Selection of Nonlinear Features and Interactions in Random Fourier Basis Regression

Tristan Pollner

Abstract

Prediction of an outcome variable from multiple observed features is a central question in modern computational statistics with numerous applications, and feature selection for prediction is an important way to facilitate model interpretation and can play a critical role in improving modeling for data with high dimensionality. Here we consider the most general model, with considerations for both non-linearity and non-additivity. Building upon the recently proposed "Random Kitchen Sinks" approach to efficient non-linear modeling, and using the Fourier basis to implement both additive and non-additive features, we first present a new application of the full group lasso. Expanding on this for efficiency, we construct models with stochastically generated interaction terms, and use the statistical lasso to regularize parameter estimates. We show how to perform variable selection in a manner that identifies the set of features that together best predict the outcome variable, as well as which specific terms interact and propose a connection to the group lasso approach. The approach is implemented in an R package, and we benchmark it on a variety of simulated and empirical datasets, in areas including robot dynamics, tumor diagnosis, and global health, demonstrating improvements over existing methods for prediction and feature selection.

1 Introduction

Prediction and variable selection problems are critically important, both in research and industry, especially as data scientists routinely consider problems with thousands of potential independent variables. Here, we represent our set of predictors with $\mathbf{X}$, an n by p matrix where n is the number of instances of p features, where the $\mathbf{X}_{i,j}{}^{\text{th}}$ element gives the i^{th} observation of feature j, and we represent our response $\mathbf{Y}$ as a vector of length n. $\mathbf{X}_i$ denotes the ith column of $\mathbf{X}$. Although traditional regression methods assume continuous outcome variables, we are able to model discrete outcomes through the generalized linear model, which has been previously discussed [8].

The statistical lasso (Least Absolute Shrinkage and Selection Operator) behaves as its acronym suggests: a shrinking loop catching only certain variables. First introduced in 1994 [14], it is one of the most important and frequently used techniques in variable selection. It employs regularization: limiting the number of features using penalties on the magnitudes of coefficients. The lasso, in the context of the generalized linear model, follows the below equation, as implemented in the `glmnet` R package by Friedman *et al.* [2] if β is a vector of coefficients of length p.

$$\min_{\beta} L(\mathbf{X}\beta, \mathbf{Y}) + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2) \tag{1}$$

Here L is a function representing the negative log-likelihood of the proposed model β (in least-squares regression minimizing L is equivalent to minimizing $(\mathbf{X}\beta - \mathbf{Y})(\mathbf{X}\beta - \mathbf{Y})^T$), and α, called the *elastic net penalty*, controls our penalty function. The standard setup in this project is $\alpha = 1$ which gives pure lasso regression (see Figure 3 for a graphical representation of lasso regression at different alpha values.)

In situations where it is natural to think of predictors as belonging to disjoint groups and where group selection (rather than individual feature selection) is of importance, the standard lasso will not suffice. The group lasso, introduced by Yuan and Lin [15], provides a framework to answer this question. It is necessary to extend the notation discussed earlier, and we follow the standard used by Simon *et al.* [13]. This notation considers L groups of variables, each containing s_i features where $1 \leq i \leq L$. We split our matrix of predictors $\mathbf{X}$ into L matrices $\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_L$, consisting of the features for the corresponding group, and similarly split our coefficient vector β into L vectors $\beta_1, \beta_2, ...\beta_L$. The relevant minimization is then:

$$\min_{\beta}\left(\|\mathbf{Y}-\sum_{i=1}^{L}\mathbf{X}_i\beta_i\|_2^2+\lambda\sum_{i=1}^{L}\sqrt{s_i}\|\beta_i\|_2\right) \tag{2}$$

In standard lasso and group lasso regression for data $\mathbf{X}$, only linear fits are considered, but the real world is seldom linear, and our models should reflect this. Here we show how this can be achieved by applying the standard lasso and the group lasso to random Fourier basis regression, demonstrating empirical improvements over existing techniques in predictive accuracy and variable selection.

There is no canonical technique for nonlinear regression, and many proposed methods work well only for certain types of problems. We consider the general case of non-linear and non-additive modeling: where the change $\mathbf{Y}$ with respect to $\mathbf{X}_i$ can depend on the value of $\mathbf{X}_i$ (i.e. the model is non-linear), and where the change in $\mathbf{Y}$ with respect to $\mathbf{X}_i$ can depend on $\mathbf{X}_j$ (i.e. the model is non-additive). Recently a novel method proposed for nonlinear regression has demonstrated promising results in fitting a large variety of data efficiently [6], and in exhibiting reasonable extrapolation behavior when relevant. Its particular strength is in modeling highly nonlinear data, which is a weakness of classic regression methods. The technique is based on the Fourier series, where we write

$$f(x)=\sum_{i=1}^{\infty}(a_n\cos nx+b_n\sin nx)$$

for an integrable function $f(x)$ with closed form formulas for a_n and b_n. Thus a series of sines and cosines can be used to approximate any $f(x)$, assuming $f(x)$ follows the Dirichlet conditions. Rather than using the standard coefficients a_n and b_n for a known $f(x)$, we seek to approximate an unknown, smooth $f(x)$ that is only sampled through noisy observations, using sines and cosines of varying amplitudes and frequencies. Selection of these can be done through random kitchen sinks.

Rahimi and Recht proposed random kitchen sinks as a technique "replacing minimization with randomization in learning" to increase computational efficiency [10]. When minimizing something of the form $\sum_k f(x;w_k)\alpha_k$, "rather than jointly optimizing over α and w, the following algorithm first draws the parameters of the nonlinearities randomly from a pre-specificied distribution p. Then with w fixed, it fits the weights α optimally via a simple convex optimization". In the model given by Equation 3, the coefficients $A_{i,j}$ have the same role as α_k and can be learned with the group lasso, and the technique proposed by Rahimi and Recht provides us with an efficient way of selecting frequencies (to fill the matrix R, in Equation 3 as well).

Applications to a Fourier basis have been considered in the previous literature recently, though never in the context of variable selection. Lopez-Paz *et al.* have used a normal distribution $\mathcal{N}(0, s)$ for frequency selection, noting similarities to other machine learning methods [5]. We follow this convention in the standard case, but also extend this, noting that we may wish to use a different p to exploit any prior beliefs about the data, especially any form of possible periodicity.

In this paper we develop proposed mechanisms behind feature selection in such a context, for both additive and non-additive models. Instead of only providing a list of relevant variables, we also provide insight into structure with the selection of specific features/interactions. We explain the usage of a group lasso for a full search of pairwise interactions, and describe an approach based on the randomized inclusion of interactions in basis functions, eventually connecting the two approaches.

