Mathematical Sciences

Individual Report

# Wave Energy Converters

*Submitted By :*
You ZUO, Jingzhuo HUI

PICMath

# 1 Description of Context

The original research of our data sets investigates the placement optimization of oscillating buoy-type wave energy converters (WEC). Optimizing the buoy positions is a challenging research problem because of very complex interactions (constructive and destructive) between buoys. The primary purpose of the original research is maximizing the power output of the farm through the placement of 16 buoys in a size-constrained environment.

During the research, four potential sites on the southern coast of Australia are considered: Adelaide, Perth, Tasmania and Sydney. From related research reports, we can know these wave regimes vary in terms of total energy and directional distributions of the waves.

# 2 Attribute Information

The attributes of our data sets can be roughly divided into three categories:

(1) WECs position $\{X_1, X_2, \dots, X_{16}; Y_1, Y_2, \dots, Y_{16}\}$ continuous from 0 to 566(m).

(2) WECs absorbed power: $\{P_1, P_2, \dots, P_{16}\}$

(3) Total power output of the farm: *Powerall*

Here $[X_i, Y_i]$ is the position of the $i^{th}$ buoy, $P_i$ the corresponding absorbed power generated by the $i^{th}$ buoy independently, which means it has neither a constructive nor a destructive interaction from any other buoy. *Powerall* is the sum of mean power output by buoys positioned in an area at x-positions $x = [X_1, X_2, \dots, X_{16}]$ and corresponding y-positions $y = [Y_1, Y_2, \dots, Y_{16}]$, and it is the target of regression.

The repository indicates that there are missing values in our data set, but there is in fact no obvious NA value in the data. However, we found some values in Sydney_Data beyond the boundary of $[0, 566]$:

Table 1: Values beyond the boundary of Sydney_Data

| value | nbrow | nbcolumn |
|---|---|---|
| 566.0008 | 71981 | 28 |
| 566.0008 | 71982 | 28 |
| -0.0345 | 71990 | 11 |
| -0.0345 | 71991 | 11 |
| -0.0001 | 71995 | 13 |
| -0.0001 | 71996 | 13 |

We removed the outliers and then made a preliminary simple analysis of the data. Since we implemented all our methods to the four datasets exactly the same way, we choose arbitrarily the dataset Adelaide_Data here for demonstration. We firstly made some box plots of Adelaide's WECs positions and their absorbed power:

