Machine Learning (2019) 108:1613–1634
https://doi.org/10.1007/s10994-019-05820-3

CADET: interpretable parametric conditional density estimation with decision trees and forests

Cyrus Cousins[1] · **Matteo Riondato**[2]

Received: 22 October 2018 / Revised: 11 June 2019 / Accepted: 20 June 2019 / Published online: 26 June 2019
© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2019

Abstract
We introduce CADET, an algorithm for *parametric Conditional Density Estimation* (CDE) based on decision trees and random forests. CADET uses the *empirical cross entropy* impurity criterion for tree growth, which incentivizes splits that improve predictive accuracy more than the regression criteria or estimated mean-integrated-square-error used in previous works. CADET also admits more efficient training and query procedures than existing tree-based CDE approaches, and stores only a bounded amount of information at each tree leaf, by using *sufficient statistics* for all computations. Previous tree-based CDE techniques produce complicated uninterpretable distribution objects, whereas CADET may be instantiated with easily interpretable distribution families, making every part of the model easy to understand. Our experimental evaluation on real datasets shows that CADET usually learns more accurate, smaller, and more interpretable models, and is less prone to overfitting than existing tree-based CDE approaches.

Keywords Parametric models · Random forests · Sufficient statistics

When I became a cadet, I immediately decided I wanted to be an undercover cop because I don't like uniforms. — Ron Stallworth

Editors: Karsten Borgwardt, Po-Ling Loh, Evimaria Terzi, Antti Ukkonen.

Part of the work done while the authors were affiliated with Two Sigma Investments, LP.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10994-019-05820-3) contains supplementary material, which is available to authorized users.

✉ Matteo Riondato
mriondato@amherst.edu

Cyrus Cousins
cyrus_cousins@brown.edu

[1] Department of Computer Science, Brown University, Providence, USA

[2] Department of Computer Science, Amherst College, Amherst, USA

1 Introduction

Conditional Density Estimation (CDE) is a fundamental statistical task. Given a domain $\mathcal{X}$, a codomain $\mathcal{Y}$, and a joint Probability Density Function[1] (PDF) $\rho(\cdot, \cdot)$ over $\mathcal{X} \times \mathcal{Y}$, the CDE task is to estimate, for each $x \in \mathcal{X}$, the *Conditional (probability) Density Function* $\rho(\cdot|x)$. CDE estimators are inductively learned from a training set Z, which is a collection of n pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from the distribution arising from $\rho(\cdot, \cdot)$.

Classical regression estimates only the *conditional expectation* $\mathbb{E}[y|x]$, whereas CDE estimates the *conditional distribution* of y given x. Depending on the application, conditional density estimates can be used as-is, or their quantiles, moments, and other statistics can be computed, making CDE more flexible than regression. Regressors often assume homoskedasticity, while CDE methods handle heteroskedasticity, and can thus describe complicated phenomena like skew or multimodality. CDE can be also used when regression is meaningless, e.g., when conditional densities exhibit *heavy tailed* power-law distributions with undefined *expectation*.

We focus our attention on *interpretable CDE*, where the task is to train an accurate CDE model such that *both the model and its estimates* are easy to understand by a human analyst. Interpretability is difficult to quantify, but in our context, having *small representation size* and *low query complexity* is a necessary condition, and a reasonable proxy for interpretability, as the analyst should be able to conceptualize or visualize the entire model, and mentally follow the process by which queries are answered. Decision trees and random forests naturally satisfy these requisites. It is also necessary that *density estimates* are interpretable, as understanding the model and query process is only beneficial if the analyst can also understand the actual predictions. Simple parametric distributions are interpretable at a glance, but large mixture models, non-parametric estimates, and complex graphical models, while computationally convenient, are largely uninterpretable.

Existing tree-based[2] CDE techniques learn uninterpretable models, and often select splits that do not yield even local improvements to CDE accuracy. These techniques must store *all training labels* associated with each leaf in order to answer queries, yielding high storage costs and query time complexities. Probabilistic graphical models with tree structure address some of these issues, and bear some resemblance to decision trees, but inference on them is far more complicated, and the learned models are less interpretable to the human analyst.

Contributions We present CADET, a CDE algorithm based on decision trees and random forests that overcomes the above limitations of existing tree-based CDE approaches, and produces *interpretable parametric* conditional density estimates.

- CADET trees are standard decision trees that use parametric distributions stored at the leaves to answer conditional density queries. While parametric CDE methods are less expressive than non-parametric methods, they usually require less training data, better leverage domain knowledge, and are more interpretable, as they store fewer parameters and produce simpler estimates.
- CADET trees use the *empirical cross-entropy* impurity criterion for tree growth, which directly incentivizes splits that lead to more accurate estimates than the criteria used by existing tree-based CDE techniques. We show that CADET generalizes information-gain classification trees and mean-square-error-minimizing regression trees to a broad family of parametric CDE estimators.

[1] We make no assumptions on the measure space beyond the existence of densities, thus our model covers continuous, discrete, and mixed spaces.

[2] We use "tree-based" to describe both decision-tree- and random-tree-based techniques.

– CADET is the first tree-based CDE technique that can answer queries without requiring complex graphical model operations or iterating over stored training labels. Instead, each leaf stores a fixed-size *sufficient statistic* for the labels of training points mapped to it, which allows CADET to perform Maximum Likelihood Estimation (MLE) as though it had access to the training labels.
– By selecting parametric families with appropriate support, CADET can handle both *univariate and multivariate* CDE, as well as CDE on more exotic spaces, such as directional spaces, probability simplices, and mixed spaces, whereas non-parametric methods are generally restricted to particular domain types.
– Our experimental evaluation on real datasets shows that CADET produces models that are generally more accurate, less prone to overfitting, and are more interpretable than existing tree-based CDE techniques.

Outline The paper is organized as follows. An introduction to decision trees and random forests is given in Sect. 2. We discuss related work in Sect. 3. CADET, is presented in Sect. 4, followed by extensions to the basic algorithm in Sect. 5. We present our experimental comparison of CADET to existing tree-based CDE techniques in Sect. 6. Some conclusions complete the work in Sect. 7.

2 Decision trees and random forests

We now define the key concepts about decision trees and random forests. Our description of these data structures and the learning procedure is sufficiently general to encompass various learning tasks, including regression, classification, and CDE.

Decision trees As in the introduction, consider a domain $\mathcal{X}$ and a codomain $\mathcal{Y}$, and let Z be the training set, which is a collection of n pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. A *decision tree* T is a *strict rooted binary tree* such that:

1. each non-leaf node v stores a *split rule* s_v that maps each element of $\mathcal{X}$ to either the left or the right child of v, splitting $\mathcal{X}$ into two. For any node u (leaves included), there is a subset of $\mathcal{X}$ that is mapped to u. Any $x \in \mathcal{X}$ is mapped to all the nodes found by walking down the tree T starting from the root and following the split rule at each encountered non-leaf node. For any node u, we denote with $\mathsf{T}(u; T, Z)$ the subset of Z that is mapped to u. For any $x \in \mathcal{X}$ we denote with $\mathsf{tl}(x; T)$ the *leaf* of T that x is mapped to;
2. each leaf ℓ stores some information $\mathsf{L}(\ell; T)$, a set of values whose role we describe below. $\mathsf{L}(\ell; T)$ is a function of $\mathsf{T}(\ell; T, Z)$, the elements of Z that T maps to ℓ.