2 Feature Selection in an Additive Nonlinear Model[*]

2.1 The Basic Model

Here we outline our approach, both for prediction and variable selection, beginning with the simpler additive case. Relevant parameters include the number of sinusoidal basis functions per variable, and the frequencies and amplitudes of each sinusoid, which together control the flexibility of the function we seek to model. The model with $2k$ basis functions is given below, where ϵ is a noise vector of length n sampled from a normal distribution, R is a p by $2k$ matrix of chosen frequencies, and A is a p by $2k$ matrix of learned coefficients. Note that in practice, we often also add identity basis functions that preserve the original inputs to allow purely linear relationships, where necessary.

$$\mathbf{Y} = \sum_{i=1}^{p} \left(\sum_{j=1}^{k} A_{i,j} \sin\left(R_{i,j}\mathbf{X}_i\right) + \sum_{j=k+1}^{2k} A_{i,j} \cos\left(R_{i,j}\mathbf{X}_i\right) \right) + \ \epsilon \tag{3}$$

With only one variable (i.e. $p = 1$), the question of variable selection is irrelevant, but the lasso is still useful, controlling the flexibility of the sum of the sinusoids and preventing

[*] This section, in combination with sections 3 and 4, may be regarded as equivalent to a “Materials and Methods” section.

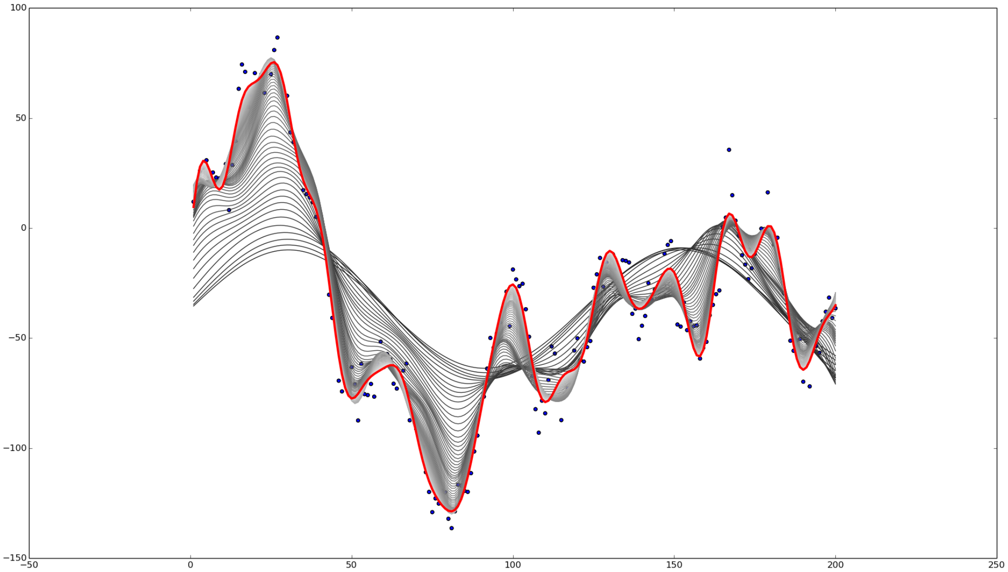


Figure 1: The Lasso for a Fourier Basis for a random walk. The red curve represents the least regularized fit, and successive curves have higher λ values and simpler fits.

over-fitting. Figure 1 demonstrates the effect of the lasso on data simulated under a random walk with a single independent variable, where $\alpha = 1$ (but see Figure 3 for other values of α).

Of course, the general and more important case that this paper addresses is $p > 1$. Each independent variable is expanded into multiple sinusoids. It thus becomes natural to group the sinusoids from each together and apply group penalties (Equation 2). Here $L = p$ (each of the L groups corresponds to exactly one of the p variables), $s_i = 2k$ for all $1 \leq i \leq L$, and β is a vector of all $2pk$ coefficients formed by concatenating all rows of A. Following the notation of Equation 2 we therefore can regularize as follows:

$$\min_{\beta} \left(\|\mathbf{Y} - \sum_{i=1}^{p} \sum_{j=1}^{k} A_{i,j} \sin(R_{i,j}\mathbf{X}_i) - \sum_{i=1}^{p} \sum_{j=k+1}^{2k} A_{i,j} \cos(R_{i,j}\mathbf{X}_i)\|_2^2 + \lambda \sum_{i=1}^{p} \sqrt{2k} \|\beta_i\|_2 \right) \tag{4}$$

We are assuming an additive nonlinear model (without interactions between variables), so each column $\mathbf{X}_i$ is getting pushed through a set of $2k$ basis functions and we take a sum over all variables. When considering the data after it has gone through the basis functions, a standard lasso is clearly not desirable because it will exclude individual basis functions from

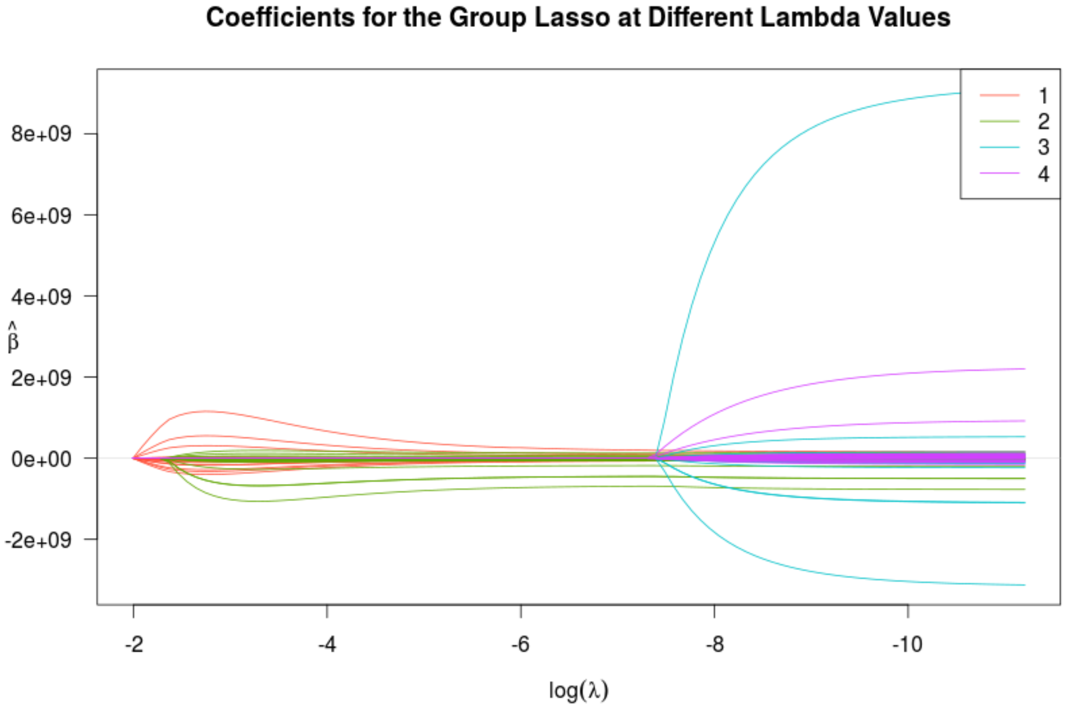


Figure 2: Using the Group Lasso for Variable Selection in an Additive Model with the Fourier Basis.

