

Unsupervised Feature Selection for Pattern Discovery in Seismic Wavefields

Andreas Köhler, Matthias Ohrnberger, Carsten Riggelsen, and Frank Scherbaum

Institut für Geowissenschaften, Universität Potsdam,
Karl-Liebknecht-Str. 24, 14476 Golm, Germany
akoehler@uni-potsdam.de,
{mao, riggelsen, fs}@geo.uni-potsdam.de

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.

Key words: Unsupervised Learning, Feature Selection, Self-Organizing Maps, Seismology

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 ban treaty (CTBT) [1–4]. 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 (6.5

s). 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 [11] |
| Spatial coherency (3 frequency bands and 3 spatial components): 9 features |
| 2 Spatial averaged autocorrelation method [12] |
| Real and imaginary autocorrelation coefficient (3 frequency bands and 3 spatial components): 18 features |
| 3 Complex 3c-covariance matrix [13–15] |
| Several degree of polarization measures, ellipticity, angle of incidence (3 frequency bands): 39 features |
| 4 Complex seismic trace analysis [16] |
| Instantaneous attributes (polarization, frequency, polarization directions, 3 frequency bands and 3 spatial components): 42 features |
| 5 Spectral attributes |
| Normalized horizontal and vertical spectra (10 frequency bands), dominant frequency, bandwidth: 25 features |
| 6 Spectra of polarization ellipsoid [17] |
| Normalized semi-major and semi-minor axis of polarization ellipsoid (10 frequency bands): 20 features |
| 7 Amplitude ratios |
| 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 [5, 6]. 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 criteria [7]. 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 [8] 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 [9]. This method reduces feature redundancy by grouping features based on a pairwise feature similarity measure called maximum information compression. Both approaches, [9] and [8], are combined by [10] suggesting a two-level filter technique. Feature selection is done by first reducing redundancy and then assessing relevance for clustering of each feature using a fuzzy feature evaluation criteria.

Since an exhaustive wrapper search based on repeated clustering is not optimal for our real-world problem with up to 163 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 pattern in their time history, what is typical e.g. for earthquakes. Therefore, using a similar idea as [10] in this study, we introduce a multi-level feature selection procedure. We use significance testing using the Wald-Wolfowitz runs test [18] as a temporal context dependent feature relevance measure and Self-Organizing Maps (SOM) [19] for redundancy reduction.

Self-Organizing Maps (SOMs) [19] 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 e.g. clustering or feature grouping can be done and easily be validated.

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 clustering of synthetic data and apply it to real-world data in Section 4.

2 Methods

In this section we introduce different techniques which are used for our feature selection procedure. We explain the significance test developed by Wald and Wolfowitz [18] and the Davies-Bouldin cluster validity index [20]. Furthermore, we introduce the Self-Organizing Maps learning algorithm [19].

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

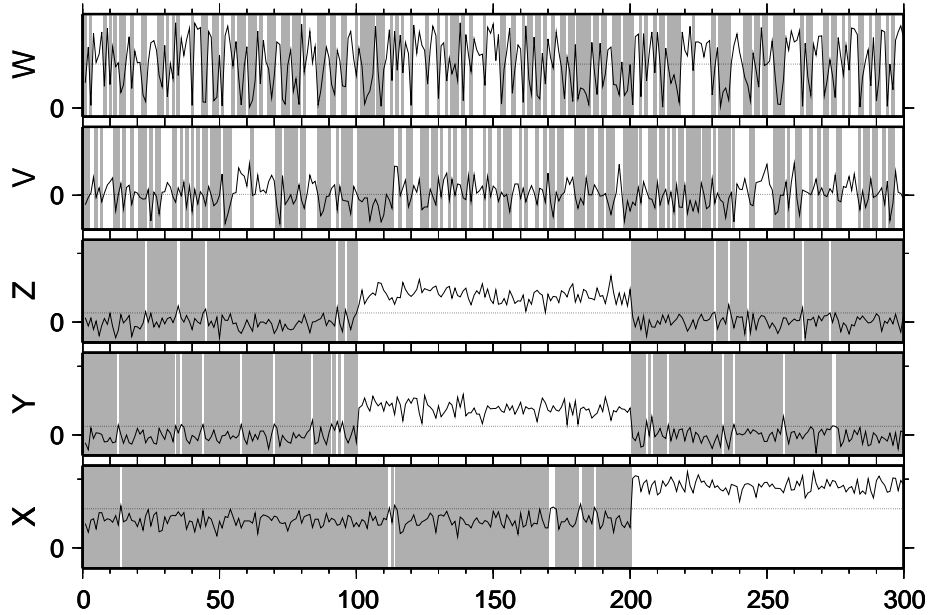


Fig. 1. Demonstration of runs test for 5 time histories. Horizontal lines correspond to median. Background colorings highlight values above and below median. Whenever coloring changes with time, a new “run” is starting.