As an example, in standard regression trees with numeric features, a split rule s_v at a non-leaf node v is a *univariate threshold function*, which is an indicator function for an inequality on the value of a single feature, such as "age ≤ 4." Elements that satisfy the condition are mapped to the left child of v, the others to the right child. In the same scenario, the information $\mathsf{L}(\ell; T)$ stored at a leaf ℓ is the mean of the $\mathcal{Y}$ components of $\mathsf{T}(\ell; T, Z)$.

We are usually interested in the leaf information or in the set of training points associated with the leaf containing some query point $x \in \mathcal{X}$, so we abuse notation, taking $\mathsf{L}(x; T)$ to mean $\mathsf{L}\big(\mathsf{tl}(x; T); T\big)$, and $\mathsf{T}(x; T, Z)$ to mean $\mathsf{T}\big(\mathsf{tl}(x; T); T, Z\big)$.

Query answering Decision trees and random forests are used to answer *queries*. For a decision tree T (we discuss forests later) that makes predictions in some codomain $\mathcal{U}$, queries are answered with the function $\mathsf{q}(\cdot; T) : \mathcal{X} \to \mathcal{U}$, where for $x \in \mathcal{X}$, $\mathsf{q}(x; T)$ is computed

using the information $\mathsf{L}(x; T)$ stored at the leaf to which T maps x. In univariate regression, $\mathcal{U} = \mathcal{Y} = \mathbb{R}$, and the query response is simply $\mathsf{q}(x; T) = \mathsf{L}(x; T)$, but in general $\mathcal{U}$ may be different than $\mathcal{Y}$ (e.g., in probabilistic classification, $\mathcal{Y}$ is a discrete set, and $\mathcal{U}$ contains *distributions over* $\mathcal{Y}$), and the leaf information may be used in various ways to respond to queries.

Impurity criterion The split rule in each non-leaf node is learned using the training set. Before describing the learning procedure, we introduce *impurity criteria*, which are functions $\mathsf{m} : \mathcal{Y}^n \to \mathbb{R}$. For a set of training labels $Y \in \mathcal{Y}^n$, $\mathsf{m}(Y)$ is the *impurity value* of Y, which is usually a proxy for the *average loss* that any constant prediction would incur over Y. We often abuse notation, taking $\mathsf{m}(Z)$ for $Z \in (\mathcal{X} \times \mathcal{Y})^n$ to simply ignore $\mathcal{X}$, and compute the impurity over the $\mathcal{Y}$ components of Z.

The *Mean Square Error* (MSE) *impurity*, used in regression trees (Breiman et al. 1984), is

$$\mathsf{m}_{\mathsf{mse}}(Y) = \frac{1}{|Y|} \sum_{y \in Y} \left(y - \bar{Y}\right)^2, \text{ where } \bar{Y} = \frac{1}{|Y|} \sum_{y \in Y} y \ . \tag{1}$$

Taking $\hat{\mathbb{P}}(i)$ to be the *sample frequency* (in Y) of $i \in \mathcal{Y}$, the (discrete) *entropy impurity*, used in information-gain classification trees (Quinlan 1986), is

$$\mathsf{m}_{\mathsf{H}}(Y) = -\sum_{i \in \mathcal{Y}} \hat{\mathbb{P}}(i) \ln\left(\hat{\mathbb{P}}(i)\right) \ . \tag{2}$$

These impurities correspond to the *square loss* and *cross entropy* loss of regressors and probabilistic classifiers, respectively, though they may also be interpreted as measures of dispersion at the leaves of a decision tree. Under either interpretation, by selecting splits to minimize total leaf impurity, decision trees seek to explain as much variation in $\mathcal{Y}$ through $\mathcal{X}$ as possible.

Some tree-based CDE methods (Pospisil and Lee 2018) use the *Mean Integrated Square Error* (MISE) impurity, defined as

$$\mathsf{m}_{\mathsf{mise}}(Y) = \frac{1}{|Y|} \sum_{y \in Y} \int_{\mathcal{Y}} \left(\hat{\rho}_B(y) - \rho(y|x)\right)^2 \mathrm{d}y,$$

where $\hat{\rho}_B(\cdot)$ is the *estimated density* computed using B, and $\rho(\cdot|x)$ is the *true conditional density* given x. While this impurity criterion incentivizes returning $\rho(\cdot|x)$ as the estimate, computing $\mathsf{m}_{\mathsf{mise}}(Y)$ requires knowledge of $\rho(\cdot|x)$ itself, so Pospisil and Lee (2018) approximate these true densities with a cosine or tensor basis non-parametric estimate. Since estimating $\rho(\cdot|x)$ is the goal of CDE, using the MISE impurity creates a cyclic dependency that is not easily resolved.

Learning procedure The learning procedure builds the tree starting from the root. It chooses a split rule s_v for the current node v and creates its two children. To choose s_v, it finds the partitioning of $\mathsf{T}(v; T, Z)$ into L and R that *maximizes*, over some family of partitionings (such as the univariate thresholds mentioned above for regression trees), the *impurity reduction* w.r.t. $\mathsf{m}(\cdot)$, defined as

$$\left|\mathsf{T}(v; T, Z)\right| \mathsf{m}\left(\mathsf{T}(v; T, Z)\right) - \left(|R|\mathsf{m}(R) + |L|\mathsf{m}(L)\right) \ . \tag{3}$$

The split rule s_v stored at v is then chosen in such a way as to be consistent with the partitioning of $\mathsf{T}(v; T, Z)$ into L and R. The procedure recursively splits each child v until a stopping criterion is met. Example criteria include the depth of v exceeding a user-specified threshold,

the impurity reduction falling short of a user-specified threshold, or the family of partitionings for v being empty.

Random forests A *random forest* F (Breiman 2001) is a collection of trees $T_1, \ldots, T_t$, where each tree is trained using the procedure described above, but each tree uses a *resampled* training set. This *bagging* of the original training set Z is done with the goal of increasing *diversity* and lowering *variance*. To further promote diversity among the forest, a random subset of the family of partitionings is searched at each node. Given a query point $x \in \mathcal{X}$, the leaf information $\mathsf{L}(x; T_j)$ for each tree $T_j \in F$ is used to compute an *ensemble response* to queries. In the running example of regression trees, the query response is a simple average of tree predictions, namely

$$\mathsf{q}(x; F) = \frac{1}{t} \sum_{T \in F} \mathsf{L}(x; T) \ .$$

3 Related work

Rosenblatt (1969) first describes CDE with *kernel CDE*, which applies *Kernel Density Estimation* (KDE) to the CDE problem, by reporting $\hat{\rho}(y|x) = \frac{\hat{\rho}(y,x)}{\hat{\rho}(x)}$, for each $\hat{\rho}(\cdot)$ estimate on the RHS made with KDE. Kernel CDE and many other nonparametric estimators require that the joint density is absolutely continuous to ensure that densities exist and that densities over $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$ exist. *Generalized Linear Models* (GLM) (Nelder and Wedderburn 1972) are CDE methods that essentially generalize linear regression beyond the fixed-variance Gassian case. They do not require absolute continuity, although $\mathcal{Y}$ must be continuous. Low-dimensional GLM are generally interpretable but inflexible, while generalizations like import vector machines (Zhu and Hastie 2002) are flexible but uninterpretable.