Here $\mathbf{Y} = 10(\mathbf{X}_1^{2.5} - \mathbf{X}_1^{1.5}) + 7(\mathbf{X}_2^{2.5} - \mathbf{X}_2^{1.5}) + 0\mathbf{X}_3 + 0\mathbf{X}_4 + \epsilon$, and we can see that for small lambda all 4 predictors are included as the model predicts noise. With more regularization only the first two predictors are included, and at the highest lambda we only include $\mathbf{X}_1$. Note how groups of basis functions drop to zero at the same lambda in an appropriate order.

the model as opposed to features. However, if we collect these basis functions into groups of size $2k$ and penalize L2-norms of these groups as in the group lasso, we can encourage these groups of basis functions, corresponding to features, to drop to zero. In our work we used the implementation provided by a `grLasso` function in the `grpreg` package [1].

As a final note for variable selection in an additive model, we stress the importance of including all potential predictors in the group lasso. Checking pairwise correlation between $\mathbf{X}_i$ and $\mathbf{Y}$ is a technique that although occasionally used in practice will not catch important relationships in both additive and interactive models; examples are given in “Results”.

2.2 Frequency Selection

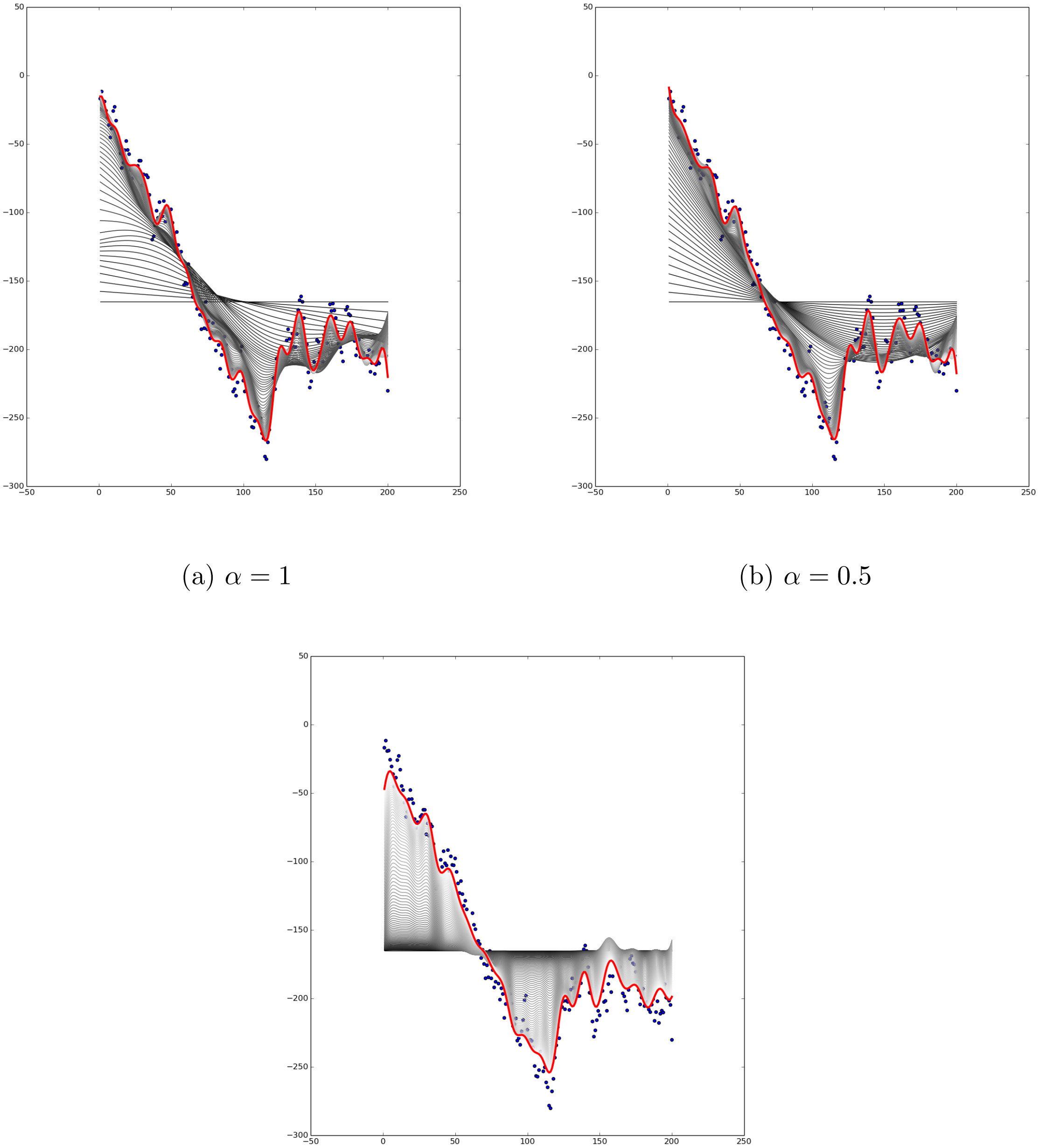
We follow the work of Lopez-Paz *et al.* in using a normal distribution as a standard in frequency selection [5], but also note that there are natural extensions. If, for example, we believe that $\mathbf{Y}$ may be periodic in $\mathbf{X}_i$ with period t, we can set some of the relevant frequencies in R to be close to $\frac{2\pi}{t}$ (this may be implemented through the frequency distribution). Figure 4 demonstrates such an approach for data from the Mauna Loa Observatory in Hawaii, which monitors and releases monthly carbon dioxide concentrations. The data exhibit semi-periodic properties because the amount of the gas fluctuates according to the season and exhibits a highly nonlinear trend, but our model with very few parameters provides a reasonable fit and sensible extrapolation behavior.

