Energy Trees: Regression and Classification With Structured and Mixed-Type Covariates

Riccardo Giubilei  
Department of Statistics, Sapienza University of Rome, Rome, Italy  
Department of Economics and Finance, Luiss Guido Carli, Rome, Italy  
rgiubilei@luiss.it

Tullia Padellini  
DG Economics, Statistics and Research, Bank of Italy, Rome, Italy  
tullia.padellini@bancaditalia.it

Pierpaolo Brutti  
Department of Statistics, Sapienza University of Rome, Rome, Italy  
pierpaolo.brutti@uniroma1.it

Abstract

The increasing complexity of data requires methods and models that can effectively handle intricate structures, as simplifying them would result in loss of information. While several analytical tools have been developed to work with complex data objects in their original form, these tools are typically limited to single-type variables. In this work, we propose energy trees as a regression and classification model capable of accommodating structured covariates of various types. Energy trees leverage energy statistics to extend the capabilities of conditional inference trees, from which they inherit sound statistical foundations, interpretability, scale invariance, and freedom from distributional assumptions. We specifically focus on functional and graph-structured covariates, while also highlighting the model’s flexibility in integrating other variable types. Extensive simulation studies demonstrate the model’s competitive performance in terms of variable selection and robustness to overfitting. Finally, we assess the model’s predictive ability through two empirical analyses involving human biological data. Energy trees are implemented in the R package etree.

Keywords: nonparametric methods, supervised learning, functional data, graphs, complex data

1 Introduction

Increasingly often in data analysis, quantities of interest are complex objects living in non-Euclidean spaces, including curves, graphs, shapes, images, and strings. A popular approach to analyzing these objects is the translation into Euclidean feature vectors in order to apply “standard” statistical techniques (Jain and Obermayer, 2009). The main limitation is that a universally valid way of obtaining such a representation does not exist (Jain and Obermayer, 2009), leading to arbitrary choices and loss of information.

As the urgency of analyzing complex variables is growing, several frameworks for analyzing structured data without further simplification have been developed (Wang and Marron, 2007; Jain and Obermayer, 2009; Marron and Alonso, 2014). Particularly relevant is the
case of object-oriented data analysis (OODA) (Wang and Marron, 2007), which has been initially applied to tree-structured data objects (Wang and Marron, 2007), becoming later successful in functional data analysis (Sangalli et al., 2009). Additional examples in the field of structured data object analysis include graphs (Ginestet et al., 2017; Zhou and Müller, 2022), persistence diagrams (Bendich et al., 2016), manifolds (Lila et al., 2016; Lila and Aston, 2020), sounds (Pigoli et al., 2018; Tavakoli et al., 2019), images (Benito et al., 2017), covariance matrices (Dryden et al., 2009; Pigoli et al., 2014), and probability distributions (Chen et al., 2021; Petersen et al., 2021).

While “analyzing data objects directly we avoid loss of information that occurs when data objects are transformed into numerical summary statistics” (La Rosa et al., 2016, p. 1), most existing contributions focus solely on single-type data objects. Consequently, the techniques are often domain-specific and cannot be easily extended to other data types. Even more critically, they do not support the joint analysis of multiple complex sources.

The pioneering work of Balakrishnan and Madigan (2006) considers structured and mixed-type covariates, but it focuses on a single type of structured variables, requires domain-specific expertise, and lacks any concept of statistical significance. Brandi (2018) and Nespoli (2019) also moved in the direction of a unifying learning framework for structured and mixed-type data, proposing two models that partially share building principles with this article. However, they have both proposed a very specific version that does not account for mixed-type data—not even including traditional types—and only allows for one type of structured covariates: functions in the first case, graphs in the second. Additionally, these works lack in-depth investigations into the model’s structure, design principles, properties, and performance.

In this work, we introduce energy trees as a new and more general class of decision trees. The model has sound statistical foundations and provides a unifying framework to perform classification and regression with structured and mixed-type covariates. Energy trees draw essential features from conditional inference trees (Hothorn et al., 2006) and energy statistics (Szekely and Rizzo, 2013). The tree structure facilitates the analysis of mixed-type variables, and the use of association tests, as in conditional trees, ensures statistically principled splits. Energy statistics enable assessing the association between variables of different and possibly structured types.

The article is organized as follows. Section 2 describes the structure, the properties, and the algorithm of energy trees. Specifically, Section 2.2 and Section 2.3 focus on the two crucial steps of variable selection and splitting, respectively. Section 3 explains how to leverage the model’s flexibility to accommodate any type of covariate. The properties and the performance of energy trees are demonstrated through various numeric applications with simulated (Section 4) and empirical (Section 5) data. Section 6 includes a brief recapitulation and ideas for future work.

1. In this article, mixed-type does not refer to the dichotomy between numeric and categorical, but more generally to the case with covariates of different types. As a further matter of terminology, numeric and categorical types are referred to as traditional, which is intended as opposed to structured, including any other type.
2 Energy Trees

Energy trees belong to the class of recursive partitioning models, or trees. They share a similar structure with conditional trees (Hothorn et al., 2006), which find splits employing permutation tests to estimate the conditional distribution of statistics measuring the association between dependent and explanatory variables. This approach allows overcoming two critical issues of traditional tree-based methods: selection bias and overfitting. In energy trees, the association is evaluated using energy tests of independence (Székely et al., 2007) from the energy statistics framework. This fairly general class of tests allows assessing the association between variables defined in spaces that are not necessarily Euclidean and not necessarily the same.

Energy trees have several advantageous properties that derive from their constituent parts. As a tree-structured model, they are easily interpretable, scale-invariant, do not require preprocessing or distributional assumptions, can simply handle missing values, and provide automatic feature selection. Grounded in a conditional inference framework similar to that of conditional trees, they have statistically sound foundations, avoid selection bias, and exhibit robustness to overfitting. Finally, energy statistics enable the analysis of covariates that are potentially structured and of different types, positioning energy trees as a unifying framework for regression and classification with any kind of variable.

Energy trees are implemented in the R package etree, which is available on CRAN at https://CRAN.R-project.org/package=etree and on GitHub at https://github.com/ricgbl/etree.

2.1 Structure and Algorithm

Energy trees take as input a learning sample \( L_n = \{(Y_i, X_{1i}, \ldots, X_{Ji}); i = 1, \ldots, n\} \) consisting of a response variable with support \( \mathcal{Y} \) and a set of covariates \( X = (X_1, \ldots, X_J) \). For regression, \( \mathcal{Y} \subseteq \mathbb{R} \); in the case of classification, \( \mathcal{Y} \) is a discrete set of labels \( \{1, \ldots, K\} \). The domain of the \( j \)-th covariate, \( j = 1, \ldots, J \), is denoted with \( X_j \) and does not necessarily have to be a vector space. This means that the model can handle not only traditional variables but also structured ones in the form of complex objects such as strings, graphs, or functions.

When growing any type of tree based on a generic \( L_n \), the observations \( 1, \ldots, n \) are recursively partitioned into nodes that eventually determine which units have similar behavior regarding the response variable. Each node can be split into two or more kid nodes to whom observations are assigned based on the value taken in a single covariate. The initial node is called root, while the ones without kid nodes are defined terminal nodes. Each node in the tree is represented by a vector of case weights \( w = (w_1, \ldots, w_n) \), whose generic element \( w_i \) is a non-negative integer that indicates the number of times the \( i \)-th observation appears in the node. Consequently, within node \( w \), the response variable is \( Y^w = ([Y_1]_{xw_1}, \ldots, [Y_n]_{xw_n}) \), and the \( j \)-th covariate is \( X^w_j = ([X_{j1}]_{xw_1}, \ldots, [X_{jn}]_{xw_n}) \), where \( [a]_{xb} \) indicates the repetition of element \( a \) for \( b \) times (or its absence if \( b = 0 \)).