By their inherently probabilistic nature, graphical models are well-suited for CDE. *Cutset networks* (Rahman et al. 2014; Di Mauro et al. 2017) are OR trees with tractable probabilistic models at their leaves, and *mixed-sum product networks* (Molina et al. 2018) are graphical models with tree structure for mixed data. Each bears some resemblance to decision trees, and admits more efficient induction and inference than general graphical models. However, they must be large enough to represent conditional density relationships *between all variables*, since they make no distinction between *features* and *labels*. Answering CDE queries on these models requires, despite their tree structure, a complicated *global process* of marginalization, conditioning, and related operations, often spanning *the entire network*. This procedure is less efficient and more recondite to the human analyst than standard decision tree queries, which occur *locally* along a single root-to-leaf path.

Decision trees are lauded for their *simplicity*, *efficiency*, and *interpretability*, but current tree-based CDE techniques lack these properties. Chaudhuri et al. (2002) propose the first tree-based *Conditional Quantile Estimation* (CQE) technique, and Meinshausen (2006) introduces the first tree-based CQE approach, *Quantile Regression Forest* (QRFs). QRFs minimize standard regression impurity criteria to select split rules, which essentially only consider the *means* of the target variable y in the subsets resulting from the split, rather than taking into account the entire *sample distribution* of the target variable. These impurity criteria are ill-suited for CDE, as they do not incentivize splits that improve CDE estimates (discussed further in Sect. 4.2). Pospisil and Lee (2018) introduce *Random Forests for Conditional Density Estimation* (RFCDE), which are largely equivalent to CQE, except they use estimated MISE impurity (whose issues were discussed in Sect. 2), and output KDE (effectively kernel-

smoothed quantile estimates). RFCDE and QRF queries operate on the training labels mapped to each leaf, which must be stored and processed explicitly, incurring high storage and query costs.

Hothorn and Zeileis (2017) propose the *transformation forest* (TF), which chooses split rules using *null-hypothesis testing*. It is not clear that conservatively chosen splits benefit forests, as ensemble methods thrive on *diverse weak learners*. TFs fit distributions using *transformation families*: given a fixed univariate PDF ψ they pick an invertible *transformation function* $\phi : \mathbb{R} \to \mathbb{R}$, producing the density estimate $(\psi \circ \phi)(y) = |\frac{\mathrm{d}}{\mathrm{d}y}\phi(y)|^{-1}\psi\big(\phi(y)\big)$. The learned ϕ can be complicated, yielding uninterpretable models even for simple ψ, and TFs must also store and process raw labels to answer forest queries (see also Sect. 5).

CADET overcomes these limitations with a *parametric approach*, learning *interpretable trees* that make *parametric density estimates*. It uses the *empirical cross-entropy* impurity criterion, which incentivizes effective splits for CDE. CADET attains low storage and query costs by storing *sufficient statistics* of the training labels associated with each leaf, requiring bounded memory and computation. CADET estimates parametric densities within a user-selected family, which are generally more interpretable, and learning them requires fewer samples than nonparametric estimates. Finally, as CADET makes no assumptions on the underlying probability space, it can be instantiated directly on arbitrary probability spaces (including multivariate, mixed, and other exotic cases).

4 CADET: interpretable parametric CDE with trees and forests

CADET is a specific instantiation of the decision tree and random forest models (Sect. 2). It makes heavy use of *sufficient statistics*, so we first discuss this concept.

4.1 Sufficient statistics

Let $\mathcal{F}$ be a *parametric* family of PDFs over $\mathcal{Y}$, with parameter space Θ, and take $\theta \in \Theta$. The member of $\mathcal{F}$ identified by θ is denoted as $\rho(\cdot; \mathcal{F}, \theta)$. We omit $\mathcal{F}$ from this and other notation when clear from context.

Let $Y \in \mathcal{Y}^n$ for some sample size n, sampled i.i.d. from the distribution arising from some unknown $\rho(\cdot; \theta) \in \mathcal{F}$. A *sufficient statistic* for Θ (alternatively referred to as a sufficient statistic for $\mathcal{F}$) is a vector-valued function $\mathsf{w}^{(n)} : \mathcal{Y}^n \to \mathbb{R}^{\dim(\mathsf{w})}$ (where $\dim(\mathsf{w})$ is the codomain dimension of $\mathsf{w}(\cdot)$) such that $\mathsf{w}^{(n)}(Y)$ is *as informative as* Y for the purpose of estimating the unknown θ that determines the unknown PDF $\rho(\cdot; \theta)$ (Casella and Berger 2002, Sect. 6.2). For example,

$$\mathsf{w}^{(n)}(Y) = \left(\sum_{y \in Y} y, \sum_{y \in Y} y^2 \right)$$

is a sufficient statistic for the Gaussian family, with MLE mean and variance $\hat{\mu} = \frac{\mathsf{w}_1^{(n)}(Y)}{n}$ and $\hat{\sigma}^2 = \frac{\mathsf{w}_2^{(n)}(Y)}{n} - \hat{\mu}^2$. A sufficient statistic for the Pareto family is

$$\mathsf{w}^{(n)}(Y) = \left(\min(Y), \prod_{y \in Y} y \right). \tag{4}$$

The Fisher-Neyman factorization theorem (Halmos et al. 1949) shows that for any PDF $\rho(\cdot;\theta) : \mathcal{Y} \to \mathbb{R}_{0+}$ from a family $\mathcal{F}$ with sufficient statistic $\mathsf{w}^{(1)}(\cdot) : \mathcal{Y} \to \mathbb{R}^{\dim(\mathsf{w})}$, there exists a *base measure* $\mathsf{h}(\cdot) : \mathcal{Y} \to \mathbb{R}_{0+}$ and a *factorization function* $\mathrm{F}(\cdot;\cdot) : \mathbb{R}^{\dim(\mathsf{w})} \times \Theta \to \mathbb{R}_{0+}$ such that

$$\rho(\cdot;\theta) = \mathsf{h}(\cdot)\mathrm{F}\big(\mathsf{w}^{(1)}(\cdot);\theta\big) \ . \tag{5}$$

We now define $\mathsf{p}^{(n)}(\cdot) : \mathbb{R}^{\dim(\mathsf{w})} \to \Theta$ to be the function that selects $\theta \in \Theta$ to maximize the likelihood of an i.i.d. sample $Y \in \mathcal{Y}^n$ given $\mathsf{w}^{(n)}(Y)$:

$$\mathsf{p}^{(n)}\Big(\mathsf{w}^{(n)}(Y)\Big) = \arg\max_{\theta\in\Theta} \prod_{y\in Y} \rho(y;\theta) = \arg\max_{\theta\in\Theta} \sum_{y\in Y} \ln\big(\mathrm{F}(\mathsf{w}^{(n)}(Y);\theta)\big), \tag{6}$$

where the rightmost equality follows from (5). We omit the sample-size superscript from both $\mathsf{w}(\cdot)$ and $\mathsf{p}(\cdot)$ when clear from context, and further abuse notation when discussing trees, letting $\mathsf{w}^{(n)}(Z)$ ignore the $\mathcal{X}$ elements of a sample $Z \in (\mathcal{X}\times\mathcal{Y})^n$.