3 Variable Selection with Pairwise Nonlinear Interactions between Predictors

The model described in Equation 5 is sufficient when data can be effectively described by an additive model

$$\mathbf{Y} = \sum_{i=1}^{p} f_i(\mathbf{X}_i) + \epsilon$$

for general nonlinear and smooth functions f_i. In many data sets, especially those with high noise and/or important omitted predictors, this is enough. However, when data are more precise, a different approach is necessary. For example, if we limit ourselves to two predictors $\mathbf{X}_1$ and $\mathbf{X}_2$, we may have $\mathbf{Y} = f(\mathbf{X}_1, \mathbf{X}_2)$ for a function f that is nonlinear and non-additive. Figure 6 shows simulated data demonstrating this property, and further examples can be



(a) $\alpha = 1$

(b) $\alpha = 0.5$

(c) $\alpha = 0$

Figure 3: Lasso regression with the Fourier basis on simulated data for different alpha values. At $\alpha = 1$ we see a much faster drop to the zero function as basis function quickly drop out of the model – this is the behavior we desire for feature selection.

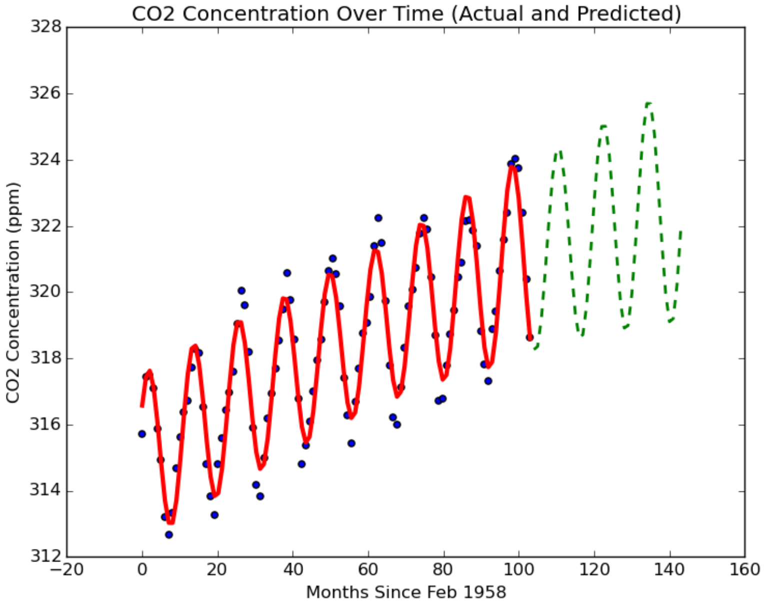


Figure 4: Extrapolation for Semi-Periodic Data from the Fourier Basis

found in the "Results" section. We refer to such a nonlinear function of v variables as a v-way interaction, assuming it cannot be represented as a sum of $(v-1)$-way interactions.

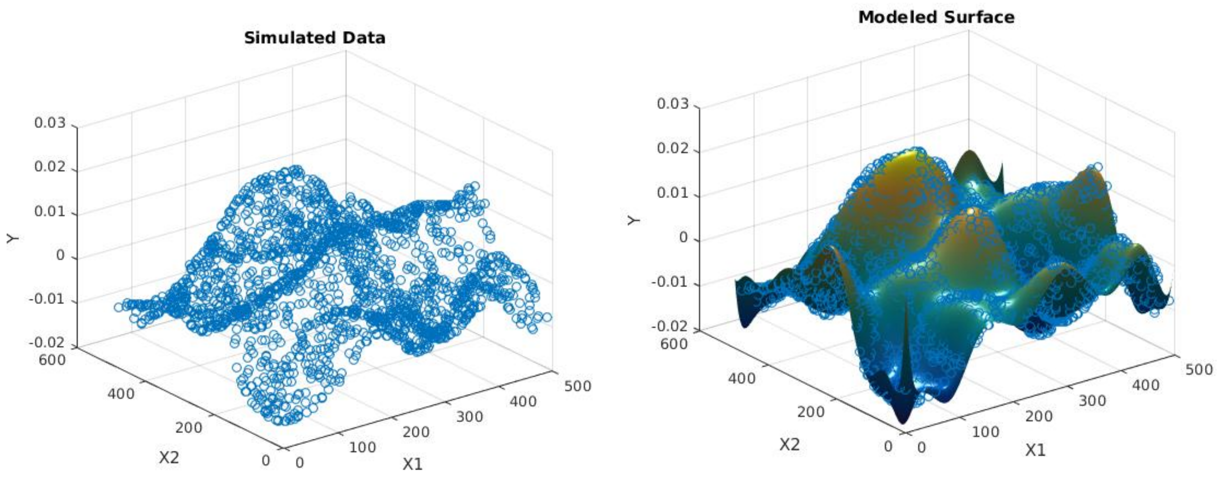
It is of course a theoretical possibility that the desired model is a p-way interaction between all p variables (i.e. $\mathbf{Y} = f(\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_p) + \epsilon$). These functions would however be theoretical curiosities above all else - in actual data sets one would be hard-pressed to find anything higher than a 3-way interaction. For computational considerations that will be formalized later, we first present a technique for selection of only pairwise interactions along with individual components (sufficient for large amounts of data if p is not too large). We assume that the data can be modeled in the form

$$\mathbf{Y} = \sum_{i=0}^{p} f_i(\mathbf{X}_i) + \sum_{1 \leq i \neq j \leq p} g_{i,j}(\mathbf{X}_i, \mathbf{X}_j) + \ \epsilon$$

First, we describe how to model interactions in a Fourier series context. In [4], the multivariate Fourier series is given by

$$f(\mathbf{x}) = \sum_{\mathbf{c}} a_{\mathbf{c}} \sin(\mathbf{c} \cdot \mathbf{x}) + b_{\mathbf{c}} \cos(\mathbf{c} \cdot \mathbf{x})$$

for a vector $\mathbf{x}$ of length p and all vectors $\mathbf{c}$ in $\mathbb{N}^p$ (for the purposes of this paper we will not



(a) Initial data with 2 predicting variables

(b) A surface modeled via the random Fourier series

Figure 5: An example of a nonlinear interaction between variables $\mathbf{X}_1$ and $\mathbf{X}_2$, and the proposed model. The data was generated as a random surface from code posted publicly [9].