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 (see dotted lines in Fig. 1). In order to find the features that show significant, non-random temporal patterns, we evaluate the test statistic of the runs test:

$$Z_{test} = \frac{r - E[R]}{\sqrt{\text{Var}[R]}}, \quad (1)$$

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:

$$E[R] = \frac{2N^-N^+}{N} + 1, \quad (2)$$

and the variance

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

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 pattern and therefore has no information content.

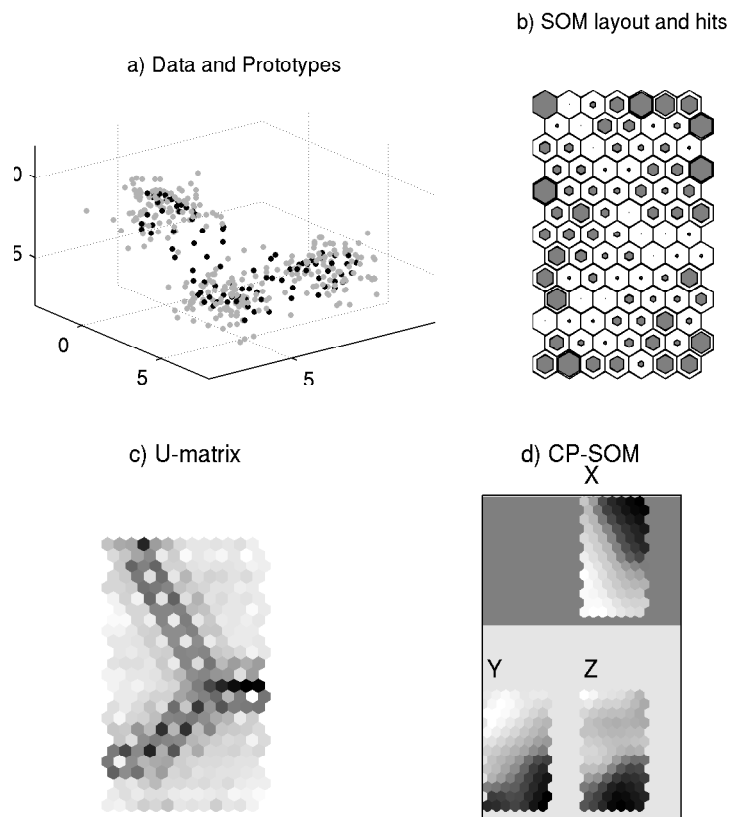


Fig. 2. Example for Self-Organizing Maps applied to a simple 3D-data set.

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. In the following we use the Davies-Bouldin (DB) index [20]:

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

where d is the distance between cluster centroids, 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 (Fig. 2b). Usually, SOMs are built on a regular, hexagonal grid. Each grid unit n is represented by a prototype vector \mathbf{m}_n . For each data sample \mathbf{x}_t (gray symbols in Fig. 2a and 2b) the closest prototype vector \mathbf{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 \mathbf{x}_t :

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

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 [21].

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 [20, 22].

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 \mathbf{m}_n . The components of the absolute correlation matrix \mathbf{A} between all CPs is defined as:

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

As proposed by [23] the correlation matrix can be used as input data for the training of a second SOM on a rectangular grid. The data vector \mathbf{x}_t is then defined as:

$$\mathbf{x}_t \stackrel{def}{=} \mathbf{a} \cdot \mathbf{j}, \quad (7)$$

where $\mathbf{a} \cdot \mathbf{j}$ 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 [24, 25] (coloring of base map in Fig. 2d).