Exponential class A *natural sufficient statistic* is a sufficient statistic for $\mathcal{F}$, such that for i.i.d. samples $Y \in \mathcal{Y}^n$, $Y' \in \mathcal{Y}^{n'}$, and their *concatenation* $Y \uplus Y'$, it holds (Casella and Berger 2002, Thm. 6.2.10) that

$$\mathsf{w}^{(n+n')}(Y \uplus Y') = \mathsf{w}^{(n)}(Y) + \mathsf{w}^{(n')}(Y') \ . \tag{7}$$

A distribution family $\mathcal{F}$ with parameter space Θ and support $\mathcal{Y}$ is said to be in the *exponential class* if it admits a factorization into a *natural sufficient statistic* $\mathsf{w}^{(1)}(\cdot) : \mathcal{Y} \to \dim(\mathsf{w})$, *base measure* $\mathsf{h}(\cdot) : \mathcal{Y} \to \mathbb{R}_+$, *parameter function* $\eta(\cdot) : \Theta \to \mathbb{R}^{\dim(\mathsf{w})}$, and *log-partition function* $\mathsf{A}(\cdot) : \Theta \to \mathbb{R}$, such that any PDF $\rho(\cdot;\theta) \in \mathcal{F}$ can be written as

$$\rho(\cdot;\theta) = \mathsf{h}(\cdot)\exp\big(\eta(\theta)\cdot\mathsf{w}^{(1)}(\cdot) - \mathsf{A}(\theta)\big), \tag{8}$$

The exponential class contains many well-known (thus interpretable to a human analyst) distribution families, including the Gaussian, exponential, gamma, beta, Dirichlet, geometric, and Poisson families. *Sufficient statistics* and *combination functions* like (7) are key to the performance guarantees of CADET, so naturally one might wonder under which conditions they exist. The Pitman-Koopman-Darmois theorem (Koopman 1936) shows that if a family $\mathcal{F}$ has fixed support and a bounded-dimensional sufficient statistic, then $\mathcal{F}$ is in the exponential class.

Among variable-support families with a bounded-dimensional sufficient statistic $\mathsf{w}(\cdot)$, some admit a *combination function* $\mathsf{g}(\cdot,\cdot)$ such that

$$\mathsf{w}^{(n+n')}(Y \uplus Y') = \mathsf{g}\big(\mathsf{w}^{(n)}(Y), \mathsf{w}^{(n')}(Y')\big), \tag{9}$$

which generalizes (7) beyond the exponential class. The Pareto and uniform interval families admit such $\mathsf{g}(\cdot,\cdot)$; the reader is invited to derive one for the Pareto family, starting from the sufficient statistic in (4). CADET estimates conditional densities by storing sufficient statistics at each leaf of the decision tree, which through (6), are isomorphic to MLE distribution estimates.

4.2 Decision trees for interpretable parametric CDE

CADET is an instantiation of the decision tree model described in Sect. 2. It is parameterized by a parametric family $\mathcal{F}$, which determines the class of densities that a CADET tree or forest can predict. Bounded-dimensional sufficient statistics and combination functions are needed to *efficiently* train CADET trees and to aggregate tree information into forest queries,

so here we assume these exist for $\mathcal{F}$. Their nonexistence does not impact the theory behind CADET, thus with minor changes, CADET may be instantiated for parametric families lacking bounded-dimensional sufficient statistics or combination functions, although in this case, training time, forest memory, and forest query time costs may be higher.

Impurity criterion Let $\mathcal{F}$ be a parametric family of PDFs, with bounded-dimensional sufficient statistic w and parameter space Θ. CADET minimizes the *Empirical Cross Entropy* (ECE) impurity, defined as

$$\mathsf{m}_{\mathsf{ece}}(Y;\mathcal{F}) = -\frac{1}{|Y|}\sum_{y\in Y}\ln\big(\rho(y;\mathsf{p}(\mathsf{w}(Y)))\big)\ . \tag{10}$$

The ECE impurity is *parametric* in the sense that it depends on the hyperparameter $\mathcal{F}$ (omitted when clear from context). This dependence is key, as it allows $\mathsf{m}_{\mathsf{ece}}$ to incentivize splits that lead to the data being well-fit *by* $\mathcal{F}$. The ECE impurity should be contrasted with the MSE loss from (1), which Hothorn and Zeileis (2017) argue is ineffective for CDE, as it is not sensitive to changes over $\mathcal{X}$ of the *conditional distribution* of $\mathcal{Y}$, but only to changes of the *conditional expectation* of $\mathcal{Y}$.

The ECE is the impurity-criterion counterpart of the *cross entropy loss*, often used in neural networks (Goodfellow et al. 2016, Ch. 5.5) and binomial regression models (Weisberg 2005, Ch. 12). Cross entropy loss is theoretically motivated, both from decision-theoretic and coding-theoretic perspectives. In decision theory, a *strictly proper scoring rule* is a loss function that is uniquely minimized by predicting the true density. The cross entropy (often called the *logarithmic scoring rule*), is the only such rule (up to affine transformation) that is also *local*, meaning that given label y and estimated distribution $\hat{\rho}(\cdot)$, it may be computed as a function of $\hat{\rho}(y)$ (Shuford et al. 1966). From a coding theory perspective, cross entropy is a measure of the degree of inefficiency of using one distribution to encode symbols from another. The source coding theorem (Shannon 1948) shows that maximal efficiency is attained when the encoding distribution matches the true distribution.

The entropy of a PDF ψ with support $\mathcal{Y}$ is[3]

$$\mathsf{H}(\psi) = -\int_{\mathcal{Y}} \psi(y)\ln\big(\psi(y)\big)\,\mathrm{d}y\ .$$

We now show that ECE impurity and the entropy of the MLE distribution often coincide in the exponential class.