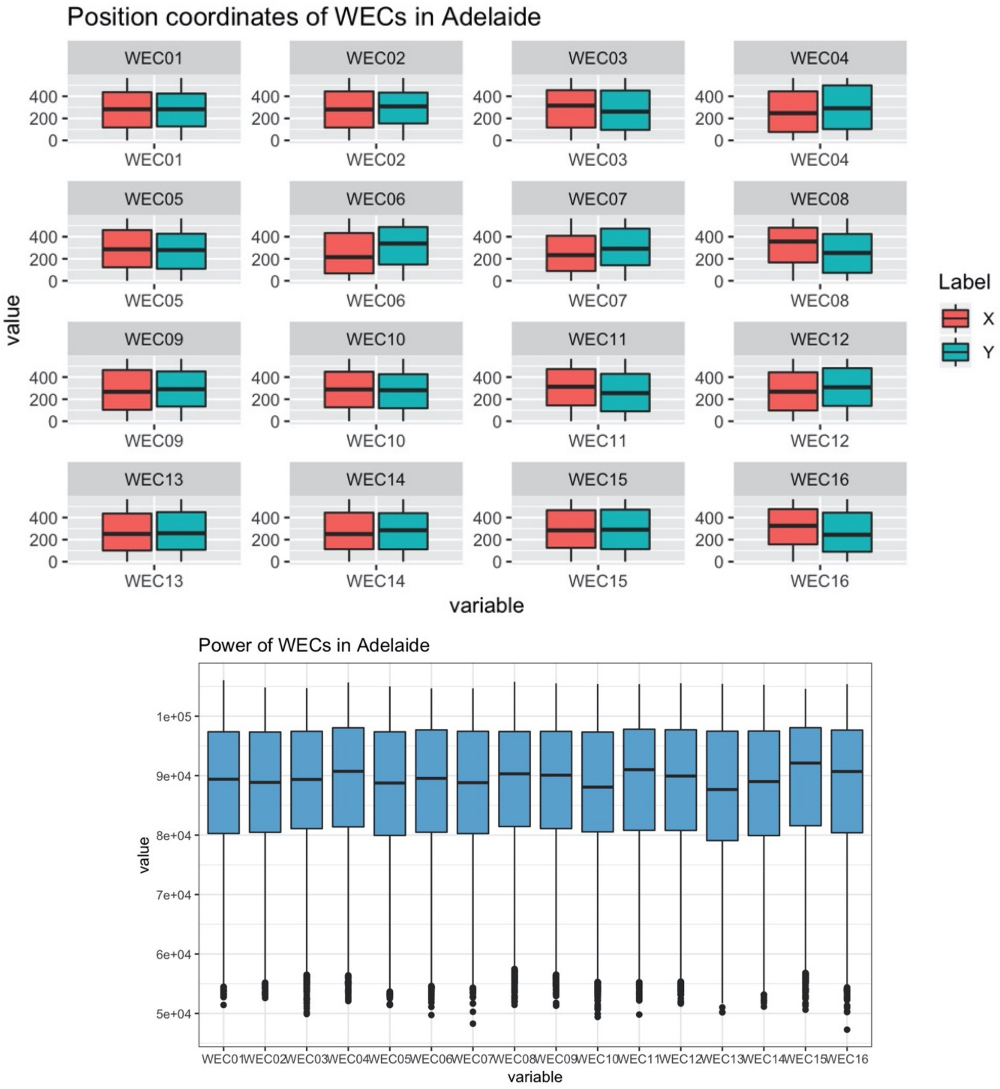
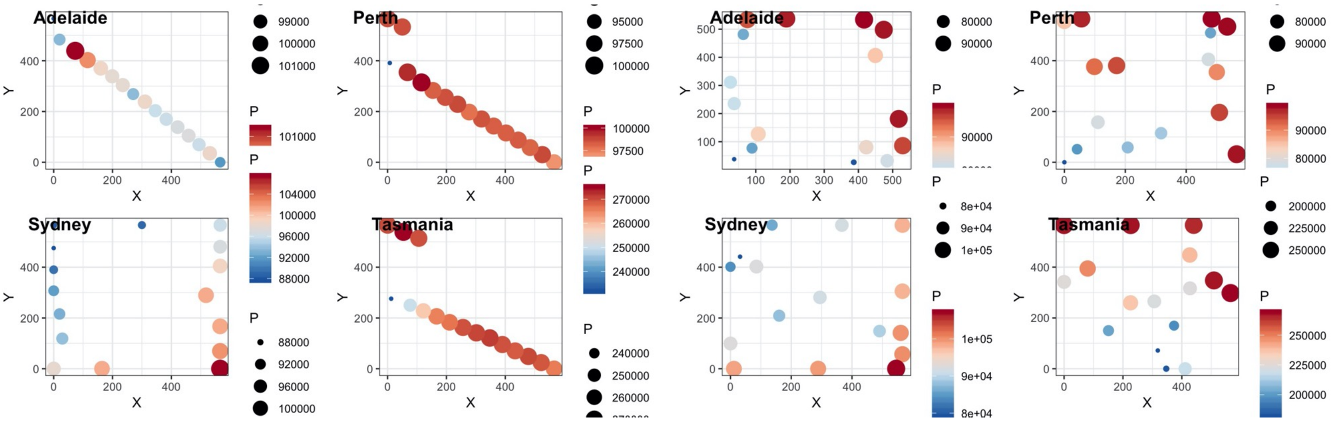


Figure 1: Boxplots for Adelaide_Data

From the eight boxplots we have made, including the two above in Figure 1, we can see that the distribution of the buoys in Adelaide, Perth and Tasmania are stable, while in Sydney it is relatively unstable and it has a more massive range. In fact, the third report given in the repository indicated that the layout for the Sydney wave scenario differs markedly in its spacing and orientation. The wave environment is more varied in terms of wave direction. As a result, the optimization runs produced layouts where the buoys are well-spaced, which minimizes destructive interactions.

Also, we found that the collected energy is also distributed differently on different coasts. Tasmania has the highest energy of about 230,000 watts, followed by Sydney's farm, the energy there collected by each buoy is about 90,000 to 97,000 watts. Finally, Perth and Adelaide's energy values are somehow similar, which is about 90,000 watts.