worry about closed-form expressions for the coefficients $a_{\mathbf{c}}$ and $b_{\mathbf{c}}$). This suggests that to model an interaction between $\mathbf{X}_a$ and $\mathbf{X}_b$ with $2k$ basis functions, we consider the sum

$$\sum_{j=1}^{k} A_j \sin\left(R\mathbf{X}_a + R\mathbf{X}_b\right) + \sum_{j=k+1}^{2k} A_j \cos\left(R\mathbf{X}_a + R\mathbf{X}_b\right)$$

Note that for simplicity we have labeled the sinusoid frequencies (i.e. the coefficients of $\mathbf{X}_a$ and $\mathbf{X}_b$) with R, a generic placeholder that will generally be pulled from a normal distribution. Following this convention, a full expression for our model is therefore

$$\begin{aligned}\mathbf{Y} = \sum_{1\le a\ne b\le p} \left(\sum_{j=1}^{k_2} A_{a(p+1)+b,j} \sin\left(R\mathbf{X}_a + R\mathbf{X}_b\right) + \sum_{j=k_2+1}^{2k_2} A_{a(p+1)+b,j} \cos\left(R\mathbf{X}_a + R\mathbf{X}_b\right) \right) \\ + \sum_{i=1}^{p} \left(\sum_{j=1}^{k_1} A_{i,j} \sin\left(R\mathbf{X}_i\right) + \sum_{j=k_1+1}^{2k_1} A_{i,j} \cos\left(R\mathbf{X}_i\right) \right) + \epsilon\end{aligned} \tag{5}$$

Above, we consider p individually additive variables, each with $2k_1$ basis functions and $\binom{p}{2}$ 2-way interactions, each with $2k_2$ basis functions. The above equation simply describes the mechanics of expanding our initial matrix $\mathbf{X}$ to a matrix with $2pk_1 + (p^2 - p)k_2$ columns and then considering a group-regularized linear fit on our expanded matrix. We do this by grouping p groups of $2k_1$ functions along with $\binom{p}{2}$ groups of $2k_2$ functions, corresponding to additive and interactive components.

4 Selection among Stochastically Generated Interactions

For many data sets, considering up to 2-way interactions will suffice. However, the two main problems are extreme non-additivity (3-way interactions or higher) and computational feasibility as we increase p. A third problem is that, as the number of interaction terms combinatorially explodes, there is a greater risk of one of those interactions fitting noise by chance, thus appearing important. A possible solution to consider higher order interactions might be to use basis functions of the form $\sin(c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + ... + c_p\mathbf{X}_p)$ and $\cos(c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + ... + c_p\mathbf{X}_p)$, which are effective for modeling a full p-way interaction. However, an nth order basis requires n^p basis functions in the Fourier series, so even in the case where only a few variables are involved this is computationally infeasible.

However, we can safely restrict our study to 3-way or at most 4-way interactions - in practical scenarios, a model with this degree of non-additivity will be more than sufficient, and other data would simply be a simulated curiosity. The approach from section 3 is not sufficient, as considering a full group lasso for this many combinations is impractical for even moderate values of p.

To address this, instead of modeling a set of p-way interactions as mentioned above, we limit the variables included in each basis function through a binomial model. Continuing with the notation above, for each coefficient $c_1, c_2, ...c_p$ we include it in the model with a probability q, so the number of variables included follows a binomial distribution with center pq. The parameter q controls the size of the interactions we consider – larger q values will catch higher degrees of nonlinearity, but only if we include sufficient basis functions according to the size of p. A complete model is therefore given by

$$\begin{aligned}\mathbf{Y} &= \sum_{j=1}^{k} A_j \sin(c_{1,j}\mathbf{X}_1 + c_{2,j}\mathbf{X}_2 + ... + c_{j,p}\mathbf{X}_p) + \sum_{j=k+1}^{2k} A_j \cos(c_{1,j}\mathbf{X}_1 + ... + c_{p,j}\mathbf{X}_p) + \ \epsilon \\ &= \sum_{j=1}^{2k} f_j(\mathbf{X}) + \ \epsilon\end{aligned} \tag{6}$$

where A is a coefficient vector of length $2k$ learned through linear regression (consistent with earlier notations) and

$$c_{i,j} = \begin{cases} \texttt{randn}(0, s) & \text{with probability } q \\ 0 & \text{else} \end{cases}$$

Note that for simplicity of later notation we have abbreviated the first part of Equation 6 with a sum of $2k$ functions f_j.

It is not obvious how to impose variable selection on this, but it is natural to consider a standard lasso, as in Equation 1, for learning the coefficients in A, as this will often reduce the full set of $2k$ basis functions to a much smaller set that explains the model with essentially the same accuracy. However, there is no reason to believe that the included variables will stand out in this set, and empirically we can confirm that this is not the case - even with simulated data there is an insignificant difference. A small improvement would be to look at the basis functions whose coefficients are large in magnitude, under the assumption that these basis functions are contributing more to the model. This works much better, but there are problems inherent to the assumption, as two functions with opposite coefficients can cancel each other out. A different approach is necessary.

To gain insight into the important variables/interactions of the model, we utilize our knowledge of the structure of the function we are using, assuming that it is modeling the data accurately. We know that if a v-way interaction ($v \geq 1$) is of importance, those v variables must be inside the same sine/cosine – if they are distributed among different functions we can model the data with a sum of smaller interactions. A weak and fairly obvious assertion we have is therefore that any interaction between $x_1, x_2, ...x_v$ is irrelevant if it does not appear inside a sinusoid.

A stronger statement, going off the above, is that we can judge the importance of an interaction by measuring how our predicted $\mathbf{Y}$ changes with its removal *from only the basis functions where all variables are included.* In measuring the importance of an interaction with v variables defined by indices $V \subset \{1, 2, ...p\}$, for each function $f_j, 1 \leq j \leq 2k$, we consider the set of inner frequencies $c_{i,j}$. If $c_{i,j} > 0$ for each $i \in V$, create a new set of inner frequencies $C_{i,j}$ where $C_{i,j} = c_{i,j}$ for $i \notin V$ and $C_{i,j} = 0$ for $i \in V$. If this is the case, define:

$$F_j = \begin{cases} A_j \sin(C_{1,j}\mathbf{X}_1 + C_{2,j}\mathbf{X}_2 + ... + C_{j,p}\mathbf{X}_p), & j \leq 2k \\ A_j \cos(C_{1,j}\mathbf{X}_1 + C_{2,j}\mathbf{X}_2 + ... + C_{j,p}\mathbf{X}_p), & j > 2k \end{cases}$$

Else if $c_{i,j} = 0$ for some $i \in V$, we define $F_j = f_j$. We then consider

$$\texttt{cor}\left(\sum_{j=1}^{2k} f_j(\mathbf{X}), \sum_{j=1}^{2k} F_j(\mathbf{X})\right)$$

where `cor` refers to Pearson's r, and use this number as a measure of the importance of the interaction to our model. This is consistent with previously mentioned material as we are able to consider only the effects of that interaction based on the specific structure of our model (taking out all instances of the v variables would also remove their additive components).