The recursive partitioning algorithm for energy trees can be summarized as follows:

1. **Stopping criterion.** For node \( w \), test the null hypothesis of global independence between the response variable \( Y^w \) and all the covariates \( X^w_j \), \( j = 1, \ldots, J \), as \( H_0 = \cap_{j=1}^{J} H_0^j \), where the \( J \) partial hypotheses \( H_0^j : D(Y^w|X^w_j) = D(Y^w) \) regarding the
distribution $D(\cdot)$ are verified using energy tests of independence (Szétkely et al., 2007). If $H_0$ is not rejected at a pre-specified level $\alpha$, stop.

2. **Variable selection.** Select the $j^*$-th covariate $X_j^w$ that exhibits the strongest association with the response $Y^w$, based on the energy test of independence for $H_0^j$.

3. **Split.** Determine a non-empty set $A^* \subset X_j^*$ to partition $X_j^*$ into $A^*$ and $X_j^* \setminus A^*$. The resulting case weights $w_{\text{left}}$ and $w_{\text{right}}$ define the two subgroups (one for each kid node) and have their elements computed as $w_{\text{left},i} = w_i \cdot I(X_j^* \in A^*)$ and $w_{\text{right},i} = w_i \cdot I(X_j^* \notin A^*)$ for all $i = 1, \ldots, n$, where $I(\cdot)$ is the indicator function.

4. Repeat steps 1, 2 and 3 on nodes $w_{\text{left}}$ and $w_{\text{right}}$, respectively.

Despite the similarity with the algorithm of conditional trees, two key differences enable the generalization to structured and mixed-type covariates. In step 1, the $J$ partial hypotheses $H_0^j : D(Y^w|X_j^w) = D(Y^w)$ are verified using energy tests of independence instead of permutation tests; in step 3, determining the set $A^* \subset X_j^*$ is remarkably more complicated. Section 2.2 describes in greater detail steps 1 and 2, while Section 2.3 addresses step 3.

### 2.2 Variable Selection

Steps 1 and 2 of the algorithm involve verifying the $j$-th partial hypothesis $H_0^j : D(Y^w|X_j^w) = D(Y^w)$ of independence between $Y^w$ and $X_j^w$, for $j = 1, \ldots, J$. Each hypothesis is verified with an energy test of independence between the response and the $j$-th covariate, using the standard choice of $nV_n^2$ as the test statistic, where $n$ is the sample size and $V_n$ is the sample distance covariance (Szétkely et al., 2007). Thus, for each node $w$, the sample distance covariance needs to be calculated between each covariate and the response.

Traditionally, the sample distance covariance is obtained by computing, for each variable, the Euclidean distance between couples of observations (Szétkely et al. 2007; Szétkely and Rizzo, 2013). However, energy trees allow for covariates in more complex spaces, requiring a generalized notion of distance between observations. Formally, consider node $w$ and denote its size with $m$, i.e., $\sum_{i=1}^n w_i = m$. The goal is to verify the $j$-th partial hypothesis $H_0^j : D(Y^w|X_j^w) = D(Y^w)$. Let $X_jk$, $k = 1, \ldots, m$, be the generic element of $X_j^w$, and $Y_k$, $k = 1, \ldots, m$, that of $Y^w$. Take a distance $\delta(\cdot)$ on the covariate’s space $X_j^*$ and the Euclidean norm $|\cdot|$ on $\mathcal{Y} \subseteq \mathbb{R}$. The observed test statistic $mV_m^2(X_j^w, Y^w)$ for node $w$ is obtained as follows:

1. Compute the distance matrices defined by
   
   $$(a_{kl}) = \delta(X_{jk}, X_{jl}), \quad k, l = 1, \ldots, m,$$
   
   $$(b_{kl}) = |Y_k - Y_l|, \quad k, l = 1, \ldots, m;$$

2. For both matrices, calculate row means, column means, and global mean, e.g.,

   $$\bar{a}_k = \frac{1}{m} \sum_{l=1}^m a_{kl}, \quad \bar{a}_l = \frac{1}{m} \sum_{k=1}^m a_{kl}, \quad \bar{a}_* = \frac{1}{m^2} \sum_{k,l=1}^m a_{kl};$$
3. Compute the centered distances, e.g.,

\[ A_{kl} = a_{kl} - \bar{a}_k - \bar{a}_l + \bar{a}_\cdot, \quad k, l = 1, \ldots, m; \]

4. Calculate the square of the sample distance covariance \( V_m(X^w, Y^w) \) as

\[ V_m^2(X^w, Y^w) = \frac{1}{m^2} \sum_{k,l=1}^{m} A_{kl} B_{kl}; \]

5. Compute the value of the observed test statistic for node \( w \) as \( m V_m^2(X^w, Y^w) \).

The procedure involves the calculation of \( m(m-1)/2 \) distances for both the covariate and the response variable, resulting in a computational time complexity of \( O(m^2) \).

The sampling distribution of the test statistic \( m V_m^2(X^w, Y^w) \) under the null hypothesis depends on the unknown joint distribution of \( X^w \) and \( Y^w \). Hence, it is estimated using permutation tests, that is, computing replicates of the test statistic under random reshuffles of the indices of \( Y^w \). The procedure is repeated for each covariate, yielding a p-value \( P_j \) for each partial hypothesis \( H_j^0 \).

In step 1 of the energy trees algorithm, the test of global independence between the response variable \( Y^w \) and all the covariates \( X^w_j, j = 1, \ldots, J \), is formalized as \( H_0 = \cap_{j=1}^J H_j^0 \). The p-value of the global test can be computed starting from \( P_j, j = 1, \ldots, J \), and using p-values adjustment techniques for multiple testing procedures. Energy trees adopt the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). As opposed to Bonferroni correction—used in conditional trees—that controls the probability of at least one false rejection, FDR regulates the expected proportion of false rejections among all rejections, so it has less conservative control of type I error but also greater power. After correction, \( H_0 \) is rejected if the minimum of the adjusted p-values is less than a pre-specified nominal level \( \alpha \); otherwise, the recursion stops. Hence, as in conditional trees, \( \alpha \) may be interpreted not only as the nominal level controlling type I error in each node but also as a tunable hyperparameter determining the size of energy trees.

If the hypothesis of global independence is not rejected in step 1, the algorithm proceeds to step 2 (variable selection). The covariate selected for splitting is the one that yields the smallest p-value, i.e., \( X^w_{j^*} \) is such that \( j^* = \arg\min_{j=1, \ldots, J} P_j \).

### 2.3 Splits

The goal of step 3 is to use the splitting variable \( X^w_{j^*} \) to divide observations into two subgroups. Many methods can accomplish this when \( X^w_{j^*} \) is either numeric or categorical; Energy trees perform an energy test of independence for each possible split point. For numeric covariates, the independence test is conducted between the response variable and a binary vector that indicates which units would be assigned to the first kid node based on that split point. Note that the construction of the binary vector leverages the natural ordering among numeric values. In the case of nominal covariates, a natural ordering may not exist. Hence, the binary vector is obtained by considering each instance of all non-trivial combinations of the splitting variable’s categories. In both cases, the optimal split point is
the one yielding the strongest association in terms of p-value with the response. The idea of performing the test using a binary vector representing the split follows Hothorn et al. (2006), and it induces a two-sample statistic that maximizes the discrepancy between the distribution of the response in the two kid nodes.

Formally, the goal is to find the optimal $A^* \subset X_j^*$, with $A^* \neq \emptyset$, to partition the splitting variable’s domain $X_j^*$ into two sets: $A^*$ and $X_j^* \setminus A^*$. For traditional variables, $A^*$ is searched among the non-empty and proper subsets $Q$ of $X_j^*$, where $Q$ belongs to a reasonable set $Q$. In energy trees, for each $Q$ in $Q$, an energy test of independence is performed between the response variable and the binary vector given by

$$\ell_Q = \left[ I(X_{j^*1} \in Q) \right]_{x_1}, \ldots, \left[ I(X_{j^*n} \in Q) \right]_{x_n}. \quad (1)$$

The optimal $A^*$ is selected as the $Q$ whose corresponding $\ell_Q$ shows the strongest association in terms of p-value with $Y^w$.