After that, we selected some observations of placement of the buoys. We have separately the placement of the optimal solutions and some randomly selected ones:



(a) optimal placements

(b) randomly selected placements

Figure 2: examples of the placement of buoys

According to these optimal placements in Figure 2(a), we can see that for Perth, Adelaide and Tasmania the layouts are similar with the buoys being oriented in a line roughly normal to the prevailing wave direction.

The layout produced for the Sydney wave scenario is very different with the buoys being placed at large distances from the others. This pattern was observed for a number of Sydney runs where the best layouts tended to contain widely dispersed buoys to minimize destructive interference.

## 3 Correlation Analysis

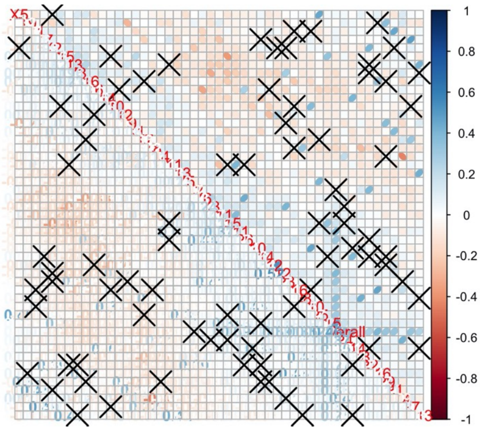


Figure 3: Corrplot for Adelaide_Data

It is clear to been seen from the Figure 3 above: since we have too many variables, it is almost possible to visualize their correlations by the plot. We should analyze it by the correlation matrix if necessary.

While, according to the results, among all 49 variables, there is only one pair of data has a correlation coefficient of more than 0.5($cor(Adelaide\_Data)[28, 44] = 0.52$), so it can be determined that the correlation between the original data is very small. Secondly, the correlation analysis can be said to be of **no practical** significance to our original data, because the research background has determined that we cannot delete the measurement value of any buoy for correlation reasons.

# 4 Presentation of new features

## 4.1 Conception of reconstruction features

As we can see, there were problems with the data structure. Cause it was hard for us to start building our regression model yet with the variable selection, so we decided to reconstruct new features with old variables. Here we picked arbitrarily one placement data of Adelaide_data to demonstrate the conception:

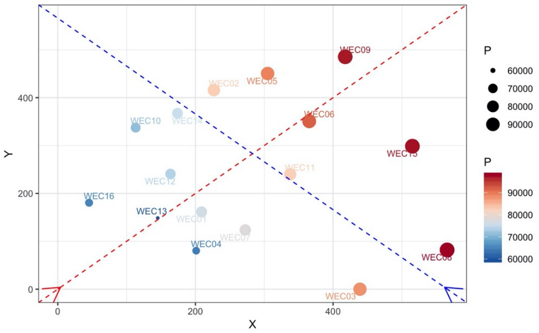


Figure 4: Construction of new features

The original data are distributed in a $[0, 566] \times [0, 566]$ coordinate system with a certain order for WECs. However, here we would like to rearrange them along the direction of two diagonal lines: the positive diagonal line, which is the blue line in *Figure4*, and the minor-diagonal line, which is the red one. The origin points for the red line and the blue line are respectively (0,0) and (566,0), then we calculated the distance from each WEC to the diagonal vertex from near to far. We have 16 position coordinates for every placement, so in this way, we get 16 distances for each diagonal direction hence 32 features in the end.

We have four dataset, each table has exactly the same structure except that they were measured in different wave scenarios, so to use all our data sets as a whole, we had to add a variable "location". The variable has 4 possibilities: Adelaide, Perth, Sydney and Tasmania. To build a multiple regression model with such a categorical variable, we introduced dummy variables which allow us to identify which of the four scenarios an individual falls into.

Besides, we deleted the variables $\{P_1, P_2, \ldots, P_{16}\}$, because, after testing, we found that Powerall is actually the sum of $P_i$. Hence our data set are in this structure: $\{Position, D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}, D_{14}, D_{15}, D_{16}, P_1,$ refers to the direction of positive diagonal, while P the negative diagonal direction).