Lemma 1 *Suppose $Y \in \mathcal{Y}^n$, and $\mathcal{F}$ a member of the exponential class, with base measure $h(\cdot)$ and sufficient statistic $\mathsf{w}(\cdot)$. Let $\theta = \mathsf{p}^{(n)}\big(\mathsf{w}^{(n)}(Y)\big)$, $\hat{B} = \frac{1}{n}\sum_{y\in Y}\ln\big(h(y)\big)$, y' drawn with density $\rho(\cdot;\theta)$, and $B = \mathbb{E}_{y'}\big[\ln\big(h(y')\big)\big]$. Then*

1. *if $\ln\big(h(\cdot)\big)$ is an* affine function *of $\mathsf{w}^{(1)}(\cdot)$, then $\mathsf{m}_{\mathsf{ece}}(Y) = \mathsf{H}\big(\rho(\cdot;\theta)\big)$; and*
2. *in general, $\mathsf{m}_{\mathsf{ece}}(Y) = (B-\hat{B}) + \mathsf{H}\big(\rho(\cdot;\theta)\big)$.*

Proof We first show Case 2, from which Case 1 follows.

$$\mathsf{m}_{\mathsf{ece}}(Y) = -\frac{1}{n}\sum_{y\ \in\ Y}\ln\big(\rho(y;\theta)\big) \qquad \text{DEFINITION OF } \mathsf{M}_{\mathsf{ECE}}(\cdot)$$

[3] This definition of entropy encompasses differential entropy for integration w.r.t. the Lebesgue measure, discrete entropy for integration w.r.t. the counting measure, and other entropies with appropriate measures.

$$
\begin{aligned}
&= -\frac{1}{n}\sum_{y \in Y} \ln\big(\mathsf{h}(y)\big) + \eta(\theta)\cdot \mathsf{w}^{(1)}(y) - \mathsf{A}(\theta) && \text{Equation 8}\\
&= (B - \hat{B}) + \mathsf{A}(\theta) - C - \eta(\theta)\cdot \frac{1}{n}\sum_{y \in Y} \mathsf{w}^{(1)}(y) && \text{Algebra}\\
&= (B - \hat{B}) + \mathsf{A}(\theta) - C - \eta(\theta)\cdot \frac{1}{n}\mathsf{w}^{(n)}(Y) && \text{Equation 7}\\
&= (B - \hat{B}) + \Big(\mathsf{A}(\theta) - C - \eta(\theta)\cdot \mathbb{E}_{y'}[\mathsf{w}^{(1)}(y')]\Big) && \text{Maximum Likelihood}\\
&= (B - \hat{B}) + \mathbb{E}_{y'}\Big[\mathsf{A}(\theta) - C - \eta(\theta)\cdot \mathsf{w}^{(1)}(y')\Big] && \text{Linearity of Expectation}\\
&= (B - \hat{B}) - \mathbb{E}_{y'}\big[\ln\big(\rho(y';\theta)\big)\big] && \text{Equation 8}\\
&= (B - \hat{B}) + \mathsf{H}\big(\rho(\cdot;\theta)\big)\ . && \text{Definition of H}(\cdot)
\end{aligned}
$$

The Maximum Likelihood step holds since in MLE, sample sufficient statistics are always preserved in the fitted distribution (this property is evident in the *maximum entropy* interpretation of MLE, where it holds by definition).

The additional hypothesis in Case 1 implies the existence of $\beta \in \mathbb{R}, \alpha \in \mathbb{R}^{\dim(\mathsf{w})}$ such that $\ln\big(\mathsf{h}(\cdot)\big) = \beta + \alpha \cdot \mathsf{w}^{(1)}(\cdot)$. It then holds that

$$
\hat{B} = \frac{1}{n}\sum_{y \in Y} \ln\big(\mathsf{h}(y)\big) = \beta + \alpha \cdot \frac{1}{n}\mathsf{w}^{(n)}(Y) = \beta + \alpha \cdot \mathbb{E}_{y'}\big[\mathsf{w}^{(1)}(y')\big] = B,
$$

and via Case 2, noting that here $B - \hat{B} = 0$, we obtain Case 1. □

Case 1 of Lemma 1 applies to many families of interest, such as the Gaussian, gamma, and Von-Mises families, where $\mathsf{h}(\cdot)$ is constant, and the beta, Dirichlet, and log-Gaussian families, where $\ln\big(\mathsf{h}(\cdot)\big)$ is an affine function of $\mathsf{w}^{(1)}(\cdot)$. When Case 1 holds, the splits chosen by CaDET are the same that would be chosen by minimizing entropy, as done in *information gain* trees. These trees select splits that explain as much variation in $\mathcal{Y}$ as possible, leading to more homogeneous leaves to which more accurate distributions can be fit. When the ECE and entropy do not coincide, an argument can be made for using either as an impurity criterion, and CaDET can be adapted to instead select entropy-minimizing splits if so desired.

A more practical consequence of Lemma 1 is that the *impurity reduction* (see (3)) w.r.t. $\mathsf{m}_{\mathsf{ece}}(\cdot)$ of any split at any node with training labels Y can computed from $\mathsf{w}(Y)$ without having to iterate over Y or knowing $\hat{B}$. Furthermore, $\mathsf{m}_{\mathsf{ece}}(Y)$ can be *computed* from $\mathsf{H}\big(\rho(\cdot;\theta)\big)$ even in Case 2, if $\hat{B}$ (the sum of log base measures) is computed along with $\mathsf{w}(Y)$. Similar results can often be derived for $\mathcal{F}$ not in the exponential class; the reader is invited to confirm that for the uniform interval distribution (over $\mathbb{R}$ or $\mathbb{Z}$), it holds that $\mathsf{m}_{\mathsf{ece}}(Y) = \mathsf{H}\big(\rho(\cdot;\theta)\big)$.

Leaf information The information $\mathsf{L}(\ell; T)$ stored at the leaf ℓ of a CaDET tree T is the number of training points mapped to the leaf $|\mathsf{T}(\ell; T, Z)|$, and the *sufficient statistics* $\mathsf{w}\big(\mathsf{T}(\ell; T, Z)\big)$ of the training elements $\mathsf{T}(\ell; T, Z)$ that T maps to ℓ. For notational convenience, we take $\mathsf{L}(\cdot;\cdot)$ to be a vector, where the 0th component is the sample size, and the remaining components are the sufficient statistic, i.e.,

$$
\mathsf{L}_0(\ell; T) = \big|\mathsf{T}(\ell; T, Z)\big|, \text{ and } \mathsf{L}_{1:\dim(\mathsf{w})}(\ell; T) = \mathsf{w}\big(\mathsf{T}(\ell; T, Z)\big),
$$

where $V_{a:b}(\cdot)$ is *vector slice notation*, corresponding to codomain indices $a, \ldots b$ of the vector-valued function $V(\cdot)$. Because CaDET stores only $\mathsf{w}\big(\mathsf{T}(\ell; T, Z)\big)$ at each leaf ℓ, it has lower storage and query time costs than current tree-based CDE methods, which must store and process raw training labels to answer forest queries.

Response to queries Given a tree T, the response $\mathsf{q}(\cdot; x, T)$ to a query at $x \in \mathcal{X}$ is the MLE PDF w.r.t. $\mathcal{F}$ on the $\mathcal{Y}$ components of $\mathsf{T}(x; T, Z)$:

$$\mathsf{q}(\cdot; x, T) = \rho\Big(\cdot; \mathsf{p}^{(N)}\Big(\mathsf{w}^{(N)}\big(\mathsf{T}(x; T, Z)\big)\Big)\Big) = \rho\Big(\cdot; \mathsf{p}^{(N)}\big(\mathsf{L}_{1:\dim(\mathsf{w})}(x; T)\big)\Big),$$

taking $N = |\mathsf{T}(x; T, Z)| = \mathsf{L}_0(x; T)$. Since CDE responses are PDFs, which are themselves functions, we write $\mathsf{q}(\cdot; x, T)$, where the first argument is an element of the domain of the PDF, the second the query point, and the third the tree.