3 An Unsupervised Feature Selection Procedure

In the previous section we introduced different technique which we will now combine for an unsupervised feature selection procedure. In order to keep significant features and reduce redundant information for a feature set generated by

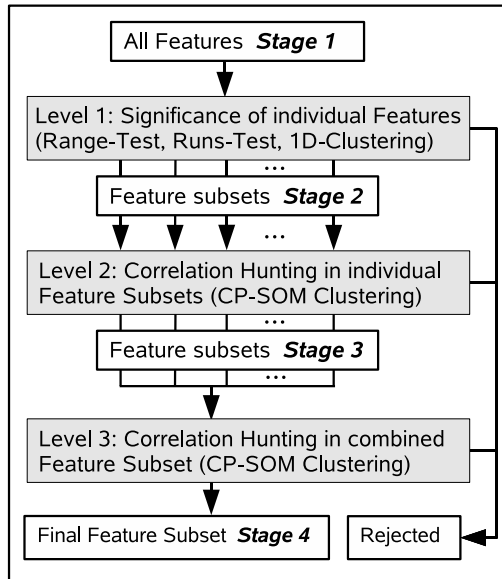


Fig. 3. Three-level feature selection procedure. Stages 1-4 correspond to different feature subsets after or before particular processing steps. Feature subsets at stage 2 and 3 correspond to different feature generation methods.

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 pattern ($Z_{test} < Z_{limit}$). We found empirically that for amplitude features

from generation methods 5, 6 and 7 $r_{limit} = 0$ (accepting all features) is optimal. 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 pattern is longer, increasing this value may improve the performance.

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

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.

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. The forward search based on a normalized cluster scatter separability criterion as proposed by [7] results in a best feature subset including X , Y and Z , thus, correctly rejecting the random features. However, no redundancy reduction is obtained and maximum number of clusters has to be set to 3.

Table 2. Results of Feature selection 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 | - | - |

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 on a synthetic data set [7]. The validation is based on hierarchical clusterings of synthetic seismic waveforms computed for different receivers. 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 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, [7]) 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 3, 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 pattern 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 3. 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 ±9.4 | 45.8 ±10.3 | 30.1 ±11.9 | 25.2 ±12.0 | 22.5 ±7.7 | 35.0 ±8.3 | 31.4 ±11.4 | 17.1 ±4.3 |
| Stage 2 | 29.7 ±9.4 | 45.8 ±10.3 | 30.1 ±11.9 | 23.2 ±11.0 | 20.0 ±7.5 | 34.0 ±5.6 | 31.4 ±11.4 | 16.2 ±4.9 |
| Stage 3 | 27.6 ±7.0 | 36.9 ±6.2 | 25.8 ±10.2 | 22.4 ±8.3 | 21.5 ±7.3 | 41.5 ±7.0 | 39.2 ±6.4 | 15.8 ±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

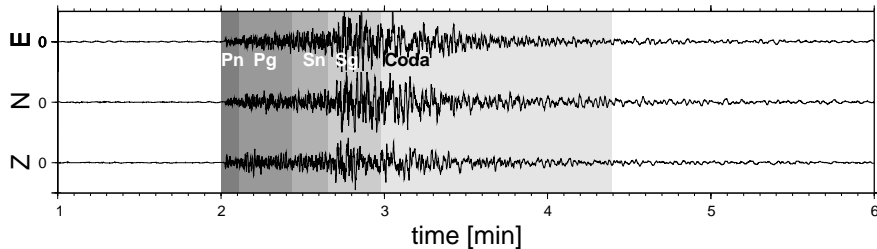


Fig. 4. Time histories of all three spatial components for an earthquake record. Background coloring corresponds to theoretical phase labels.

distinguish between different phases of the arriving waves. 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 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 ($N_{clus} = 5$, $r_{limit} = 0.2$, $Z_{limit} = 4.0$).

Finally, a SOM is trained using these 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 discrimination power using the 6 theoretical class labels (Pn, Pg, Sn and Sg phases, coda of event, noise). For this purpose, 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 CV_k are computed for individual classes k . In case a class is not present on the SOM after labeling, CV_k is set to 1. Finally, the mean over all classes (CV) is penalized by the ratio R_{amb} between number of ambiguous and all SOM units:

$$CV_{final} = CV + (1 - CV) * R_{amb} . \quad (8)$$

The discrimination power S is computed using the normed scatter separability criteria [7] for the SOM prototype vector clusters given by the 6 class labels.

Considering S , CV_{final}^+ (false positive) and CV_{final}^- (false negative), Table 4 shows that the wave phases are clearly better discriminated with the feature vector of significant lower dimension compared to the complete feature set. In this context, we also tested the influence of parameters N_{clus} , r_{limit} , Z_{limit} , choice of clustering index and length of time window for feature generation. Results for classification errors show that we used the optimal values. For detailed results and discussion on the sensitivity of parameters see the JMLR version of this paper.

Table 4. Results of Feature selection applied on real-world data.

| | no Feature selection | Feature selection |
|--------------------|----------------------|-------------------|
| CV_{final}^+ | 0.16 | 0.06 |
| CV_{final}^- | 0.19 | 0.10 |
| S | 40.8 | 212.1 |
| Number of features | 129 | 9 |

5 Conclusions and Outlook

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. 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.