## 4.2 Correlation analysis

Before implementing any method of regression to our dataset, we want to summary the correlations between our new features:

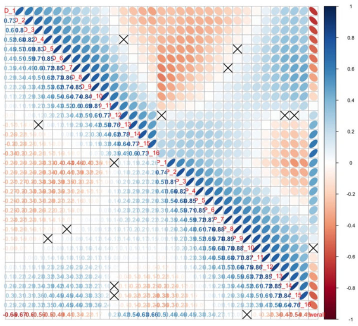


Figure 5: Corrplot of new data

And from the corrplot above, we can see that the correlations between our new features are pretty high compared with what we got from our original variables.

# 5 Baseline Models

## 5.1 Baseline model for all data sets

At first we implemented a **multiple linear regression with one categorical variable** with all of our four data set:

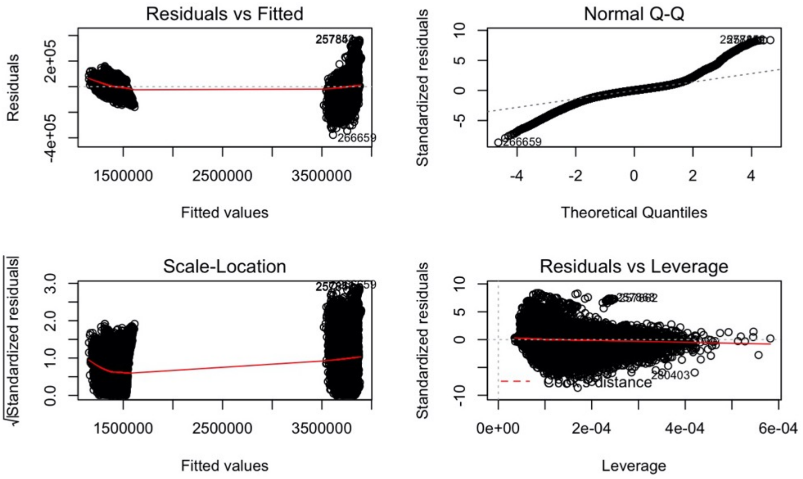


Figure 6: diagnostic plots of baseline model of all datasets

This model has pretty good results for significance level and the R-squared value which is 0.9981. However, after looking at its diagnostic plots, we thought it might not be a good idea to combine all the data from different scenarios, because from the first plot we found that there are three locations have similar wave energy levels, while the data collected from Tasmania are much higher than the others. In this way the target "Powerall" will have a large blank interval in the middle, which would lead to the inaccuracy of the model. Finally, we decided to build a linear model only for one location each time. By the way, except for the nuances of the methods due to the characteristics of the data itself, they have almost entirely similar processes, so we ended up choosing only the first set of data.

## 5.2 Baseline model for Adelaide_data

For data set of Adelaide, we firstly split the data into training data set and testing data set with a ratio of 0.7 and 0.3. We used the training data to build the regression model, the equation of result is:

$\hat{y} = 1.167421e + 06 - 1.640867e + 02D_1 - 1.591625e + 02D_2 - 5.829057e + 01D_3 - 3.212843e + 01D_4 + 8.831993e + 00D_5 - 5.152022e + 01D_6 - 3.007600e + 00D_7 - 1.473685e + 00D_8 9.108618e + 01D_9 - 6.546072e + 01D_{10} + 1.142567e + 01D_{11} + 6.198135e + 00D_{12} + 5.576902e + 01D_{13} + 7.386615e + 01D_{14} + 1.673023e + 02D_{15} + 2.720101e + 02D_{16} + 9.714379e + 01P_1 + -1.042784e + 01P_2 + 2.448317e + 01P_3 + 4.055566e + 01P_4 - 2.734366e + 01P_5 + 2.942442e + 01P_6 + 1.639010e + 01P_7 + -3.465341e + 01P_8 + 8.422266e + 00P_9 + -6.374848e + 01P_{10} + 4.364991e + 00P_{11} + 7.591393e + 00P_{12} + 3.817306e + 00P_{13} - 9.911260e + 00P_{14} - 9.231525e + 01P_{15} - 7.303094e + 01P_{16}$

