–3) and random noise in between (class 4). The class labels are only used for the error computation.

Level 1 of the feature selection procedure ($N_{clus} = 5$, $r_{limit} = 0.1$, $Z_{limit} = 1.96$) is applied on the complete data set (training and test data) in order to keep the temporal context for the runs test. After feature selection, SOM training and clustering using the training data, each cluster is classified with respect to the most frequent class label within. For the testing we compute the BMUs, and thus the cluster-memberships, of the test data set on the training data set SOM. A class error is computed as the percentage of misclassified data compared to the total number of samples of the test data set for each fold. Finally, the (mean) cross-validated classification error (CVCE, Dy and Brodley (2004)) is calculated.

In order to quantify the improvements made by our new feature selection approach, we compute the CVCE for several feature subsets obtained at four stages of the procedure (see Fig. 3) and for particular feature generation methods (Table 1). It should be noted, that we do not expect to achieve a CVCE tending to zero, since the transition between seismic wave types and noise can be continuous, although we introduced a threshold for the class labeling.

Considering the overall trend for each feature generation method in Table 4, the classification errors slightly decrease with number of features and therefore with stage of feature selection. Furthermore, comparing the methods, the CVCE decreases significantly when all feature generation methods are combined at each stage compared to the individual feature subsets. Focussing on individual methods, method 5 (spectral features) seems to provide the best discriminative power for clustering. For method 7 (amplitude ratios) and method 6 (spectra of polarization ellipsoid) the CVCE increases at stage 3. The best performance (15.8%) is achieved with about 57 features from all methods at stage 3. However, after assessing correlation between all feature generation methods at stage 4, the CVCE is still within the range of standard deviations of stages 1 to 3 for the combination of methods. Due to the relatively simple synthetic wavefield, most features show significant patterns and are therefore accepted in feature selection Level 1. However, assessing the correlations between features in Level 2 and 3, significantly reduced the set of features for all feature generation methods. The reduction in Level 2 and 3 does not worsen the classification rate, except for feature generation methods 6 and 7, where probably the number of features becomes too low.

From the cross-validation we conclude that it is sufficient to consider only the finally reduced feature subset combining features from different methods (stage 4). The dimensionality, and therefore computation time and model complexity, is reduced considerably for further analysis of the data set, without significantly losing discriminative power.

Table 4: Results of cross-validation for a synthetic seismic wavefield. Cross-validated Classification Error (CVCE) and Averaged number of features for different stages of feature selection and different feature generation methods (see Fig. 3 and Table 1).

| Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | all |
|---|---|---|---|---|---|---|---|---|
| | | | | **Percent CVCE** | | | | |
| Stage 1 | 29.7 | 45.8 | 30.1 | 25.2 | 22.5 | 35.0 | 31.4 | 17.1 |
| | ±9.4 | ±10.3 | ±11.9 | ±12.0 | ±7.7 | ±8.3 | ±11.4 | ±4.3 |
| Stage 2 | 29.7 | 45.8 | 30.1 | 23.2 | 20.0 | 34.0 | 31.4 | 16.2 |
| | ±9.4 | ±10.3 | ±11.9 | ±11.0 | ±7.5 | ±5.6 | ±11.4 | ±4.9 |
| Stage 3 | 27.6 | 36.9 | 25.8 | 22.4 | 21.5 | 41.5 | 39.2 | 15.8 |
| | ±7.0 | ±6.2 | ±10.2 | ±8.3 | ±7.3 | ±7.0 | ±6.4 | ±5.9 |
| Stage 4 | - | - | - | - | - | - | - | 16.9 |
| | - | - | - | - | - | - | - | ±5.6 |
| | | | | **Averaged Number of Features** | | | | |
| Stage 1 | 9 | 18 | 39 | 42 | 25 | 20 | 6 | 159 |
| Stage 2 | 9 | 18 | 39 | 37 | 24 | 15 | 6 | 148 |
| Stage 3 | 5.0±0.0 | 6.7±1.5 | 12.5±1.8 | 14.5±2.3 | 10.5±1.6 | 6.5±1.2 | 2.9±0.5 | 57.9±3.9 |
| Stage 4 | - | - | - | - | - | - | - | 22.2±2.3 |

### 4.2 Real-world Data

In this section we apply our procedure to earthquake recordings in order to find suitable features which allow to detect the temporal onset of an event, and also to distinguish between different phases of arriving waves. First, we use three similar events which were recorded at the same receiver and occurred at different times in the same source region. In Fig. 4 for one event the three-component seismogram is shown. The labels and the background coloring on top indicate different wave phases which can be identified using theoretical arrival times and expert knowledge of seismologists. Except of generation method 1 and 2, which require more than one receiver, all features are computed and the feature selection procedure is applied ($N_{clus} = 5$, $r_{limit} = 0.2$, $Z_{limit} = 4.0$).