The form of $Q$ depends on the type of the splitting variable. In the numeric case, consider the $k$ unique values of $X_j^*$ for which $w_i \neq 0$, and let $x(1), \ldots, x(k)$ be the corresponding sorted vector. The set $Q$ is a right-closed interval, $Q = (-\infty, q]$, where $q = x(1), \ldots, x(k-1)$. In other terms,

$$Q = \{(-\infty, x(1)], \ldots, (-\infty, x(k-1)]\}. \quad (2)$$

In the nominal case, let $M = \{1, \ldots, m\}$ represent the levels of $X_j^*$ for which $w_i \neq 0$. In this case, $Q$ is any element of

$$Q = \mathcal{P}(M) \setminus \{\emptyset, M\}, \quad (3)$$

where $\mathcal{P}(M)$ is the power set of $M$, and the trivial cases $Q = \emptyset$ and $Q = M$ are excluded. Moreover, due to the complementary nature of kid nodes’ subspaces in binary partitioning models, additional cases can be ignored.

The problem becomes more complicated if $X_j^*$ is structured, such as in the case of curves, graphs, shapes, images, or strings. These types of observations do not have a natural ordering or an obvious way to split them into two subgroups. Therefore, alternative strategies for splitting must be considered, as discussed in the remaining part of this section.

### 2.3.1 Feature vector extraction

The first presented method is called feature vector extraction. It consists of finding the split after applying a transformation to switch from the complex sample space $X_j^*$ of structured data objects to a more tractable Euclidean feature space. The name of the method derives from the popular practice to represent structured objects through Euclidean feature vectors (Jain and Obermayer, 2009).

In the specific context, the transformation depends on the type of the splitting variable and is a function $g_j : X_j^* \rightarrow \mathbb{R}^{s_j}$. It involves expanding a structured covariate into real-valued coefficients, or components, similarly to basis expansion in linear algebra or for the particular case of functional variables (Ramsay and Silverman, 2005). Hence, the transformation of the splitting variable into coefficients is referred to as coefficient expansion in the following.
Let $X_{j,k}$, $k = 1, \ldots, m$, be the generic element of $X_{j}^{w}$, and denote by $b_{k}^{j}$ the $s_{j} \times 1$ vector resulting from the transformation $g_{j}$ applied to $X_{j,k}$, i.e., $g_{j}(X_{j,k}) \mapsto b_{k}^{j} = (b_{k1}^{j}, \ldots, b_{ks_{j}}^{j})^{T}$. If $X_{j}^{w}$ consists of $m$ objects, the transposed collection of vectors $B^{j} = [b_{1}^{j} \cdots b_{m}^{j}]^{T}$ is a $m \times s_{j}$ matrix. The matrix $B^{j}$ can be also expressed as the collection of the components it contains, i.e., $B^{j} = [b_{1}^{j} \cdots b_{s_{j}}^{j}]$; in other words, for $s = 1, \ldots, s_{j}$, $b_{s}^{j} = (b_{s1}^{j}, \ldots, b_{sm}^{j})^{T}$ is the $s$-th component resulting from the transformation of $X_{j}^{w}$ via $g_{j}$.

Feature vector extraction solves the issue of splitting structured covariates, but the resulting $B^{j} = g_{j}(X_{j}^{w})$ is an $m \times s_{j}$ matrix that includes the $s_{j}$ components $b_{s}^{j}$, $s = 1, \ldots, s_{j}$. To handle this multidimensional problem, the approach replicates the setup of multiple (originally) numeric covariates. First, the most associated component $b_{s}^{j}$, is selected through an energy test of independence between the response variable $Y^{w}$ and each component $b_{s}^{j}$, $s = 1, \ldots, s_{j}$. This process is as described in Section 2.2, except that no stopping criterion is used here. Then, the split point for the selected real-valued $b_{s}^{j}$, is determined similarly to numeric covariates. The optimal $A^{*} \subset X_{j}^{*}$ is the subspace $Q$ corresponding to the binary vector $\ell_{Q}$ from Equation (1) that, among all $Q$ in $Q$ as defined in Equation (2), yields the strongest association with the response variable $Y^{w}$ in an energy test of independence.

While the use of feature vector extraction for splits may initially seem counterintuitive in a context where the goal is to analyze data in their original form, “it can be quite hard to directly understand population structure using the object space alone. Thus, it is useful to simultaneously consider the (closely linked) feature space as well” (Marron and Alonso, 2014, p. 734). In other words, combining the two perspectives may lead to improved results and more meaningful interpretations (Marron and Dryden, 2021, ch. 3).

### 2.3.2 Clustering

Feature vector extraction is a valuable technique; however, it necessarily implies loss of information. Since the splitting step implies partitioning the observations, a natural alternative strategy is to use clustering, as suggested by Balakrishnan and Madigan (2006). Specifically, distance-based clustering methods allow keeping the splitting variable in its original form and involve two steps: first, identify two (or more) representative data objects, or medoids; second, assign other observations to the cluster that minimizes the distance from the corresponding medoid.

Let $X_{j}^{w}$ be the splitting covariate, $X_{j}^{*}$ its sample space, and $\delta(\cdot)$ the distance defined on $X_{j}^{*}$. The optimal $A^{*} \subset X_{j}^{*}$, with $A^{*} \neq \emptyset$, for splitting $X_{j}^{*}$ into two sets $A^{*}$ and $X_{j}^{*} \setminus A^{*}$ is determined by first identifying two medoids $C_{1}$ and $C_{2}$, one for each kid node. Then, other observations are assigned to the kid node that corresponds to the closest medoid in terms of $\delta(\cdot)$. Specifically, the kid node to which the $k$-th element $X_{j,k}$ of $X_{j}^{w}$ is assigned is given by

$$\text{argmin}_{c \in \{C_{1}, C_{2}\}} \delta(X_{j,k}, c).$$

(4)

Note that Equation (4) implies that the optimal $A^{*}$ can be defined as the Voronoi region associated with one of the two medoids; e.g., for $C_{1}$,

$$A^{*} = \{x \in X_{j}^{*} : \delta(x, C_{1}) \leq \delta(x, C_{2})\}.$$

(5)
Many clustering techniques are based on finding representative observations among a set of structured data objects. Energy trees employ partitioning around medoids (PAM) (Kaufmann and Rousseeuw, 1987), also known as \textit{k-medoids}. The computational complexity of PAM is $O(n^2)$, similarly to many other distance-based clustering algorithms. Among these, it is worth mentioning k-groups (Li and Rizzo, 2017), which lies within the energy statistics framework. However, PAM is preferred because it has well-established faster variants such as CLARA (Kaufman and Rousseeuw, 2008) and FastPAM (Schubert and Rousseeuw, 2019).

The clustering approach to splits offers the advantage of working directly with data objects without using arbitrary transformations. It enables the implementation of multiway splits, which may be particularly useful in multi-class problems, by considering more than two representative observations simultaneously. Its computational time is preferable to the $O(n^3)$ complexity of feature vector extraction. However, an important drawback is the lack of any concept of statistical significance for the splitting step. This is in contrast to feature vector extraction, where splits are based on statistical tests. Moreover, feature vector extraction provides enhanced interpretability by allowing focused analysis of specific aspects concerning single components. The conclusion is that no dominant strategy exists, and the optimal approach should be determined application-wise.

### 3 Definition of Distances and Coefficient Expansion Methods

Energy trees are designed to accommodate covariates of various types. In the following, \textit{traditional} refers to numeric and categorical variables, while \textit{structured} encompasses any other type, such as functional and in the form of graphs. Each type of covariate requires an appropriate distance, which is necessary for computing the test statistic in the energy tests of independence employed for variable selection. It is also needed for structured covariates when using the clustering approach to splitting. Alternatively, when using feature vector extraction for splits, structured covariates necessitate a suitable method for coefficient expansion. Finally, any type of traditional covariate requires the distance for categorical variables to determine the split point, as outlined in Equation (1).