This response is well-motivated, as $\mathsf{T}(x; T, Z)$ should be an *approximately independent* sample from *approximately* the conditional distribution at x. The "approximate" qualification is needed as split choice induces some dependence, and the conditional distribution changes as $\mathcal{Y}$ varies throughout the leaf. Ignoring the approximation, it is then reasonable to return the MLE estimate for this sample.

The careful reader may notice that one could just store this PDF at the leaf, in place of the sufficient statistic of the training set mapped to this leaf. For trees, either suffices, but we will require sufficient statistics to answer queries with forests.

4.3 Random forests

Consider a random forest F composed of CaDET trees $T_1, \ldots, T_t$, with training sets $Z_1, \ldots, Z_t$, and shared distribution family $\mathcal{F}$. Here the response $\mathsf{q}(\cdot; x, F)$ to the query at $x \in \mathcal{X}$ is

$$\mathsf{q}(\cdot; x, F) = \rho\left(\cdot\,; \mathsf{p}^{(N)}\left(\mathsf{w}\left(\biguplus_{i=1}^{t} \mathsf{T}(x; T_i, Z_i)\right)\right)\right) = \rho\left(\cdot\,; \mathsf{p}^{(N)}\left(\sum_{T \in F} \mathsf{L}_{1:\dim(\mathsf{w})}(x; T)\right)\right),$$

where $$N = \sum_{i=1}^{t} \big|\mathsf{T}\big(x; T_i, Z_i\big)\big| = \sum_{T \in F} \mathsf{L}_0(x; T),$$

and for exponential-class $\mathcal{F}$, the sum is from (7), and must be replaced by repeated applications of $\mathsf{g}(\cdot, \cdot)$ from (9) for $\mathcal{F}$ not in the exponential class.

If each training set Z_i for each T_i were drawn i.i.d., then sample concatenation across the trees would be well-motivated, since for any $x \in \mathcal{X}$, by the same reasoning as in the tree case, each $\mathsf{T}(x; T_i, Z_i)$ is an approximately i.i.d. sample from the true conditional density at x, thus the MLE estimator for their sample concatenation should be better than any of the individual trees estimates. When instead each Z_i is created by bagging the original training data, the samples at each leaf are more dependent (duplicates are more likely), and MLE should behave similarly to a *parametric bootstrap estimate*, but the same reasoning of combining small approximately i.i.d. samples into one large sample and performing MLE holds.

4.4 Discussion

On domains and parametric families CaDET can be instantiated with any parametric family with a bounded-dimensional sufficient statistic over any $\mathcal{Y}$. In contrast, non-parametric techniques are generally tied to a particular codomain, often $\mathbb{R}^d$. Although many spaces (e.g., discrete, simplicial, spherical, or cyclic) can be embedded in $\mathbb{R}^d$, interpreting a nonparametric model over $\mathbb{R}^d$ in $\mathcal{Y} \subseteq \mathbb{R}^d$ may invalidate density estimates, as densities over $\mathbb{R}^d$ are not necessarily densities over $\mathcal{Y}$.

Specifically, if $\mathcal{Y} \subseteq \mathbb{R}^d$, but densities in $\mathcal{Y}$ are interpreted w.r.t. the Lebesgue or Borel measures in $\mathbb{R}^d$, then often the total mass over $\mathcal{Y}$ is less than 1. Furthermore, if $\mathbb{R}^d$ and $\mathcal{Y}$ do not share a measure (as in simplicial or spherical domains, where $\mathcal{Y}$ is $(d-1)$-dimensional), the total mass of estimated densities can even exceed[4] 1. Workarounds like *transformation functions* exist, though they have their own issues (see Sect. 5), whereas CADET can handle tasks directly in their original space, using simple probabilistic models designed to work well for a particular setting.

Parametric versus non-parametric sample complexity CADET's restriction to parametric families with bounded-dimensional sufficient statistics necessarily limits the representative power of its CDEs: if $\mathcal{F}$ poorly models true conditional densities, then nonparametric CDE trees may outperform CADET given enough training data. However, CADET will generally perform better with small sample sizes, as MLE exhibits faster convergence than nonparametric techniques.[5] We show an example of this behavior in Sect. 6.

This faster rate is particularly important in CDE-trees, since *each leaf* requires enough data to accurately estimate conditional densities. CADET trees thus require fewer samples at each leaf than nonparametric methods, allowing them to better model *conditional* density structure with additional splits. Even with additional splits, CADET generally remains more interpretable than nonparametric methods, as splits are easily understood, whereas complicated nonparametric distribution estimates are not.

Generalizing prior art Let $\mathcal{F}_\mathsf{c}$ be the *categorical family*, and $\mathcal{F}_\mathsf{G}$ the *unit-variance Gaussian family*. It holds by Lemma 1 that $\mathsf{m}_{\mathsf{ece}}(\cdot; \mathcal{F}_\mathsf{c}) = \mathsf{m}_\mathsf{H}(\cdot)$, and $\mathsf{m}_{\mathsf{ece}}(\cdot; \mathcal{F}_\mathsf{G}) \propto \mathsf{m}_{\mathsf{mse}}(\cdot)$. Thus, with these family choices, CADET makes the same splits as entropy-minimizing classification trees (Quinlan 1986) and MSE-minimizing regression trees (Breiman et al. 1984), respectively. CADET therefore generalizes two classic decision-tree models to a broad class of parametric estimation problems.

4.5 Training time complexity

Consider the training of a decision tree using a training set $Z \in (\mathcal{X} \times \mathcal{Y})^n$, where splits are chosen from all univariate threshold functions over a constant number of features to minimize either $\mathsf{m}_\mathsf{H}(\cdot)$ (for classification) or $\mathsf{m}_{\mathsf{mse}}(\cdot)$ (for regression). The time necessary for the training is in the best case $\boldsymbol{\Theta}\big(n \log n\big)$, and in the worst case $\boldsymbol{\Theta}\big(n^2\big)$. TF (Hothorn and Zeileis 2017) and RFCDE (Pospisil and Lee 2018) require $\boldsymbol{\Omega}\big(|\mathsf{T}(v; T, Z)|\big)$ time to evaluate each potential split of node v, thus training them takes time $\boldsymbol{\Omega}\big(n^2\big)$ in the best case and $\boldsymbol{\Omega}\big(n^3\big)$ in the worst case. These times are worse than the ones mentioned above by a factor $\tilde{\boldsymbol{\Omega}}(n)$.