Furthermore, for the experiment on real-world data, we found 9 features suitable for earthquake detection and wave type discrimination for 3 recorded events. By comparing classification errors for the corresponding SOM, we found that this feature set provided better discrimination between seismic wave types than using all potential features. We plan to investigate the generalization capability of our procedure for larger earthquake data sets using cross-validation.

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 [4] which are able to classify the event type and to detect seismic phases.

Acknowledgment

Parts of this work has been made possible under the EU-Project NERIES contract no. 026130.

References

1. Joswig, M., 1990, Pattern recognition for earthquake detection, *Bulletin of the Seismological Society of America*, **80**(1), 170–186
2. Dai, H. & MacBeth, C., 1995, Automatic picking of seismic arrivals in local earthquake data using an artificial neural network, *Geophysical journal international*, **120**(3), 758–774
3. Ohrnberger, M., 2001. *Continuous Automatic Classification of Seismic Signals of Volcanic Origin at Mt. Merapi, Java, Indonesia*. Dissertation, University of Potsdam

4. Riggelsen, C., Ohrnberger, M. & Scherbaum, F., 2007, Dynamic Bayesian Networks for Real-Time Classification of Seismic Signals, *Lecture Notes in Computer Science*, **4702**, 565–572
5. Bellman, R., 1961. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ
6. Bishop, C., 2006. *Pattern recognition and machine learning*. Springer
7. Dy, J. & Brodley, C., 2004, Feature Selection for Unsupervised Learning, *The Journal of Machine Learning Research*, **5**, 845–889
8. Basak, J. & De, R.K., & Pal, S.K., 1998, Unsupervised feature selection using a neuro-fuzzy approach, *Pattern Recognition Letters*, **19**(11), 997–1006
9. Mitra, P. & Murthy, C.A., & Pal, S.K., 2002, Unsupervised Feature Selection Using Feature Similarity, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(3), 301–312
10. Li, Y. & Lu, B.L., & Wu, Z.F., 2007, Hierarchical fuzzy filter method for unsupervised feature selection, *Journal of Intelligent and Fuzzy Systems*, **18**(2), 157–169
11. Kvaerna, T. & Ringdahl, F., 1986, Stability of various FK estimation techniques, *Semianual technical summary, 1 October 1985 - 31 March 1986, NOR SAR Scientific Report, Kjeller, Norway*, **1-86/87**, 29–40
12. Aki, K., 1957, 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
13. Vidale, J., 1986, Complex polarization analysis of particle motion, *Bulletin of the Seismological Society of America*, **76**(5), 1393–1405
14. Park, J., Vernon III, F. & Lindberg, C., 1987, Frequency dependent polarization analysis of high-frequency seismograms, *Journal of Geophysical Research*, **92**(B12), 12664–12674
15. Jurkevics, A., 1988, Polarization analysis of three-component array data, *Bulletin of the Seismological Society of America*, **78**(5), 1725–1743
16. Taner, M., Koehler, F. & Sheriff, R., 1979, Complex seismic trace analysis, *Geophysics*, **44**, 1041–1063
17. Pinnegar, C., 2006, Polarization analysis and polarization filtering of three-component signals with the time-frequency S transform, *Geophysical Journal International*, **165**(2), 596–606
18. Wald, A. & Wolfowitz, J., 1940, On a Test Whether Two Samples are from the Same Population, *The Annals of Mathematical Statistics*, **11**(2), 147–162
19. Kohonen, T., 2001. *Self-Organizing Maps*. Springer
20. Davies, D. & Bouldin, D., 1979, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**(2), 224–227
21. Vesanto, J. et al., 2000. *SOM Toolbox for Matlab 5*. Helsinki University of Technology
22. Vesanto, J. & Alhoniemi, E., 2000, Clustering of the self-organizing map, *Neural Networks, IEEE Transactions on*, **11**(3), 586–600
23. Vesanto, J. & Ahola, J., 1999, 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*
24. Vesanto, J. & Sulkava, M., 2002, Distance Matrix Based Clustering of the Self-Organizing Map, *Proc. International Conference on Artificial Neural Networks–ICANN 2002* 951–956
25. Barreto, M.A. & Pérez-Urbe, A., 2007, Improving the Correlation Hunting in a Large Quantity of SOM Component Planes, *Lecture Notes in Computer Science: Artificial Neural-Networks–ICANN 2007*, **4669**, 379–388