Table 1 provides the choices of distance and coefficient expansion for each type of covariate considered in this paper’s applications. While standard methods apply for numeric, nominal, and functional variables, no natural technique exists in the literature when data are in the form of graphs (Marron and Dryden, 2021, p. 69). Energy trees use the edge difference distance (Hammond et al., 2013), defined as the Frobenius norm of the difference between the adjacency matrices of the two graphs, and the shell distribution (Carmi et al., 2007), obtained via $k$-cores (Seidman, 1983), $s$-cores (Eidsaa and Almaas, 2013), and $d$-cores (Giatsidis et al., 2013), for binary, weighted, and directed graphs, respectively. Further details can be found in Appendix A.

One notable advantage of energy trees is their high flexibility. The choices of distance and coefficient expansion can be adjusted according to personal preferences or application-specific needs. Moreover, the model can easily accommodate other types of covariates as long as the corresponding distance is specified. This implies that many types of variables equipped with their own distance can be incorporated into the framework defined by energy trees. Examples include shapes and covariance matrices with Procrustes distance,
Trees With Structured and Mixed-Type Covariates

| Type      | Distance       | Coefficient exp. |
|-----------|----------------|------------------|
| Numeric   | Euclidean      | NA               |
| Nominal   | Gower          | NA               |
| Functional| $L^2$-norm     | Cubic B-splines  |
| Graphs    | Edge difference| Shell distribution|

Table 1: Choices of distance and coefficient expansion for the four types of covariates considered in this work.

...manifolds with Riemannian distance, images with image Euclidean distance (Wang et al., 2005), time series and sounds with dynamic time warping, probability distributions with $f$-divergences, strings with edit distances, and data objects in general metric spaces with Gromov-Hausdorff distance or fused Gromov-Wasserstein distance (Vayer et al., 2020). It is important to mention that distance covariance characterizes independence only for metric spaces of strong negative type (Lyons, 2013), though it can be extended to semimetric spaces of negative type (Sejdinovic et al., 2013). However, even when these conditions are not met, distance covariance can still be interpreted as a loose measure of association.

4 Simulation Study

Energy trees are unbiased, robust to overfitting, and select meaningful covariates. This section presents three simulation scenarios that empirically validate these properties. The setup extends the original work on conditional trees (Hothorn et al., 2006) to the case of structured and mixed-type covariates. Experiments are conducted using 10,000 replications and calculating 95% confidence intervals through the normal approximation to the binomial distribution.

4.1 Unbiasedness

In a recursive partitioning model, *unbiasedness* refers to the selection of covariates $X_1, \ldots, X_J$ with equal probabilities of $1/J$ under the null hypothesis of global independence between the response variable and predictors (Hothorn et al., 2006). To empirically demonstrate unbiasedness, it is necessary to verify if each covariate is selected with approximately the same relative frequency under independence. In this case, where no association exists between the response and the covariates, the root split is forced by removing any stopping criterion. The response variable $Y$ follows a standard normal distribution $\mathcal{N}(0,1)$, while the covariates are specified as:

- $X_1$. Numeric: uniformly distributed between 0 and 1;
- $X_2$. Nominal: binary variable with uniformly-sampled values;
- $X_3$. Functions: Gaussian random processes over 100 evaluation points ranging from 0 to 1, with a mean of 0 and the identity matrix as the covariance matrix;
Energy trees are compared to decision trees and conditional trees in terms of performance. However, since these models cannot handle structured covariates, feature vector extraction is employed to transform them into numeric components (see Section 3 and Appendix A for details). Additional adjustments are explained in Appendix B. The results of the scenario are presented in Table 2. Decision trees are known to exhibit bias towards covariates with many possible splits: the least selected covariates are $X_2$, which takes only two values, and $X_3$, whose components have few split points. Conditional trees display far less bias but fail to include the reference probability 0.25 within any of the approximated 95% confidence intervals. Energy trees yield estimates that are very close to 0.25 for all covariates and include this value within the approximated 95% confidence intervals. They are unbiased regardless of the measurement scale or type of covariates. Consequently, energy trees stand out as the only model among the three that exhibits unbiasedness in the general case of structured and mixed-type covariates.

| Covariate    | Decision trees | Conditional trees | Energy trees |
|--------------|----------------|-------------------|--------------|
| $X_1$ (Numeric) | 0.4846 (0.4748, 0.4944) | 0.2604 (0.2518, 0.2690) | 0.2505 (0.2420, 0.2590) |
| $X_2$ (Nominal) | 0.0293 (0.0260, 0.0326) | 0.2741 (0.2654, 0.2828) | 0.2492 (0.2407, 0.2577) |
| $X_3$ (Functions) | 0.3954 (0.3858, 0.4050) | 0.1901 (0.1824, 0.1978) | 0.2503 (0.2418, 0.2588) |
| $X_4$ (Graphs) | 0.0907 (0.0851, 0.0936) | 0.2754 (0.2666, 0.2842) | 0.2500 (0.2415, 0.2585) |

Table 2: Simulated point estimates and approximated 95% confidence intervals for the relative frequencies of variable selection under independence between the response and the covariates when no stopping criterion is applied.

4.2 Overfitting and Selection of Meaningful Covariates

In recursive partitioning models that employ statistical tests of independence for variable selection, power represents the probability of selecting any covariate, rather than stopping, under the alternative hypothesis of association with the response. Power can be examined by analyzing the behavior of the probability of selecting any variable (without forcing the split) as the association between the response and a specific covariate increases. When independence holds, the probability should be close to zero. As the association grows, it can be properly interpreted as the power of the independence test, so larger values indicate higher power. Another relevant quantity is the conditional probability of selecting the associated covariate, given that any variable is chosen for splitting. In the case of independence, each covariate should have an equal probability of being selected, resulting in a conditional
probability of \( 1/J \) for all \( j = 1, \ldots, J \). As the association between the response and a covariate grows, the conditional probability should also increase. The power analysis assesses the robustness to overfitting, focusing on the ability to perform splits only when necessary. On the other hand, the conditional probability analysis concentrates on the selection of meaningful covariates.

The simulation scheme for both analyses is similar to Section 4.1, except for \( Y \) and one covariate. The response variable \( Y \) follows a normal distribution with unit variance and a mean \( \mu = 0 \) for half the observations, while the other half have \( \mu \in [0, 1] \). The association between the response and one explanatory variable is induced by increasing the value of \( \mu \) within the interval \([0, 1]\). Since the focus of this work is on structured covariates, the associated predictor is initially the functional variable \( X_3 \) and then the graph-structured variable \( X_4 \). In the first case, \( X_3 \) is defined using the same two groups of \( Y \): half of the observations are realizations of a Gaussian random process over 100 evaluation points from 0 to 1, with a mean of 0 and the identity matrix as the covariance matrix, while the other half has a mean of 0.5. In the second case, half of the observations are Erdős–Rényi random graphs with 100 vertices and a connection probability of 0.2, while the other half have a connection probability of 0.8. When the associated covariate is \( X_j \), where \( j = 3, 4 \), the configuration of \( X_i \), for \( i = 1, \ldots, 4 \) and \( i \neq j \), is the same as in Section 4.1.

The results of the power and conditional probability analyses for decision trees, conditional trees, and energy trees are presented in Figure 1. When the associated covariate is \( X_3 \), the estimated probability of selecting any covariate for \( \mu = 0 \) is bounded above by the critical threshold \( \alpha = 0.05 \) for conditional trees (0.0432) and decision trees (0.0453), and approximately for energy trees (0.0507). However, the power curve for energy trees widely dominates the other two models across the entire range. The conditional probability for \( \mu = 0 \) closely approximates the reference value of 0.25 only for energy trees (0.2702), while it is lower for conditional trees (0.2231) and substantially higher for decision trees (0.4585). In this case, the curve for energy trees is uniformly greater than the others since \( \mu = 0.2 \); before this value, the superiority of decision trees is artificially induced by their bias. When the associated covariate is \( X_4 \), the estimated probability of selecting any covariate for \( \mu = 0 \) is bounded above by the critical threshold \( \alpha = 0.05 \) only for conditional trees (0.0406) and energy trees (0.0436), and approximately for decision trees (0.0511). As the association increases, the power curve for energy trees dominates the other two. The conditional probability for \( \mu = 0 \) is relatively close to the reference value of 0.25 for energy trees (0.2110) and conditional trees (0.2894), while it is substantially lower for decision trees (0.0554). When the response is associated with the graph-structured covariate, the conditional probability curve for energy trees is uniformly higher than the other two models.