Our feature selection procedure finds 9 features out of a set of 129:

- Normalized horizontal spectra for frequency bands 6, 8 and 9.
- Normalized vertical spectra for frequency band 2.
- Planarity of polarization for frequency band 3.
- Component-averaged instantaneous frequency for frequency band 3.
- Normalized semi-minor axis of polarization ellipsoid for frequency band 1 and 3.
- Difference of semi axis of polarization ellipsoid for frequency band 1.

Table 5: Classification errors and discriminative power for real-world data using all features and subsets (feature generation methods), with and without applying feature selection. $S$ is computed with respect to complete feature selection using all features ($S_{FSall}$). No feature from method 7 passed Level 1. Instead results for a random feature set are shown within the lower panel.

| **No Features Selection** | | | | | | |
|---|---|---|---|---|---|---|
| | All | Meth. 3 | Meth. 4 | Meth. 5 | Meth. 6 | Meth. 7 |
| $CE^{+}_{final}$ | 0.16 | 0.12 | 0.15 | 0.23 | 0.16 | 0.16 |
| $CE^{-}_{final}$ | 0.19 | 0.15 | 0.19 | 0.25 | 0.11 | 0.21 |
| $S_{FSall} - S$ | 171.3 | -37.7 | 46.8 | 72.7 | 5.1 | 4.9 |
| Number of features | 129 | 39 | 36 | 25 | 20 | 9 |
| **Features Selection** | | | | | | |
| | All | Meth. 3 | Meth. 4 | Meth. 5 | Meth. 6 | Random |
| $CE^{+}_{final}$ | 0.06 | 0.22 | 0.13 | 0.06 | 0.15 | 0.18 |
| $CE^{-}_{final}$ | 0.10 | 0.29 | 0.10 | 0.09 | 0.15 | 0.24 |
| $S_{FSall} - S$ | 0.0 | 1.2 | 6.8 | 18.2 | 12.2 | 52.8 |
| Number of features | 9 | 2 | 2 | 8 | 8 | 9 |

Finally, a SOM is trained using the selected features each weighted with its $Z_{test}$ value. For a quantitative evaluation of our method, we compute classification errors (false positive and false negative) and a measure for discriminative power. For this purpose, we use the theoretical class labels (Pn, Pg, Sn and Sg phases, coda of event, noise). The most frequent class label, resulting from the projecting of the labeled data on the SOM, is assigned to each SOM unit. Ambiguous units (same number of BMU hits for two or more classes) are counted. First, classification errors $CE_k$ are computed for individual classes $k$. In case a class is not present on the SOM after labeling, $CE_k$ is set to 1. Finally, the mean over all classes ($CE$) is penalized by the ratio $R_{amb}$ between number of ambiguous and all SOM units:

$$CE_{final} = CE + (1 - CE) \cdot R_{amb} \; . \tag{8}$$

The discriminative power between wave phases is measured by the normed scatter separability criterion $S$ (Dy and Brodley, 2004) for the data clustering given by the BMUs and their class labels on the SOM (without noise). Note, that $S$ is a relative measure which is computed with respect to a second clustering.

The first column of Table 5 shows that the best discrimination of wave phases in terms of $S$, $CE^{+}_{final}$ (false positive) and $CE^{-}_{final}$ (false negative) is obtained after feature selection with a feature vector of significant lower dimension (9). For comparison, using 9 randomly selected features results in clearly higher classification errors. Considering the feature generation methods separately, classification errors for method 5 are similar compared to feature selection using all methods, although discriminative power is worse. Thus, features representing the time-frequency content of the wavefield seem to be most suitable. This confirms classification approaches in seismology like Joswig (1990) and Riggelsen et al. (2007) where spectral features are employed a priori.