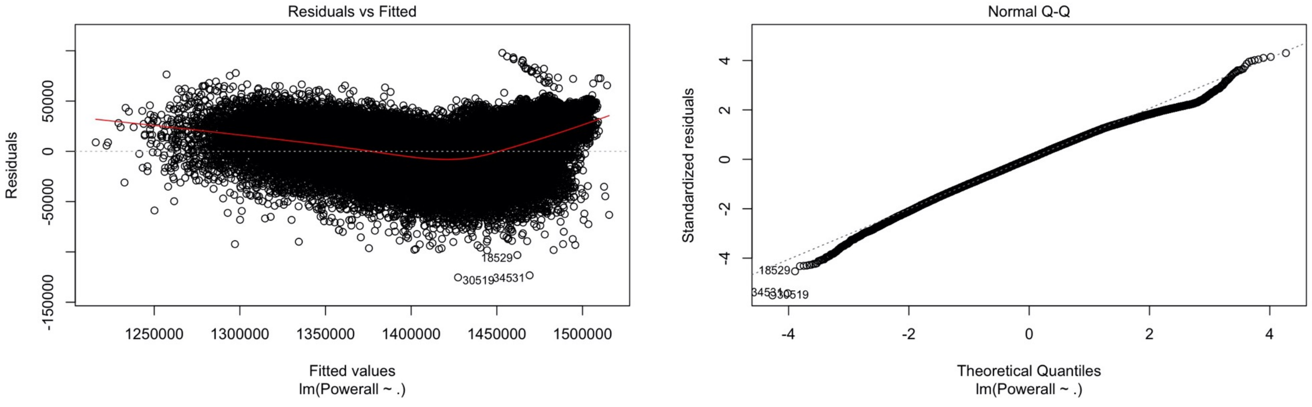


Figure 7: diagnostic plots of baseline model of Adelaide_data

The R-squared for this model is 0.8346, and some of the variables have very high p-value, which indicates that they have respectively low significance level and could be removed. For diagnostic plots, the residuals are well distributed in a normal distribution. However, the curved line in the first plot infers that the model needs to be improved.

# 6 Variable Selection

## 6.1 Step by Step Selection

We used these four methods direction ”backward” with criteria ”AIC” and ”BIC,” direction ”forward” with criteria ”AIC” and ”BIC,” and we also used ”stepwise” direction. However, because its results are the same as those of the above, we will not show its results here.

1. for both direction ”backward” and ”forward”, AIC criterion selected 25 variables from all 32 variables while BIC selected only 21 of them
2. All the models have 22770 as their RSE
3. models of AIC have similiar value of R-squared equal to 0.8345, models of BIC 0.8344
4. models of AIC have similiar value of R-squared equal to 0.8353, models of BIC 0.8352

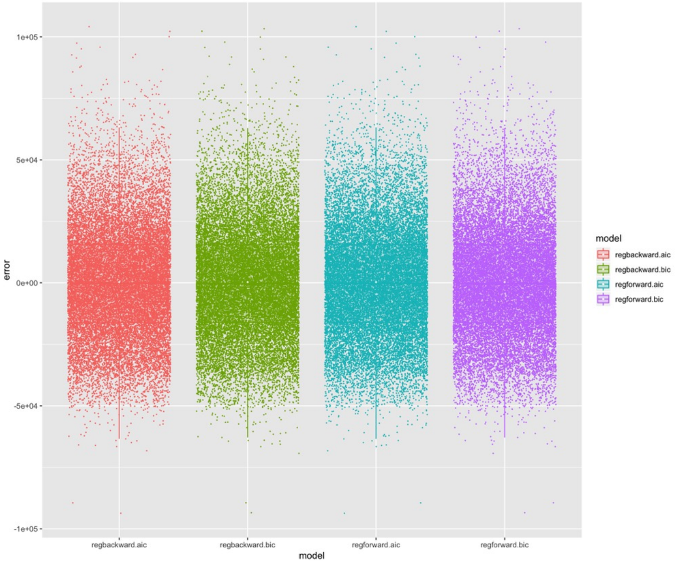


Figure 8: Boxplots of residuals of 4 different step models

As you can see, these variable selection methods did not optimize so much our initial model. So decided to continue exploring the method of penalty functions to optimize our model.

## 6.2 Penalized Regression

### RIDGE

We first used the ridge method, but the results were not so ideal. The coefficient of the penalty term needs to be large to make the coefficient of the variable slightly change, and the change of the GCV is very small as the $\lambda$ changes.