For both types of structured covariates, energy trees limit the proportion of incorrect decisions in the root node to \( \alpha \) when the response is independent of the predictors. Additionally, in this case, the proportion of selection for the covariate of interest given an incorrect split is approximately equal to the reference probability of 0.25. When the response is associated with one of the covariates, energy trees exhibit higher power and more frequently select the correct covariate (excluding low levels of association) compared to the competitors. Consequently, the two analyses demonstrate that energy trees are not only robust to overfitting and capable of selecting meaningful variables when dealing with structured and mixed-type covariates but also preferable to the other two models in these respects.
Figure 1: Simulated power and conditional probability curves using four covariates of different types. First row: functional covariate associated with the response. Second row: graph-structured covariate associated with the response. For each model, point estimates and approximated 95% confidence intervals are shown. The horizontal dashed lines represent the significance level of $\alpha = 0.05$ on the left and the reference probability of 0.25 on the right.
5 Real-World Data Applications

The predictive ability of energy trees is validated through two empirical analyses using real-world data. The first one focuses on a classification task that aims to detect knee osteoarthritis based on bones’ shape and demographic information. The second analysis is a regression problem whose goal is predicting the intelligence quotient using multimodal brain connectomes and demographic information. These analyses not only substantiate the flexibility of the energy tree model but also demonstrate its wide applicability across various fields, including human biology and medicine.

5.1 Knee Osteoarthritis Classification

Knee osteoarthritis (OA) is a painful and debilitating condition with a poorly understood etiology. However, the shape of the bones is a relevant risk factor because it directly influences the biomechanics of the joint. Two previous studies (Shepstone et al., 2001; Ramsay and Silverman, 2007) used the shape of the femur’s intercondylar notch, a deep fossa between two protrusions on the femur end closer to the knee joint, to distinguish between 21 OA femora and 75 non-OA (NOA) control femora. In the binary classification task conducted by Ramsay and Silverman (2007), the shape of the \( j \)-th intercondylar notch was transformed into two functional covariates, \( X_j(t) \) and \( Y_j(t) \), representing the longitudinal and latitudinal coordinates. These covariates were discretized to capture functional values at 50 equally-spaced points \( (t_1,\ldots,t_{50}) \) along the curves. The data set (see Figure 2) also includes the age and gender for each of the 96 notches. Age is expressed as a binary variable, indicating whether the individual is older than 45 or not, and gender is also binary.

Ramsay and Silverman (2007) considered two models that can only handle functional data, hence they discarded the two nominal covariates. The first model, functional linear discriminant analysis (FLDA), treated the 100 coordinates along the two curves as numeric variables, performed principal component analysis to reduce feature dimensionality, and used linear discriminant analysis on the resulting components to classify knees. The second model, mean difference projection (MPD), calculated a mean curve for each coordinate and each subpopulation (OA and NOA), and projected all the data onto the direction of the difference between the mean curves. Energy trees are implemented using feature vector extraction as the splitting method, and considering different combinations of significance level \( \alpha \) and minimum number \( \nu \) of observations in each terminal node: \( \alpha \in \{0.1, 0.2, \ldots, 1\} \) and \( \nu \in \{1, 5, 10\} \). To ensure comparability, energy trees are tuned and evaluated using leave-one-out cross-validation (LOO-CV), similarly to Ramsay and Silverman (2007). Initially, only the two functional covariates are used for fitting. The results, in terms of binary classification performance metrics with OA cases as the positive class, are presented in Table 3. Energy trees with \( \alpha = 0.9 \) and \( \nu = 10 \) outperform competitors in terms of accuracy, specificity, and positive predicted value (PPV). However, they perform worse than FLDA for sensitivity, negative predicted value (NPV), and balanced accuracy. In other words, when using only the two functional covariates, energy trees are the best model for correctly classifying NOA knees (at the expense of identifying only 43% of OA knees and making 20% overall errors), while FLDA is the best model for recognizing OA knees (at the expense of incorrectly classifying as OA more knees than those correctly identified and making 27% overall errors).
Figure 2: Covariates used for knee OA classification. First row: observed values of the functional variable for the X coordinate (solid lines) and mean curves for the two groups (dashed lines). Second row: same, but for the Y coordinate. Third row: bar plots for nominal variables gender and older.
While FLDA and MPD can only handle functional covariates and were specifically designed for the particular task, one of the greatest advantages of energy trees is that they provide a unifying framework for supervised learning with structured and mixed-type data. Considering the two nominal covariates as well, energy trees with both splitting methods outperform the two competitors for all the metrics (see Table 4). The best parameter combination is $\alpha = 0.5$ and $\nu = 5$ for feature vector extraction, and $\alpha = 0.9$ and $\nu = 10$ for clustering. In the specific case, the two splitting methods focus on different aspects: feature vector extraction is preferable for recognizing OA cases, while clustering correctly classifies a larger number of NOA cases. Another study (Balakrishnan and Madigan, 2006), which employed decision trees allowing for functional covariates by performing clustering-based splits, analyzed the full set of covariates\(^2\). Their model achieved an accuracy of 80.21%, which is comparatively worse than the 82.29% and 84.38% achieved by energy trees using the two splitting strategies.

|     | Acc. | Sens. | Spec. | PPV  | NPV  | B.Acc. |
|-----|------|-------|-------|------|------|--------|
| FLDA | 72.92| 66.67 | 74.67 | 42.42| 88.89| 70.67  |
| MPD  | 64.58| 57.14 | 66.67 | 32.43| 84.75| 61.90  |
| ET   | 80.21| 42.86 | 90.67 | 56.25| 85.00| 66.76  |

Table 3: Performance metrics (%) in LOO-CV for FLDA, MPD, and energy trees (ET) using only the two functional covariates.

|     | Acc. | Sens. | Spec. | PPV  | NPV  | B.Acc. |
|-----|------|-------|-------|------|------|--------|
| ET (FVE) | 82.29| 66.67 | 86.67 | 58.33| 90.28| 76.67  |
| ET (C)  | 84.38| 57.14 | 92.00 | 66.67| 88.46| 74.57  |

Table 4: Performance metrics (%) in LOO-CV for ET with feature vector extraction (FVE) and clustering (C) as the splitting strategies, using the full set of covariates.

An example of a single energy tree fitted through LOO-CV using the full set of covariates and feature vector extraction as the splitting method is given in Figure 3. The component selected for the split in the functional covariate representing the Y coordinate is the 5-th, while the split in the variable denoting the X coordinate uses the 9-th. These are the only two splits directly producing terminal nodes with a different balance between OA and NOA cases.

### 5.2 Connectome-IQ Regression

Intelligence is commonly measured using tests that provide an intelligence quotient (IQ) score. Extensive research has demonstrated an association between high IQ scores and the coordinated activation of multiple brain regions, as observed through both structural

\(^{2}\) Cf. \url{http://archive.dimacs.rutgers.edu/Research/MMS/PAPERS/fdt17.pdf}
(Haier et al., 2004; Jung and Haier, 2007) and functional (Gray et al., 2003; Lee et al., 2006) neuroimaging techniques. Recent studies have investigated the neural basis of intelligence explicitly at the connectivity level (Li et al., 2009; Hilger et al., 2017; Dubois et al., 2018). Inspired by these findings, the current analysis aims to investigate the regression relationship between IQ scores and various variables including structural and functional connectomes. The data used in this study derive from the Rockland-sample study (Nooner et al., 2012) conducted by the Nathan Kline Institute. Collecting and matching information from various sources, we have formed a data set with 159 observations and four covariates (see Figure 4). Two of the covariates are graph-structured, representing the functional and structural connectomes of individuals as undirected and weighted graphs with 188 nodes each. The remaining two covariates are traditional variables: age (with a mean of 36.43 ± 20.09) and gender (67 females and 92 males). The response variable is the full-scale IQ, a scalar measure obtained for each subject in the Wechsler abbreviated scale of intelligence test, with a mean of 109.74 ± 12.97.