Operating as a measure of the importance of an interaction or feature, we can use this not only as an indicator of variable relevance, but as a guide to an efficient use of the group lasso. Assuming that Equation 6 provides an accurate model of the data (of which empirical confirmation is given in the Results section), we can measure the importance of specific variables/interactions and explore these further by expanding the group lasso.

In the developed pipeline, implemented in R, we consider this as the main point of the analysis, after beginning with simpler linear fits through `glmnet` and additive nonlinear regularization as described earlier. A useful component of this package is its scoring of importance of interactions and variables, which is empirically tested in the results and runs extremely quickly – the limiting factor is cross-validation in `glmnet`.

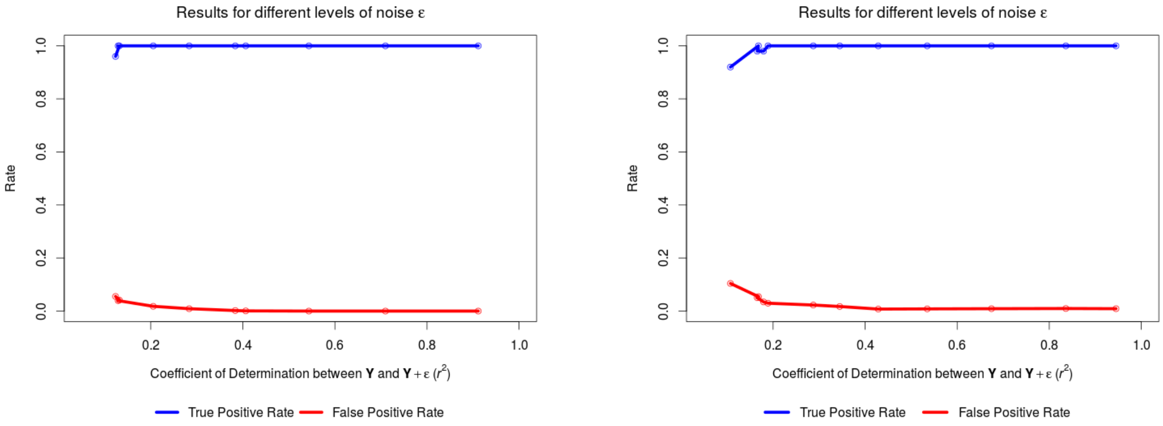
5 Results

5.1 Simulated Data

We begin with simulated data before progressing to applications, which demonstrates the correctness and scope of the technique. Data was generated as follows:

$$\mathbf{Y} = f(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5) + 0\sum_{i=6}^{150} \mathbf{X}_i + \texttt{randn}(0, k) \tag{7}$$

Here each feature $\mathbf{X}_i$ is generated as a vector of length 2000 from a uniform distribution on $[0, 2]$. $\mathbf{Y}$ depends only on a nonlinear function involving $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$, and $\mathbf{X}_5$, and the function `randn` refers to noise ϵ with $\epsilon \sim \mathcal{N}(0, k)$. The first relevant question is that of determining a subset of important variables, and our technique is extremely effective and computationally efficient in this, as well as in providing a functional model for the data.



(a) TP and FP rates for f_1

(b) TP and FP rates for f_2

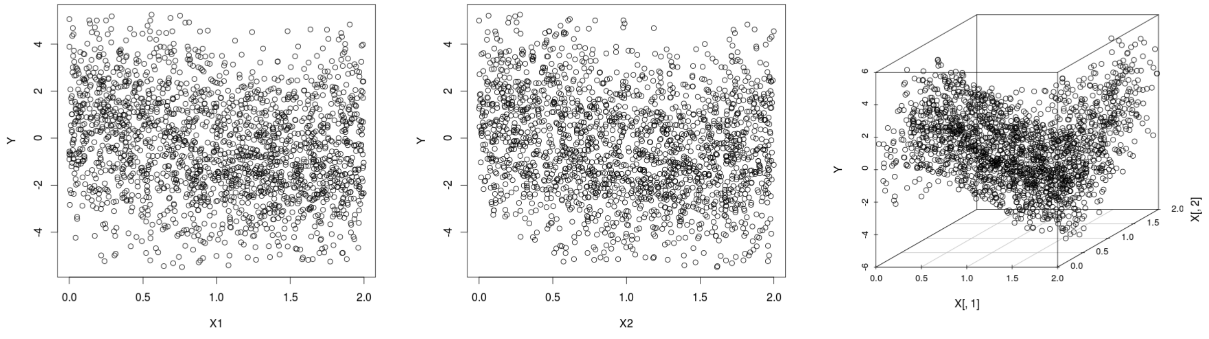
Figure 6: For f_1 and f_2, 10 simulations were conducted at each of the different noise levels considered for ϵ (points on the above graphs). Here we have plotted the True Positive and False Positive rates as a function of r^2 between $\mathbf{Y}$ and $\mathbf{Y} + \epsilon$. We checked for inclusion of variables in a final model, and found extremely accurate results even with tremendous noise. For reasonable noise, for both functions we reported a TP rate of 1 and FP rate of 0.

Functions tested included

$$\begin{aligned} f_1 &= (\mathbf{X}_1 - 1)^2\mathbf{X}_2 + (\mathbf{X}_3 - 1)^2\mathbf{X}_4 + \mathbf{X}_5^{2.5} - 3\mathbf{X}_5 \\ f_2 &= 2\sin(2\mathbf{X}_1 + 2\mathbf{X}_2) + 2\cos(2\mathbf{X}_3 + 2\mathbf{X}_4) + (\sin(2\mathbf{X}_5) + \cos(2\mathbf{X}_5)) \end{aligned} \tag{8}$$

We also compared our method to that proposed by Rosasco *et al.* [12]. Ours is more general as it can provide insight for the importance of specific interactions, but we compared to the relevant aspects of Rosasco *et al.* using the first stage of our approach consisting of modeling and variable selection (i.e. not including the final group lasso). They provide four simulations where $p = 6$ (as opposed to our examples with $p = 150$), and their full results are not reported; Table 1 reports our results including a TP (True Positive) rate of 1 and a FP (False Positive) rate of < 0.05 for variables included in the model. For each function 20 simulations were conducted on the combined training and validation sets, and although we could not directly compare efficiency because of the lack of a provided implementation in [12], our method runs very quickly due to such a small value of p.