Table 6: Sensitivity of parameters of feature selection algorithm. $S$ is computed with respect to $N_{clus} = 5$, $r_{limit} = 0.2$ and $Z_{limit} = 4.0$.

| | | **Number of Clusters** | | | | | | **Range Limit** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $N_{clus}$ | 2 | 5 | 10 | 15 | 20 | $r_{limit}$ | 0.0 | 0.1 | 0.2 | 0.5 |
| $CE^{+}_{final}$ | 0.06 | 0.06 | 0.10 | 0.8 | 0.8 | $CE^{+}_{final}$ | 0.13 | 0.13 | 0.06 | 0.17 |
| $CE^{-}_{final}$ | 0.11 | 0.10 | 0.14 | 0.10 | 0.10 | $CE^{-}_{final}$ | 0.09 | 0.09 | 0.10 | 0.20 |
| $S_{FS} - S$ | 22.6 | 0.0 | 16.9 | 7.9 | 7.9 | $S_{FS} - S$ | 12.9 | 12.9 | 0.0 | 4.0 |
| | | | | | **Runs Test Limit** | | | | | |
| $Z_{limit}$ | 1.96 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $CE^{+}_{final}$ | 0.10 | 0.09 | 0.06 | 0.09 | 0.07 | 0.06 | 0.12 | 0.08 | 0.04 | 0.07 |
| $CE^{-}_{final}$ | 0.19 | 0.13 | 0.10 | 0.11 | 0.08 | 0.05 | 0.10 | 0.06 | 0.04 | 0.07 |
| $S_{FS} - S$ | 20.2 | 13.5 | 0.0 | 21.7 | 26.1 | 15.6 | 3.7 | 23.9 | -1.0 | 5.9 |

Table 7: Sensitivity of time window length given by parameter $WINFAC$.

| $WINFAC$ | 1 | 2 | 4 | 6 | 8 |
|---|---|---|---|---|---|
| $CE^{+}_{final}$ | 0.24 | 0.20 | 0.06 | 0.07 | 0.36 |
| $CE^{-}_{final}$ | 0.23 | 0.30 | 0.10 | 0.08 | 0.36 |

## SENSITIVITY TESTS

In Table 6 we show the sensitivity of parameters $N_{clus}$, $r_{limit}$ and $Z_{limit}$ by changing one parameter while keeping constant the other two. The optimal values for our problem are $N_{clus} = 5$ and $r_{limit} = 0.2$. Increasing $N_{clus}$ seems to lead to overfitting of the data. Thus, features are selected which do not represent the actual clustering of the data. When more classes are expected, increasing $N_{clus}$ may improve results. For $Z_{limit}$ definition of an optimal value is not so clear. For $Z_{limit} > 3$ classification errors are slightly lower and quite similar. The best results are obtained for $Z_{limit} = 4$ and, slightly better, $Z_{limit} = 10$. Theoretical tests of the runs test show that $Z_{test}$ of a non-random time series ($Z_{test} > 1.96$) depends on data length and number and duration (period) of patterns. For our problem we expect a minimum pattern length of 2 samples what corresponds to relative low values ($1.96 < Z_{test} < 5$). Thus, in order to ensure that we capture all possible patterns and also to consider the test results, we use $Z_{limit} = 4$ for our investigations. However, in general, when no a priori information and is available, a value corresponding to an appropriate significance level should be used (e.g. $Z_{limit} = 1.96$ for 5%). Furthermore, we test different cluster validity criteria (Halkidi et al., 2002) instead of the DB index in feature selection Level 1. The best performance is achieved using the DB index. For instance for the $S_{Dbw}$ index (Halkidi et al., 2002) $CE^{+}_{final} = 0.16$ and $CE^{-}_{final} = 0.18$ are obtained.