Figure 3: Example of classification energy tree fitted in LOO-CV using feature vector extraction as the splitting method.
Figure 4: Covariates used in the regression analysis. First two rows: multidimensional scaling plots for the functional and structural graph covariates, respectively. Third row: bar plot for the nominal variable gender and histogram for the numeric variable age.
Similarly to Section 4, energy trees are fitted and compared with two other recursive partitioning models: decision trees and conditional trees. These competitors cannot handle structured covariates, hence requiring the transformation of graphs into feature vectors, which is performed using the shell distribution based on scores (see Section 3 and Appendix A for details). Energy trees are implemented using both splitting methods. For energy trees and conditional trees, the stopping criteria \( \nu \) and \( \alpha \) are tuned, while decision trees use \( \nu \) and a complexity parameter \( cp \) that plays a role similar to \( \alpha \) in determining the tree size. To evaluate the models in an unbiased way, a nested 5-fold cross-validation (CV) procedure is adopted: the inner folds are used for parameter tuning, and the outer folds for performance assessment. The optimal parameter combination is selected based on the root mean square error (RMSE), averaged across the inner folds. The parameter values explored for the three models are the following: \( \nu \in \{4, 7, \ldots, 25\} \), \( \alpha \in \{0.05, 0.08, \ldots, 0.20\} \), and \( cp \in \{0.005, 0.008, \ldots, 0.020\} \). Table 5 presents the results in terms of the RMSE over the outer folds. The performance of energy trees with the two splitting methods is identical, and it results in an improvement of 2% over decision trees and of 0.5% over conditional trees.

|       | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Average |
|-------|--------|--------|--------|--------|--------|---------|
| ET (FVE) | 11.69  | 13.47  | 13.98  | 11.57  | 13.70  | 12.88   |
| ET (C)   | 11.69  | 13.47  | 13.98  | 11.57  | 13.70  | 12.88   |
| DT       | 13.58  | 13.28  | 13.52  | 11.57  | 13.91  | 13.14   |
| CT       | 11.69  | 13.47  | 13.98  | 11.87  | 13.70  | 12.94   |

Table 5: RMSE over the outer folds of the nested 5-fold CV for energy trees with both splitting methods, decision trees (DT), and conditional trees (CT).

Figure 5 illustrates an example of a single energy tree fitted using the outer folds of the nested 5-fold CV, with clustering as the splitting method. The tree hierarchy and the relative frequency of splits concerning graph-structured covariates confirm the relevant role of brain connectomes in the predictive task under consideration.

6 Concluding Remarks

Energy trees offer a unifying framework for classification and regression with structured and mixed-type data. They possess several advantageous properties derived from their constituent elements. As a recursive partitioning model, they are interpretable, scale-invariant, do not require preprocessing or parametric assumptions, have built-in feature selection, and can handle missing values, as well as covariates of different types. Using independence tests for variable selection and possibly for splitting ensures statistically sound foundations. The incorporation of energy statistics enables analyzing structured covariates. Furthermore, the simulation study has shown that energy trees exhibit unbiasedness, robustness to overfitting, and the ability to identify meaningful covariates for splitting. The model’s predictive ability has been validated through two different experimental settings using real-world data. The ultimate advantage of energy trees lies in their great flexibility. This enables users to 1)
choose between two alternative splitting strategies, 2) change the distance or the coefficient expansion technique as required, and 3) accommodate any other type of covariate.

The simulation and empirical analyses not only confirm the need, as mentioned in Section 1, for a model that can handle multiple complex sources, but also demonstrate that energy trees have competitive predictive performance compared to traditional models. Two main reasons account for this outcome: firstly, the ability to leverage all available data and retain each covariate, as shown in Section 5.1, where it has made a significant impact; secondly, the decision to analyze covariates in their most natural form, avoiding the loss of information that occurs when transforming them into feature vectors, which has instead affected competitors in Section 4 and Section 5.2. The simulation scenarios have proven that energy trees are unbiased regardless of the measurement scale or the type of covariate, unlike decision trees (which are biased) and conditional trees (which are only unbiased when working with traditional types). Additionally, energy trees display greater power and more frequently select the correct covariate across various levels of association compared to the other two models. The two empirical analyses are just two examples of the wide applicability of energy trees. The model could be also employed for image recognition, sentiment analysis, disease diagnosis, fraud detection, market segmentation, credit scoring, recom-
mendation systems, and event detection, as well as for regression tasks in econometrics, finance, environmental sciences, geostatistics, social sciences, biostatistics, climate sciences, sports analytics, psychology, genetics, artificial intelligence, and many other fields. The two splitting approaches, although structurally different, have demonstrated equal validity and may even yield identical results (see Section 5.2). Yet, they may focus on different aspects (see Section 5.1), so it is preferable to compare the two and select the most suitable one based on the specific application’s characteristics and goals. Finally, the parameter $\alpha$ can be determined in a data-dependent way, as shown in Section 5, if prediction accuracy is the primary focus. Nevertheless, Section 4 has proven that the classical value of $\alpha = 0.05$ performs well compared to potential competitors.

Much room is left for improvement. The theoretical derivation of the statistical properties of energy trees is desirable but challenging because the setting naturally involves diverse complex objects and mathematical spaces. Conducting a sensitivity analysis for input parameters would be crucial to gain a deeper understanding of the model’s internal mechanisms and potentiality. The case of a structured response variable would be straightforward to implement in many respects, such as replacing the Euclidean norm with an appropriate distance, but developing a meaningful approach for summarizing each type of structured variable is essential to enable predictions. Conversely, the potential for incorporating new types of structured covariates and exploring novel applications is virtually limitless. Finally, it would be worthwhile to investigate and analyze ensemble methods, such as boosting, bagging, and random forests, that use energy trees as base learners.

Data Availability

The data supporting the findings of the analysis in Section 5.1 can be openly accessed on the companion website to Ramsay and Silverman (2007) at http://www.stats.ox.ac.uk/~silverma/fdacasebook/notchchap.html. IQ scores and phenotypical information of participants for the analysis in Section 5.2 are available at http://fcon_1000.projects.nitrc.org/indi/pro/nki.html, which is the Rockland-sample study page on the 1000 Functional Connectomes Project site managed by NeuroImaging Tools and Resources Collaboratory (NITRC). Raw DTI and fMRI neuroimaging data can be downloaded at https://www.nitrc.org/frs/?group_id=404, which is the File Release Download page on the NITRC site, under the section labeled Nathan Kline Institute. The preprocessed version of this data—in the form of structural and (no GSR) functional connectomes (Brown et al., 2012), as used in this paper—was previously available online and can now be shared upon reasonable request with the permission of the original authors (Brown et al., 2012).

Disclosure Statement

The authors declare no funding support or competing interests to disclose. The views expressed in this paper solely belong to the authors and do not involve any responsibility of the Bank of Italy and/or the Eurosystem.
Appendix A. Distance and Coefficient Expansion for Selected Types of Covariates

Section 3 and Table 1 present the distances and coefficient expansion methods for the four types of covariates considered in this paper: numeric, nominal, functional, and graph-structured. Further details and mathematical formulations are provided below.

A.1 Distance

Each type of covariate requires a specific distance to compute the test statistic in the energy tests of independence used for variable selection. Additionally, traditional covariates need the distance for categorical variables to search for the split point, while structured covariates necessitate their own distance when performing splits with the clustering approach.

For traditional types numeric and nominal, energy trees use Euclidean and Gower’s distance, respectively. Let $X_j$ be a numeric variable, meaning that observed values $X_{jk}$ and $X_{jl}$ are scalars. The distance $\delta_1(\cdot)$ between numbers is the Euclidean distance $\delta_1(X_{jk}, X_{jl}) = |X_{jk} - X_{jl}|$.