Training CADET trees with a family $\mathcal{F}$ attains the faster training time complexities of $\mathsf{m}_\mathsf{H}(\cdot)$ and $\mathsf{m}_{\mathsf{mse}}(\cdot)$ trees, as long as $\mathcal{F}$ has sufficient statistic $\mathsf{w}(\cdot)$ and combination function $\mathsf{g}(\cdot, \cdot)$ (see (9)), such that $\mathsf{g}(\cdot, \cdot)$, $\mathsf{w}^1(\cdot)$, and $\mathsf{H}\big(\rho(\cdot; \mathsf{p}(w))\big)$ for any $w \in \mathbb{R}^{\dim(\mathsf{w})}$ can all be evaluated in $\boldsymbol{\Theta}(1)$ time. CADET attains these time complexities because sufficient statistics can be updated via $\mathsf{g}(\cdot, \cdot)$ in amortized time at each potential split, and entropies can be efficiently computed (by assumption), matching the cost of computing discrete entropy or variance in classification or regression trees. Without bounded-dimensional sufficient statistics or combination functions, CADET generally must perform $\boldsymbol{\Omega}\big(|\mathsf{T}(v; T, Z)|\big)$ work to evaluate

[4] For example, if $\mathcal{Y}$ is the unit circle $\mathcal{S}_2$, the uniform distribution on $[-1, 1]^2$ with pdf ρ has mass $\oint_{\mathcal{S}_2} \rho(z)\, \mathrm{d}z = \frac{\pi}{2} > 1$ when integrated over the unit circle.

[5] Concretely, the MISE decays as $\boldsymbol{\omega}_p(n^{-1})$ for the best-known KDE bounds (Agarwal et al. 2017), but $\mathbf{O}_p(n^{-1})$ for parametric MLE (Kanazawa 1993).

a split at node v, exactly as in RFCDE and TF. In this case, CADET would then attain the slower training time complexities of these algorithms.

5 Extensions

Parametric distributions with few parameters, such as univariate Gaussians, are generally interpretable. However, the distribution families one might naturally consider in high-dimensional or unfamiliar spaces may have many parameters, thus becoming less interpretable. We now discuss three methods to construct rich parametric families over complex domains from simple constituent families over familiar domains, *without sacrificing interpretability*:

- *product families*, which are multivariate distributions built from univariate constituent distributions;
- *transformation families*, which can be used to produce distributions with restricted support to suit domain-specific requirements; and
- *union families*, which enable performing MLE over *multiple families*.

Product families We often want to estimate multivariate densities, i.e., $\mathcal{Y}$ is a *product space* with $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_d$, but have domain-specific knowledge about each $\mathcal{Y}_i$ (e.g., whether the support is discrete, real, or semireal) which standard primitive distributions, such as the multivariate Gaussian, would ignore. *Product families* compute the *joint density* over $\mathcal{Y}$ as a *product of density estimates* (thus treating each multiplicand as an independent random variable) over each $\mathcal{Y}_1, \ldots, \mathcal{Y}_d$. Computation over product families is particularly convenient, as sufficient statistics, densities, and entropies can all be computed from univariate densities, and the exponential class is closed under finite products.

Product-family CADET should be contrasted with CADET applied separately on each $\mathcal{Y}_i$, and estimating joint densities as products of univariate densities. In both cases, joint CDE are product distributions, however in the first case, CADET uses impurity reduction *across all* $\mathcal{Y}_1, \ldots, \mathcal{Y}_d$ to select splits, whereas in the second case, splits are separately learned for each $\mathcal{Y}_i$. If the conditional densities of each $\mathcal{Y}_i$ vary similarly over $\mathcal{X}$, then this additional information allows better split selection in the first case. Additionally, the product-family tree is simpler than the collection of trees for each $\mathcal{Y}_i$, thus more interpretable and less prone to overfitting.

Transformation families Often $\mathcal{Y}$ is not $\mathbb{R}^d$ or some space with a plethora of convenient well-known distribution families. For example, $\mathcal{Y}$ could be the unit sphere, unit simplex, or some compact subset of $\mathbb{R}^d$. Transformation families contain distributions over such a $\mathcal{Y}$, obtained by *transforming* familiar distributions over some isomorphic space $\mathcal{Y}'$. Such transformations can be intuitive and thus interpretable; for instance we may transform Cartesian coordinates of points on the Earth's surface to the more familiar latitude and longitude coordinates.

A *transformation function* $\phi : \mathcal{Y} \to \phi(\mathcal{Y})$ is a *differentiable invertible mapping*. Given a family $\mathcal{F}$ over $\mathcal{Y}$ parameterized by Θ, we define the *ϕ-transformed density*

$$(\rho \circ \phi)(\cdot; \theta) = \left|\mathcal{J}\big(\phi(\cdot)\big)\right|^{-1} \rho\big(\phi(\cdot); \theta\big), \tag{11}$$

and the corresponding *ϕ-transformed family*

$$\mathcal{F} \circ \phi = \{(\rho \circ \phi)(\cdot; \theta) : \theta \in \Theta\}, \tag{12}$$

where $|\mathcal{J}(\phi(\cdot))|$ is the absolute determinant of the Jacobian of ϕ. In CADET we assume the existence of a bounded-dimensional sufficient statistic, which is particularly convenient

with transformation functions through (5), as we may compute the *base measure* of $\mathcal{F} \circ \phi$ as $\mathsf{h}(\cdot; \mathcal{F} \circ \phi) = |\mathcal{J}(\phi(\cdot))|^{-1}\mathsf{h}(\phi(\cdot))$ and the *sufficient statistic* as $\mathsf{w}(\cdot; \mathcal{F} \circ \phi) = \mathsf{w}(\phi(\cdot); \mathcal{F})$.

For example, we can construct the *inverse-gamma* family from $\phi(y) = y^{-1}$ and the gamma family, and the log-normal family from $\phi(y) : \mathbb{R}_+^d \to \mathbb{R}^d = \ln(y)$ and the Gaussian family. The logarithm elicits a *domain-change*, yielding families over $\mathbb{R}_+$, which is useful for estimating positive quantities.

Transformation functions, when paired with an appropriate coordinate system, can also be used to construct distributions over sets of Lebesgue measure zero, such as the unit simplex $\Delta^d = \{y \in (0,1)^{d+1} : \|y\|_1 = 1\}$, or the unit sphere $\mathcal{S}^d = \{y \in \mathbb{R}^{d+1} : \|y\|_2 = 1\}$, which are of key importance in compositional statistics and directional statistics, respectively. E.g., for simplicial data we can apply the *Additive Log-Ratio-Transform* (ALRT) (Aitchison 1982)

$$\phi_{\mathsf{ALRT}}(y_1, \ldots, y_{d+1}) : \Delta^d \to \mathbb{R}^d = \left(\ln\left(\frac{y_1}{y_{d+1}}\right), \ln\left(\frac{y_2}{y_{d+1}}\right), \ldots, \ln\left(\frac{y_d}{y_{d+1}}\right)\right), \quad (13)$$

and for spherical data, the *stereographic projection transform*

$$\phi_{\mathsf{Stg}}(y_1, \ldots, y_{d+1}) : \mathcal{S}^d \to \mathbb{R}^d = \left(\frac{y_1}{1 - y_{d+1}}, \frac{y_2}{1 - y_{d+1}}, \ldots, \frac{y_d}{1 - y_{d+1}}\right) . \quad (14)$$

In regression under the assumption of heteroskedastic noise, where the task is to predict $\mathbb{E}[y|x]$, data transformation is unsatisfying: for some transformation ϕ, learning $\mathbb{E}[\phi(y)|x]$ is insufficient, as it does not in general hold that $\mathbb{E}[y|x] = \phi^{-1}(\mathbb{E}[\phi(y)|x])$. In contrast, in CDE, we can convert conditional densities over $\phi(\mathcal{Y})$ to conditional densities over $\mathcal{Y}$ through (11), so we retain the ability to interpret transformed variables in the untransformed space.