Another important parameter ($WINFAC$) is related to feature generation. The sensitivity of the time window length, for which a feature is computed, is investigated in Table 7. Window length is given by $WINFAC \cdot 1/f_{cent}$, where $f_{cent}$ is the center frequency of the overall frequency band

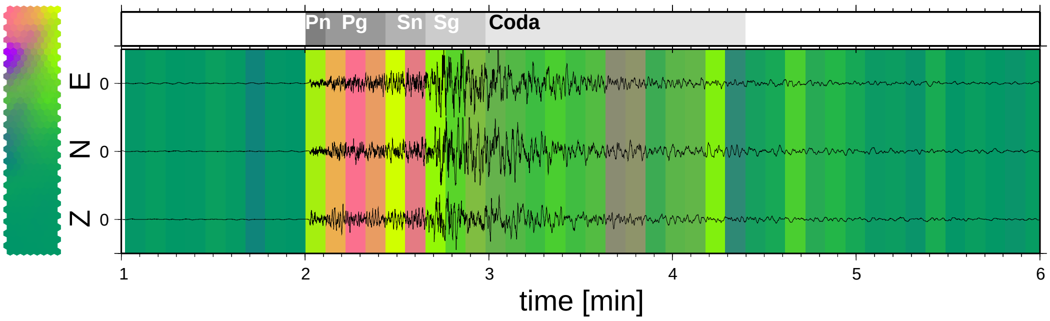

Figure 4: Time histories of all three spatial components for an earthquake record. On top different wave phases are indicated. On left hand side a SOM is shown trained after feature selection. Background coloring of seismograms corresponds to SOM coloring which is based on prototype vector similarity.

we consider for feature generation. Thus, to ensure that at least one period of the signal is present in a window, $WINFAC$ should be higher than 1. We find that $WINFAC = 6$ is optimal for wave phase discrimination and feature stability.

Fig. 4 shows a SOM visualization corresponding to the final feature set using $Z_{limit} = 10$. A color scale is spread out on top of the SOM based on prototype vector similarity. Thus, SOM units of similar prototype vectors have similar colorings. Considering the color scale of Fig. 4 as background color for the seismogram, the onset of the earthquake (Pn) as well as the different phases are clearly highlighted as different SOM units.

### CROSS-VALIDATION

In a last step we investigate the generalization capability of our procedure for a larger data set of 44 different earthquakes (Table 8, $Z_{limit} = 1.96$). We carry out a 44-fold cross-validation by leaving out one event at each fold. The previous definition for classification error (equation 8) is used. SOM labeling is done only for three classes (P wave, S wave and noise) since we are not able to identify all weak phases for all events. We obtain similar results with a slight improvement of 1-3% for the cross-validated errors $CVCE_{final}$ compared to SOM training without feature selection. However, number of features is reduced significantly and features of different generation approaches are combined. For comparison, using only features from the most common seismological approach (method 5), yields clearly higher classification errors. Again, random feature sets having dimensions similar to the best sets result in higher errors. Thus, our procedure finds the best combination of features and significantly reduces model complexity.

## 5. Conclusions

In this paper, we introduced an unsupervised feature selection procedure for seismic wavefield recordings. The features are computed from different seismic feature generation methods. The technique is based on a combination of significance testing for individual features and correlation analysis using Self-Organizing Maps for feature subsets.

Table 8: Cross-validation results for data set of 44 earthquakes using features selection (FS), all feature (noFS), features from generation approach 5 and random feature sets.

| | FS | noFS | Method 5 | Random |
|---|---|---|---|---|
| Percent $CVCE^{+}_{final}$ | 33.2 ± 1.7 | 34.6 ± 1.9 | 41.7 ± 1.5 | 43.5 ± 1.9 |
| Percent $CVCE^{-}_{final}$ | 33.1 ± 1.7 | 36.1 ± 1.2 | 41.3 ± 1.0 | 45.3 ± 0.9 |
| Number of features | 20 ± 6 | 136 | 26 | 20 |

We applied the procedure on a synthetic seismic wavefield. Cross-validating SOM-based clusterings obtained from automatically selected feature subsets showed that the best performance, considering classification error and model complexity, can be achieved with the finally selected features.

Experiments on real-world data were carried out to test feature selection for earthquake detection and wave type discrimination. By comparing classification errors for a data set of three similar events, we found that the final set of 9 features provided better discrimination between seismic wave types than using all potential features. We showed that features most suitable are those representing the time-frequency content of the seismogram. Furthermore, sensitivity of the algorithm parameters was tested. We found that a priori knowledge about number of classes and duration of temporal patterns can improve results. An optimal time window length for feature generation could be given. Furthermore, we investigated the generalization capability of our procedure for a larger earthquake data set using cross-validation. A feature set of significant lower dimension is obtained without increasing mean classification errors compared to the complete feature set. In comparison with a classical approach in seismology, results could be improved.