For benchmark simulated data, we tested on the `kin40k` dataset: highly nonlinear data intended to be an extremely difficult test even for advanced machine learning algorithms that do not consider feature selection. Pairwise interactions are not enough to explain the set, and therefore this provides a good opportunity to test the capabilities of the first stage in



(a) $\mathbf{X}_1$ vs $\mathbf{Y}$ (b) $\mathbf{X}_2$ vs $\mathbf{Y}$ (c) $\mathbf{X}_1$ and $\mathbf{X}_2$ vs $\mathbf{Y}$

Figure 7: Looking at pairwise r^2 for variable importance is inadequate, as shown by these plots from f_2. For plots (a) and (b), $r^2 = 0$, but ther is a clear nonlinear relationship between $\mathbf{X}_1$, $\mathbf{X}_1$, and $\mathbf{Y}$ as demonstrated by (c). Such data also presents an example of the importance of nonlinear interactions.

Table 1: Results for Simulated Comparison Data. Here $g_1 = (\mathbf{X}_1^4 - \mathbf{X}_1^2)(3 + \mathbf{X}_2)$, $g_2 = 2(\mathbf{X}_1^3 - \mathbf{X}_1)(2\mathbf{X}_2 - 1)(\mathbf{X}_2 + 1) + (\mathbf{X}_2^3 - \mathbf{X}_2 + 3)$ and $g_3 = -2(2\mathbf{X}_1^2 - 1)(\mathbf{X}_2)e^{-\mathbf{X}_1^2 - \mathbf{X}_2^2}$, as in Rosasco *et al.* Our perfect TP rate and very low FP rate demonstrate full success for this set of data. For g_2, we fitted a rank transform of the data due to the extreme range.

Function	True Positive Rate	False Positive Rate
g_1	1	.025
g_2	1	.025
g_3	1	0

our method to model higher order interactions. We do not report full results due to a lack of space and because feature selection is not an important part of this data, but we do note that we have obtained r^2 values higher than 0.99 with cross-validation, while other established algorithms such as support vector machines have values around 0.80 (under default settings).

5.2 Pumadyn Dataset

Both the complete group lasso for pairwise interactions and the technique for higher order interactions were tested on “a realistic simulation of the dynamics of a Puma 560 robot arm. The task in these datasets is to predict the angular acceleration of one of the robot arm’s links. The inputs include angular positions, velocities and torques of the robot arm” [3]. The set `pumadyn32nm` contains 8192 observations of 32 features, and was said to contain nonlinear

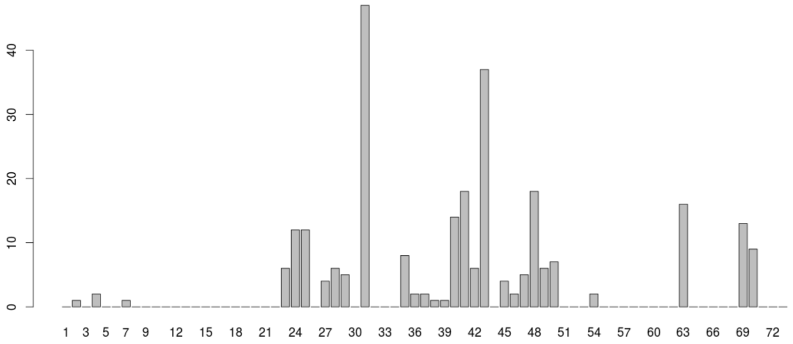


Figure 8: Feature Inclusion Counts for WHO Data over 100 runs. In such large and imperfect data, many features could have some bearing on life expectancy, and we see fairly noisy but still interesting results. In particular, features 31 and 43, “Deaths among children under five years of age due to HIV AIDS” and “Neonatal mortality rate per 1,000 live births” seem to have particular importance to the result.

data with moderate noise. We do not report full results do to a lack of space, but emphasize the success of both feature and interaction selection: in the first stage, we can notice 4 clear significant features and important interactions between $\mathbf{X}_4$ and $\mathbf{X}_5$, $\mathbf{X}_{16}$ and $\mathbf{X}_4$, and $\mathbf{X}_5$ and $\mathbf{X}_{16}$ (in particular, that between $\mathbf{X}_5$ and $\mathbf{X}_{16}$ causes the correlation, as defined in Section 4, to drop to less than .02). In this most sparse model later created/regularized through the group lasso, we achieved a Mean Squared Error of .084 (r^2 of 0.912) with only an interaction between $\mathbf{X}_5$ and $\mathbf{X}_{16}$, representing the angular position of link 5 and the torque at joint 4 respectively.

5.3 World Health Organization Data

This data set provides measures such as life expectancy, GDP, death rates, and other information for 202 countries across the world [11]. This is a very ambitious data set, there are of course missing values, so we began with a cleaning procedure, considering only the $n = 160$ countries out of the total 202 reporting life expectancy and total population data, and consider the 73 features without missing values that do not provide life expectancy information. Our method on each iteration provided a fairly sparse model that accurately explained the data; feature inclusion counts are reported in Figure 8.

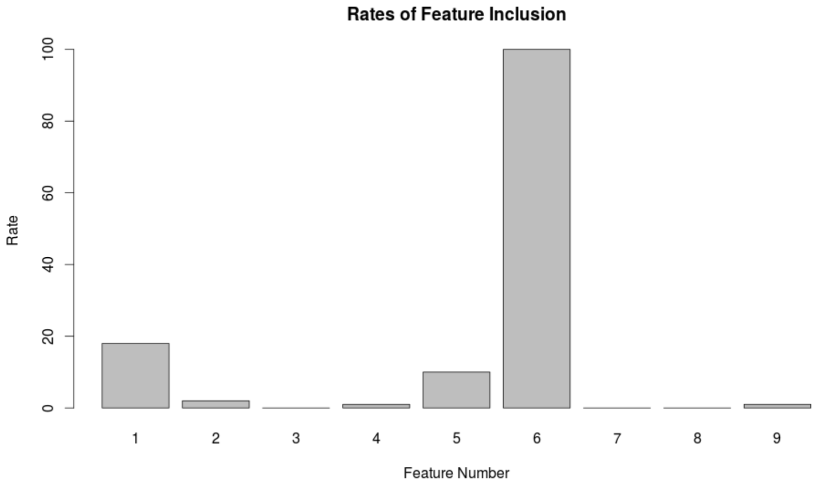


Figure 9: Feature Inclusion Rates for Breast Cancer Data. Features 1 and 6 are consistently the most important to the result, and cross-validated models with only these features explained 95.7% of the diagnoses correctly. Additive models with features 1, 5, and 6 achieved correct rate of 0.9628 (with cross-validation as well).