If $X_j$ is a nominal variable, $X_{jk}$ and $X_{jl}$ are categories from a discrete set $\{1, \ldots, K\}$. The distance $\delta_2(\cdot)$ between nominal data objects is the Gower’s distance $\delta_2(X_{jk}, X_{jl}) = I(X_{jk} \neq X_{jl})$.

When $X_j$ is functional, the $i$-th observed value $X_{ji}$ is a well-defined curve $f_i(t)$ where $t \in \mathcal{T}$ represents the evaluation points. The distance $\delta_3(\cdot)$ between functional observations $X_{ji} = f_i(\cdot)$ and $X_{jl} = f_l(\cdot)$ is the $L^2$ norm $\delta_3(f_i, f_l) = \| f_i - f_l \|_2 = \left( \int_{\mathcal{X}} |f_i(t) - f_l(t)|^2 \, dt \right)^{1/2}$, where $\mathcal{X}$ is the space over which $f_i(\cdot)$ and $f_l(\cdot)$ are defined.

If $X_j$ is a graph-structured variable, the $i$-th observed value is a graph $G_i$. Any graph $G$ can be denoted as $G = (V, E)$, where $V$ is a set of vertices and $E$ is a set of edges. Each pair of vertices $u, v \in V$ can be represented as $e = \{u, v\}$ and has an associated edge weight $w_e$, which is non-zero if and only if $e$ joins vertices $u, v \in V$, i.e., $e \in E$. Graph $G$ can be described through a $|V| \times |V|$ adjacency matrix $A$ whose generic entry is the edge weight $w_{\{u, v\}}$ between $u$ and $v$, where $u, v = 1, \ldots, |V|$. In this work, the distance $\delta_4(\cdot)$ between graph-structured observations $X_{ji} = G_i$ and $X_{jl} = G_l$ is the edge difference distance (Hammond et al., 2013), which is defined as the Frobenius norm of the difference between the two adjacency matrices. In symbols,$$
\delta_4(G_i, G_l) = \| A^i - A^l \|_F = \sqrt{\sum_u \sum_v \left| e^i_{\{u, v\}} - e^l_{\{u, v\}} \right|^2},$$
where $\| \cdot \|_F$ denotes the Frobenius norm, $A^i$ and $A^l$ are the adjacency matrices of $G_i$ and $G_l$ respectively, and $e^i_{\{u, v\}}$ and $e^l_{\{u, v\}}$ are the corresponding generic entries. Edge difference distance ensures wide applicability to binary, signed, weighted, and directed graphs, while maintaining computational efficiency and providing reasonable results in various settings.
A.2 Coefficient Expansion

Each type of structured covariate requires a specific coefficient expansion method when performing the splits through feature vector extraction. The types of structured covariates considered for this work’s applications are functional and in the form of graphs.

A functional data object $f_i(·)$ can be represented by real-valued components using a basis, which is a set of linearly independent vectors $\phi_s(·)$, with $s = 1, \ldots, S$. Bases allow approximating arbitrarily well the function as a linear combination

$$f_i(t) \approx \sum_{s=1}^{S} c_{is}\phi_s(t), \quad (6)$$

where $c_{is}$ is the coefficient of the $s$-th element of the basis for $f_i(·)$. Since bases are usually such that the vector $c_i = (c_{i1}, \ldots, c_{iS})$ fully specifies the data object, $c_i$ itself can be naturally used as the set of components.

Using the notation of Section 2.3.1, suppose that splitting variable $X_{w_j}$ is functional, meaning that its generic element $X_{j+k}$ is a function $f_k(·)$. The transformation $g_j(·)$ used for any functional covariate is such that

$$g_j(f_k) \mapsto c_k = (c_{k1}, \ldots, c_{ks_j})^T,$$

where $c_k$ is derived by representing $f_k(·)$ using Equation (6) with a basis of $s_j$ elements.

Energy trees use cubic B-splines (Ramsay and Silverman, 2005) as the basis with generic element $\phi_s(·)$ in Equation (6). Splines are commonly used for approximating non-periodic functional data because they are smooth, efficient, flexible, and parsimonious (Ramsay and Silverman, 2005). Cubic splines correspond to the lowest order that guarantees two continuous derivatives, meaning that both the representation of the function and its first derivative are smooth. Among splines, the B-splines basis system is the most popular (Ramsay and Silverman, 2005).

Techniques for performing coefficient expansion on graphs are less established than those for functional data. For the simplest case of undirected (symmetric adjacency matrix) and unweighted (binary adjacency matrix) graphs, energy trees employ the notion of $k$-core decomposition (Seidman, 1983). The method adequately captures the graph’s global connectivity structure (Seidman, 1983; Carmi et al., 2007), and is based on transforming the graph into a set of $k$-cores. The $k$-core of a graph $G$, denoted as $C_k(G)$, is defined as the maximal subgraph where every vertex has at least degree $k$.

To obtain a single vector from the set of $k$-cores, energy trees use the definition of shell index (Carmi et al., 2007): a vertex $v \in G$ has shell index $i$ if $v \in C_i(G)$, but $v \notin C_{i+1}(G)$. In other words, the shell index of a vertex $v$ represents the highest core to which $v$ belongs. Representing $k$-cores as shell indices reduces dimensionality, and these indices can be collected in a single vector that characterizes the entire graph. Specifically, a graph $G$ can be represented using a $|V|$-dimensional vector whose $j$-th entry is the number $s_j$ of vertices of $G$ that have shell index $j$, for $0 \leq j \leq |V| - 1$. Such a vector is called shell distribution of the graph $G$ and can be denoted as $s(G)$.

Suppose that the splitting variable $X_{w_j}$ is graph-structured, with the generic $X_{j+k}$ being a graph $G_k$. The transformation used for any covariate in the form of graphs is such that
\[ g_j(G_k) \mapsto s_k = (s_{k1}, \ldots, s_{k|V_k|})^T, \]

where \( s_k \equiv s(G_k) \), and \( V_k \) is the set of vertices of graph \( G_k \).

The notion of \( k \)-core was originally introduced for unweighted and undirected graphs but has been extended to both the weighted and the directed cases with \( s \)-cores (Eidsaa and Almaas, 2013) and \( d \)-cores (Giatsidis et al., 2013), respectively. On the other hand, the definitions of shell index and shell distribution remain the same. This ensures that energy trees can transform splitting variables in the form of weighted or directed graphs into real-valued components.

**Appendix B. Comparison With Traditional Competitors**

Since energy trees are introduced to overcome the limitations of traditional models in handling structured covariates, comparing their performance in simulation settings is not a straightforward task. One approach is to use feature vector extraction to transform any structured covariate into Euclidean features that competitors can handle. However, this raises two problems. Firstly, the number \( s_j \) of components deriving from feature vector extraction is, in general, different for each covariate \( j \). Secondly, the number of components for the same covariate may differ across the simulation runs due to specific variable realizations. To address these challenges, it is necessary to correct the relative frequencies of selection of any component accounting for these variations.

The same procedure applies to the unbiasedness analysis, where the root split is forced by removing any stop criterion, and to the power and conditional probability analyses, where splits are not necessarily performed. In the latter case, it suffices to condition the analysis on the actual selection of any covariate.

Let \( S \) represent the event of selecting a given component, and \( A \) denote the event of that component being available. Since \( S \subset A \), the conditional probability of selecting a specific component is \( P(S|A) = P(S)/P(A) \). This implies that estimating \( P(S|A) \) requires dividing the relative frequency of selection of the component by the corresponding relative frequency of availability. Consequently, the relative frequency of selection for the covariate can be obtained by averaging these quantities across all components.