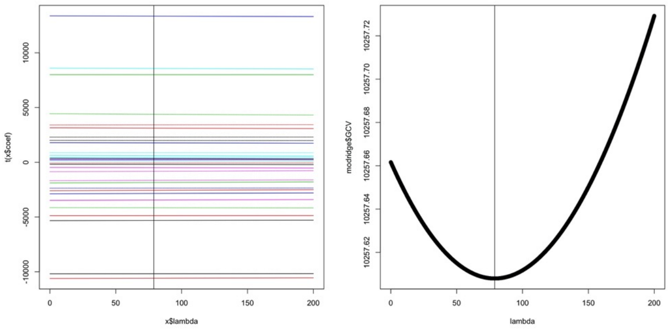
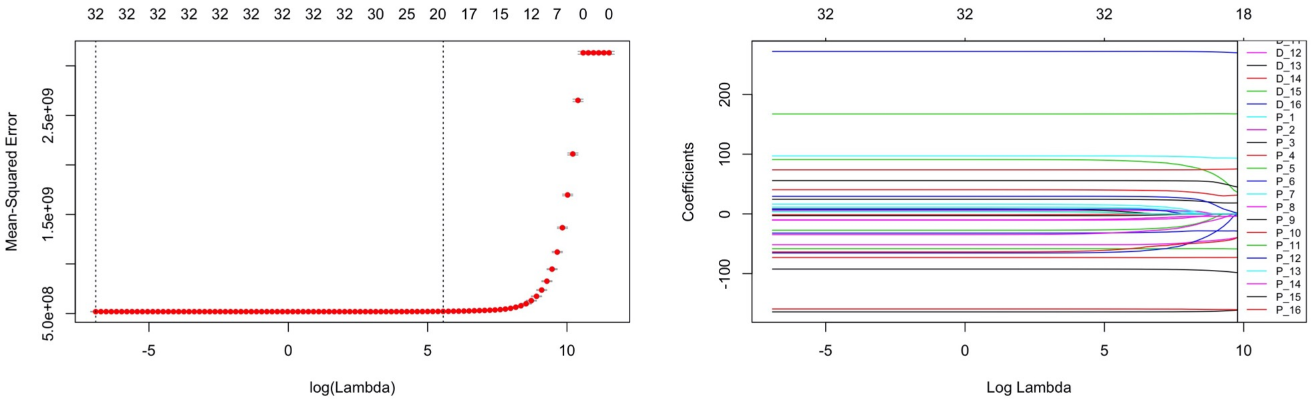


Figure 9: Results of ridge penalized regression

### LASSO

After we turned to use the lasso method, we chose $\lambda$ through a 20-fold cross-validation, and some of the coefficients shrank when *lambda* gets very big. However, the optimal $\lambda$ calculated by the algorithm is 1.707353.



(a) cross-validation results

(b) coefficients shrinkage

Figure 10: Results of lasso penalized regression

### Elastic Net

We also used Elastic Net penalty. The solution is to combine the penalties of ridge regression and lasso to get the best of both worlds. Elastic Net aims at minimizing the following loss function:

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i^2\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|)$$

where $\alpha$ is the mixing parameter between ridge ($\alpha = 0$) and lasso ($\alpha = 1$).

Previously we used the training set to build the model, and then we used the test set to make predictions and evaluations. We summarize the performance of the model in the following table, where SSR stands for sum of squared residuals:

Table 2: Performance of penalized regression

| | R-squared | SSR |
|---|---|---|
| ridge cross-validated | 0.8345725 | 1.116354e+13 |
| lasso cross-validated | 0.8345802 | 1.116359e+13 |
| elastic net | 0.8345802 | 1.116769e+13 |

# Bibliography

[1] A. Aldroubi, C. Cabrelli, U. Molter, and Sui Tang, Dynamical sampling, *Applied and Computational Harmonic Analysis*, doi:10.1016/j.acha.2015.08.014, 2016

[2] A. Aldroubi, C. Cabrelli, A. F. Cakmak, U. Molter, and A. Petrosyan, Iterative actions of normal operators, Submitted. Available at http://arxiv.org/abs/1602.04527.

[3] K. Groechenig, *Foundations of time-frequency analysis*, Birkhäuser Boston, 2001.