5.4 University of Wisconsin Breast Cancer Database

With 699 observations of 9 features, this data presents a binary classification problem in predicting the status of a tumor [7]. In this data, interactions are not of particular importance (and no interactions reported a significant r^2 drop after removal), but we run the full method to demonstrate the range of data we can fit. We ran multiple tests on additive feature importance to demonstrate stability, and results are reported below. In all 20 replicate runs, feature 6 was regarded as the most important to the fit of the model (feature 1, clump thickness, was also somewhat significant). Subsequent models where only feature 6 was included successfully predicted 91% of the observations in $\mathbf{Y}$ under a cross-validated model. Feature 6 represents a score for bare nuclei [7], and a result with such a degree of statistical significance suggests a deeper biological reason for such a result, that perhaps could be investigated further by researchers in this field. Full results, with selection rates and final classification error are presented in Figure 9.

6 Conclusion

Using the Random Kitchen Sinks technique in the context of the Fourier basis allows for a remarkable way to model nonlinear data, and (to the best of our knowledge) regularization-based methods for nonlinear feature selection in this framework have not yet been considered. Here we have proposed a use of the group lasso for variable selection in additive nonlinear models, and have extended this to nonlinear interactions, with considerations for appropriate modeling and selection of interactions, where the mechanics behind a similar use of the group lasso are detailed.

Issues with the group lasso such as computational complexity or extreme non-linearity prompt us to consider another approach in cases where this is not sufficient. We have proposed an approach that allows for variable selection that can be tuned to consider different levels of non-linearity while simultaneously providing a faster procedure for large numbers of predictors. After initial modeling where variables/interactions are included via a random binomial model, through insight into the mathematical structure of our model we can select important variables and interactions in a much more efficient way, and then use this to structure a more effective group lasso as described earlier. This allows us to tackle larger data sets more efficiently while not sacrificing any generality in variable selection. The data sets in some proposed applications also make it clear that the modeling and selection of nonlinear interactions containing two or even three variables is critical – complicated data where this is required can be found in many practical applications.

Our full technique consists of a combination of these two methods for selection of features and interactions. If we wish to consider high degrees of nonlinearity or improve efficiency, we measure the importance of variables/interactions through the correlation between predicted values after removing them, and then expand on these interactions with the group lasso. If p and n are extremely large, we may also restrict ourselves to an additive model.

We have compared this method to a variety of techniques, from a simplistic test of pairwise r^2 values to more advanced methods. We directly compared to simulated data provided by Rosasco *et al.*, concerned with selecting two relevant features from $p = 6$ and extended this for our own simulations, where the goal was to select five relevant variables from $p = 150$. In the comparison we saw results at least as good as those they reported, and demonstrated very successful results for the larger and more difficult examples we simulated.

We have also demonstrated very strong results for simulated benchmark data sets which are only concerned with prediction. However, the true significance of this work is in feature selection – in practice scientists want insight into causal factors and structure that explains the data they observe. To demonstrate potential applications, we have considered data from robot dynamics and a breast cancer database, among others. The mathematics presented here can be applied to a huge range of data sets, as we make no restrictions on whether our response is discrete or continuous, and the applications presented here are only a taste of what is possible. This approach allows us to not only model complicated data, but to understand it as well.

References

[1] Patrick Breheny and Jian Huang. "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors". In: *Statistics and Computing* (2013).

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

[3] Zoubin Ghahramani. *The* **pumadyn** *dataset*. http://www.cs.toronto.edu/~delve/data/pumadyn/pumadyn.ps.gz. Accessed: 2010-07-30. 1996.

[4] G.D. Konidaris, S. Osentoski, and P.S. Thomas. "Value Function Approximation in Reinforcement Learning using the Fourier Basis". In: *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence*. Aug. 2011, pp. 380–385. URL: http://lis.csail.mit.edu/pubs/konidaris-aaai11a.pdf.

[5] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. "The Randomized Dependence Coefficient". In: *Advances in Neural Information Processing Systems 26*. Ed. by C.J.C. Burges et al. Curran Associates, Inc., 2013, pp. 1–9. URL: http://papers.nips.cc/paper/5138-the-randomized-dependence-coefficient.pdf.

[6] David Lopez-Paz et al. "Randomized nonlinear component analysis". In: *Proc. of the 31st Int. Conf. Machine Learning (ICML)*. Ed. by Eric Xing and Tony Jebara. Curran Associates, Inc., 2014, pp. 1359–1367. URL: http://arxiv.org/pdf/1402.0119v2.pdf.

[7] O. L. Mangasarian and W. H. Wolberg. "Cancer diagnosis via linear programming". In: *SIAM News, Volume 23, Number 5* (1990), pp. 1–18.

[8] J. A. Nelder and R. W. M. Wedderburn. "Generalized Linear Models". English. In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), ISSN: 00359238. URL: http://www.jstor.org/stable/2344614.

[9] Alfonso Nieto-Castanon. *Random Gaussian Surface Generation*. https://www.mathworks.com/matlabcentral/answers/218806-random-gaussian-surface-generation. Accessed: 2015-07-20. 2015.

[10] Ali Rahimi and Benjamin Recht. "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller et al. Curran Associates, Inc., 2009, pp. 1313–1320. URL: http://papers.nips.cc/paper/3495-weighted-sums-of-random-kitchen-sinks-replacing-minimization-with-randomization-in-learning.pdf.

[11] David N. Reshef et al. "Detecting Novel Associations in Large Data Sets". In: *Science* 334.6062 (2011), pp. 1518–1524. DOI: 10.1126/science.1205438. eprint: http://www.sciencemag.org/content/334/6062/1518.full.pdf. URL: http://www.sciencemag.org/content/334/6062/1518.abstract.

[12] L. Rosasco et al. “A regularization approach to nonlinear variable selection.” In: *Proceedings of the 13 International Conference on Artificial Intelligence and Statistics* (2010).

[13] Noah Simon et al. “A sparse-group lasso”. In: *Journal of Computational and Graphical Statistics* (2013).

[14] Robert Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. In: *Journal of the Royal Statistical Society, Series B* 58 (1994), pp. 267–288.

[15] Ming Yuan and Yi Lin. *Model selection and estimation in regression with grouped variables.* 2006.