The notation introduced in Section 2.3.1 can be used to formalize these ideas. Recall that \( b_j^s \) represents the \( s \)-th component resulting from the transformation of the \( j \)-th variable \( X_j^w \) via \( g_j \). Note that \( s = 1, \ldots, s_j \), since the number of components depends on the specific transformation \( g_j \), and \( j = 1, \ldots, J \). The estimated probability of selecting the \( j \)-th covariate \( X_j^w \) is obtained as a weighted average

\[ \tilde{p}_j = \frac{1}{s_j} \sum_{s=1}^{s_j} \frac{\tilde{f}_j^s}{\tilde{a}_j^s}, \]

where \( \tilde{f}_j^s \) is the relative frequency of selecting feature \( b_j^s \), and \( \tilde{a}_j^s \) is the relative frequency of availability for feature \( b_j^s \).
However, summing $\tilde{p}_j$ from Equation (7) across $j = 1, \ldots, J$ does not necessarily yield 1. A solution is normalizing each $\tilde{p}_j$ to obtain the proper relative frequencies of selection of the $j$-th covariate,

$$\hat{p}_j = \frac{\tilde{p}_j}{\sum_{j=1}^{J} \tilde{p}_j}. \quad (8)$$

The quantities $\hat{p}_j$ from Equation (8), with $j = 1, \ldots, J$, are then used to compare the performance of traditional models with that of energy trees.

References

Suhrid Balakrishnan and David Madigan. Decision trees for functional variables. In International Conference on Data Mining, pages 798–802, 2006.

Paul Bendich, James S Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. The Annals of Applied Statistics, 10(1):198, 2016.

Mónica Benito, Eduardo García-Portugués, James S Marron, and Daniel Peña. Distance-weighted discrimination of face images for gender classification. Stat, 6(1):231–240, 2017.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1):289–300, 1995.

Marco Brandi. Classification and Regression Energy Tree for Functional Data. PhD thesis, Sapienza University of Rome, 2018.

Jesse A Brown, Jeffrey D Rudie, Anita Bandrowski, John D Van Horn, and Susan Y Bookheimer. The UCLA multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis. Frontiers in Neuroinformatics, 6:28, 2012.

Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of Internet topology using k-shell decomposition. Proceedings of the National Academy of Sciences, 104(27):11150–11154, 2007.

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. Journal of the American Statistical Association, pages 1–14, 2021.

Ian L Dryden and Kanti V Mardia. Statistical Shape Analysis: With Applications in R. John Wiley & Sons, 2016.

Ian L Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. The Annals of Applied Statistics, 3(3):1102–1123, 2009.

Julien Dubois, Paola Galdi, Lynn K Paul, and Ralph Adolphs. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1756):20170284, 2018.
Marius Eidsaa and Eivind Almaas. $s$-core network decomposition: a generalization of $k$-core analysis to weighted networks. *Physical Review E*, 88(6):062819, 2013.

Christos Giatsidis, Dimitrios M Thilikos, and Michalis Vazirgiannis. $d$-cores: measuring collaboration of directed graphs based on degeneracy. *Knowledge and Information Systems*, 35(2):311–343, 2013.

Cedric E Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2):725–750, 2017.

Jeremy R Gray, Christopher F Chabris, and Todd S Braver. Neural mechanisms of general fluid intelligence. *Nature neuroscience*, 6(3):316–322, 2003.

Richard J Haier, Rex E Jung, Ronald A Yeo, Kevin Head, and Michael T Alkire. Structural brain variation and general intelligence. *Neuroimage*, 23(1):425–433, 2004.

David K Hammond, Yaniv Gur, and Chris R Johnson. Graph diffusion distance: a difference measure for weighted graphs based on the graph Laplacian exponential kernel. In *IEEE Global Conference on Signal and Information Processing*, pages 419–422, 2013.

Kirsten Hilger, Matthias Ekman, Christian J Fiebach, and Ulrike Basten. Intelligence is associated with the modular structure of intrinsic brain networks. *Scientific Reports*, 7(1):1–12, 2017.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

Brijnesh J Jain and Klaus Obermayer. Structure spaces. *Journal of Machine Learning Research*, 10(11), 2009.

Rex E Jung and Richard J Haier. The parieto-frontal integration theory (P-FIT) of intelligence: converging neuroimaging evidence. *Behavioral and Brain Sciences*, 30(2):135, 2007.

Leonard Kaufman and Peter J Rousseew. Clustering large applications (program CLARA). *Finding groups in data: an introduction to cluster analysis*, pages 126–163, 2008.

Leonard Kaufmann and Peter Rousseuw. Clustering by Means of Medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 1987.

Patricio S La Rosa, Terrence L Brooks, Elena Deych, Berkley Shands, Fred Prior, Linda J Larson-Prior, and William D Shannon. Gibbs distribution for statistical analysis of graphical data with a sample application to fcMRI brain images. *Statistics in Medicine*, 35(4):566–580, 2016.

Kun Ho Lee, Yu Yong Choi, Jeremy R Gray, Sun Hee Cho, Jeong-Ho Chae, Seungheun Lee, and Kyungjin Kim. Neural correlates of superior intelligence: stronger recruitment of posterior parietal cortex. *Neuroimage*, 29(2):578–586, 2006.
Songzi Li and Maria L Rizzo. K-groups: a generalization of K-means clustering. *arXiv preprint arXiv:1711.04359*, 2017.

Yonghui Li, Yong Liu, Jun Li, Wen Qin, Kuncheng Li, Chunshui Yu, and Tianzi Jiang. Brain anatomical network and intelligence. *PLoS Computational Biology*, 5(5):e1000395, 2009.

Eardi Lila and John AD Aston. Statistical analysis of functions on surfaces, with an application to medical imaging. *Journal of the American Statistical Association*, 115(531):1420–1434, 2020.

Eardi Lila, John AD Aston, and Laura M Sangalli. Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *The Annals of Applied Statistics*, 10(4):1854–1879, 2016.

Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

James S Marron and Andrés M Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753, 2014.

James S Marron and Ian L Dryden. *Object Oriented Data Analysis*. Chapman and Hall/CRC, 2021.

Gabriel Nespoli. Classification and regression energy tree with network predictors. Master’s thesis, Sapienza University of Rome, 2019.

Kate Brody Nooner, Stanley Colcombe, Russell Tobe, Maarten Mennes, Melissa Benedict, Alexis Moreno, Laura Panek, Shaquanna Brown, Stephen Zavitz, Qingyang Li, et al. The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, 6:152, 2012.

Alexander Petersen, Xi Liu, and Afshin A. Divani. Wasserstein $F$-tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics*, 49(1):590 – 611, 2021.

Davide Pigoli, John AD Aston, Ian L Dryden, and Piercesare Secchi. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 2014.

Davide Pigoli, Pantelis Z Hadjipantelis, John S Coleman, and John AD Aston. The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages. *Journal of the Royal Statistical Society: Series C*, 67(5):1103–1145, 2018.

James O Ramsay and Bernard W Silverman. *Functional Data Analysis*. Springer, 2nd edition, 2005.

James O Ramsay and Bernard W Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, 2007.
Laura M Sangalli, Piercesare Secchi, Simone Vantini, and Alessandro Veneziani. A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, 104(485):37–48, 2009.

Erich Schubert and Peter J Rousseeuw. Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In *International Conference on Similarity Search and Applications*, pages 171–187. Springer, 2019.

Stephen B Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

Lee Shepstone, Jeffrey Rogers, John R Kirwan, and Bernard W Silverman. Shape of the intercondylar notch of the human femur: a comparison of osteoarthritic and non-osteoarthritic bones from a skeletal sample. *Annals of the Rheumatic Diseases*, 60(10):968–973, 2001.

Gábor J Székely and Maria L Rizzo. Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Shahin Tavakoli, Davide Pigoli, John AD Aston, and John S Coleman. A spatial modeling approach for linguistic object data: analyzing dialect sound variations across Great Britain. *Journal of the American Statistical Association*, 114(527):1081–1096, 2019.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

Haonan Wang and James S Marron. Object oriented data analysis: sets of trees. *The Annals of Statistics*, 35(5):1849–1873, 2007.

Liwei Wang, Yan Zhang, and Jufu Feng. On the Euclidean distance of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1334–1339, 2005.

Yidong Zhou and Hans-Georg Müller. Network regression with graph Laplacians. *Journal of Machine Learning Research*, 23(320):1–41, 2022.