We suggest our approach as a first learning step for advanced supervised learning techniques which rely on large, multi-dimensional time series data sets. Features selected from seismic recordings including different types of earthquakes, mining events (explosions) and other transient phenomena can be used to train e.g. context dependent learning methods like Dynamic Bayesian Networks (Riggelsen et al., 2007) which are able to classify event type and to detect seismic phases.

### Acknowledgments

Andreas Köhler is financed by a scholarship of the University of Potsdam. Parts of this work has been made possible under the EU-Project NERIES contract no. 026130.

## References

K. Aki. Space and time spectra of stationary stochastic waves, with special reference to microtremors. *Bulletin of the Earthquake Research Institute, University of Tokyo,* 35:415–456, 1957.

A. Barreto, M.A. Pérez-Uribe. Improving the Correlation Hunting in a Large Quantity of SOM Component Planes. *Lecture Notes in Computer Science: Artificial Neural-Networks–ICANN 2007*, 4669:379–288, 2007.

J. Basak, R.K. De, and S.K. Pal. Unsupervised feature selection using a neuro-fuzzy approach. *Pattern Recognition Letters*, 19(11):997–1006, 1998.

R.E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

C.M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

H. Dai and C. MacBeth. Automatic picking of seismic arrivals in local earthquake data using an artificial neural network. *Geophysical journal international*, 120(3):758–774, 1995.

D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

J.G. Dy and C.E. Brodley. Feature Selection for Unsupervised Learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

S. Guerif, Y. Bennani, V. France, E. Janvier, N.C. France, and B.B. France. $\mu$-SOM: Weighting features during clustering. *Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM 05)*, pages 397–404, 2005.

M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods: Part I and II. *SIGMOD Record*, 31(2):40–45, 2002.

M. Joswig. Pattern recognition for earthquake detection. *Bulletin of the Seismological Society of America*, 80(1):170–186, 1990.

A. Jurkevics. Polarization analysis of three-component array data. *Bulletin of the Seismological Society of America*, 78(5):1725–1743, 1988.

T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

T. Kvaerna and F. Ringdahl. Stability of various fk estimation techniques. *Semianual technical summary, 1 October 1985 - 31 March 1986, NORSAR Scientific Report, Kjeller, Norway*, 1-86/87: 29–40, 1986.

Y. Li, B.L. Lu, and Z.F. Wu. Hierarchical fuzzy filter method for unsupervised feature selection. *Journal of Intelligent and Fuzzy Systems*, 18(2):157–169, 2007.

P. Mitra, CA. Murthy, and S.K. Pal. Unsupervised Feature Selection Using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, 2002.

M. Ohrnberger. *Continuous Automatic Classification of Seismic Signals of Volcanic Origin at Mt. Merapi, Java, Indonesia*. PhD thesis, University of Potsdam, 2001.

J. Park, F.L. Vernon III, and C.R. Lindberg. Frequency dependent polarization analysis of high-frequency seismograms. *Journal of Geophysical Research*, 92(B12):12664–12674, 1987.

CR. Pinnegar. Polarization analysis and polarization filtering of three-component signals with the time-frequency S transform. *Geophysical Journal International*, 165(2):596–606, 2006.

C. Riggelsen, M. Ohrnberger, and F. Scherbaum. Dynamic Bayesian Networks for Real-Time Classification of Seismic Signals. *Lecture Notes in Computer Science*, 4702:565–572, 2007.

MT. Taner, F. Koehler, and RE. Sheriff. Complex seismic trace analysis. *Geophysics*, 44:1041–1063, 1979.

J. Vesanto and J. Ahola. Hunting for Correlations in Data Using the Self-Organizing Map. *Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA 99), Rochester, NY*, 1999.

J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on*, 11(3):586–600, 2000.

J. Vesanto and M. Sulkava. Distance Matrix Based Clustering of the Self-Organizing Map. *Proc. International Conference on Artificial Neural Networks–ICANN 2002*, pages 951–956, 2002.

J. Vesanto et al. *SOM Toolbox for Matlab 5*. Helsinki University of Technology, 2000.

J.E. Vidale. Complex polarization analysis of particle motion. *Bulletin of the Seismological Society of America*, 76(5):1393–1405, 1986.

A. Wald and J. Wolfowitz. On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.