Hothorn and Zeileis (2017) also use transformation functions in their Transformation Forests (TFs), though they fix *distributions* and parameterize *transformations*, while CADET does the opposite. For simple cases like affine transformations in location-scale families, they are equivalent, but we argue that simple distributions with complicated parameterized transformations are generally less interpretable than complicated parametric distributions with simple fixed transformations. TFs also only handle $\mathcal{Y} = \mathbb{R}$, and operate on quantiles rather than densities. Generalizing TFs to $\mathbb{R}^d$ is nontrivial, as working with multivariate quantiles or CDFs of transformations generally requires sophisticated integration, complicating interpretability and computation.

Transformation is thus intuitive, interpretable, and computationally convenient for parametric CDE. These beneficial properties put this use of transformation in stark contrast to its use in regression and quantile estimation, where it is in general difficult to interpret the output of transformed models in the original space.

Union families It is often hard to select *a priori* a parametric family to model conditional densities. One could select between models trained over multiple families, but to do so would be inefficient, and would perform poorly when the best family to fit conditional densities varies over $\mathcal{Y}$. It would be preferable to learn a model that is able to select distribution families in a *data-dependent manner*, fitting different distribution families to different regions of $\mathcal{X}$.

One could select the MLE at each leaf among multiple families, but this approach favors complexity over simplicity, and tends to overfit. Given families $\mathcal{F}_1, \mathcal{F}_2$ such that $\mathcal{F}_1 \subseteq \mathcal{F}_2$ (e.g., the exponential and gamma families), for any i.i.d. sample $Y \in \mathcal{Y}^n$, with MLE parameter estimates $\theta_1 = \mathsf{p}^{(n)}\big(\mathsf{w}^{(n)}(Y; \mathcal{F}_1); \mathcal{F}_1\big)$ and $\theta_2 = \mathsf{p}^{(n)}\big(\mathsf{w}^{(n)}(Y; \mathcal{F}_2); \mathcal{F}_2\big)$, the MLE sample densities obey

$$\rho(Y; \mathcal{F}_1, \theta_1) = \prod_{y \in Y} \rho(y; \mathcal{F}_1, \theta_1) \leq \prod_{y \in Y} \rho(y; \mathcal{F}_2, \theta_2) = \rho(Y; \mathcal{F}_2, \theta_2) \ .$$

However, the estimate $\rho(\cdot; \mathcal{F}_1, \theta_1)$ is often preferable to $\rho(\cdot; \mathcal{F}_2, \theta_2)$, for instance when they fit similarly well or n is small, as simpler distributions are more interpretable and generally less susceptible to overfitting.

We address these issues with a more nuanced approach, termed *regularized union family selection*. Given families $\mathcal{F}_1, \ldots, \mathcal{F}_m$, with parameter spaces $\Theta_1, \ldots, \Theta_m$, the *union family* $\mathcal{F} = \cup_{i=1}^m \mathcal{F}_i$ has parameter space $\cup_{i=1}^m (\{i\} \times \Theta_i)$, with

$$\rho\big(\cdot; \mathcal{F}, (i, \theta)\big) = \rho(\cdot; \mathcal{F}_i, \theta),$$

thus $\mathcal{F}$ can be used to select among distributions from several families. The sufficient statistics for each $\mathcal{F}_i$ are enough to perform MLE within each subfamily, and for exponential-class families, we may perform MLE over the entire union family given these sufficient statistics and the sample log base measures $\ln\big(\mathsf{h}(\cdot; \mathcal{F}_i)\big)$ associated with each $\mathcal{F}_i$ (see Lemma 1). However, to control for overfitting, prioritize simpler distributions, and incorporate *a priori* domain knowledge, we take *regularization hyperparameters* $\lambda = \langle \lambda_1, \ldots, \lambda_m \rangle$, and select the distribution that maximizes *regularized sample log likelihood*, defining $\mathsf{p}^{(n)}(\cdot; \mathcal{F}, \lambda)$ as

$$\mathsf{p}^{(n)}\big(\mathsf{w}^{(n)}(Y); \mathcal{F}, \lambda\big) = \operatorname{argmin}_{i \in \{1,\ldots,m\}, \theta \in \Theta_i} \lambda_i + \frac{1}{n} \sum_{y \in Y} \ln\big(\rho(y; \mathcal{F}_i, \theta)\big),$$

with corresponding *regularized empirical cross entropy* impurity criterion

$$\mathsf{m}_{\mathsf{ece}}(Y; \mathcal{F}, \lambda) = \min_{i \in \{1,\ldots,m\}} \lambda_i + \mathsf{m}_{\mathsf{ece}}(Y; \mathcal{F}_i), \tag{15}$$

where the notation explicitly references the family and regularization parameters.

There are many reasonable ways to select regularization parameters. For example, in our experiments we use the Akaike Information Criterion (AIC), setting

$$\lambda_i = \frac{\dim(\Theta_i)}{n}, \text{ for } i \in \{1, \ldots, m\}, \tag{16}$$

where $\dim(\Theta_i)$ denotes the dimension of the parameter space of $\mathcal{F}_i$.

6 Experimental evaluation

Here we present the results of our experimental evaluation of CADET, including the comparison with RFCDE (Pospisil and Lee 2018). We test the versatility of CADET by instantiating it with many parametric families, including over multivariate codomains, probability simplices, and a cyclic codomain. We also evaluate CADET with a union family and regularized ECE impurity (see Sect. 5). The accuracy of a learned model M is measured with the *Average Conditional Log Likelihood* (ACLL) of the conditional density estimations produced by M on a test set $Z' \in (\mathcal{Y} \times \mathcal{X})^{n'}$, i.e.,

$$\frac{1}{n'} \sum_{(x,y) \in Z'} \ln\big(\mathsf{q}(y; x, M)\big).$$

The ACLL is a good accuracy measure, as it can be computed from the *estimated conditional density* $\mathsf{q}(\cdot; x, M)$, and is maximized in expectation by the *true conditional density*.

Implementation Our implementation[6] of CADET extends scikit-learn (Pedregosa et al. 2011). It supports many distribution families, shown in Table 1. When building trees, it selects the

[6] Source code is provided at http://www.cs.brown.edu/people/ccousins/cadet/ under the new BSD license.

Table 1 Supported distribution families in our software package, with domain, sufficient statistic codomain dimension $\dim(\mathsf{w})$, whether the family is in the exponential class (EC), and if so, whether the base measure $\mathsf{h}(\cdot)$ is a *constant* 1, a *function* of the sufficient statistic $\mathsf{f} \circ \mathsf{w}$, or *neither* h

Domain type	Distribution family	Domain	$\dim(\mathsf{w})$	EC	$\mathsf{h}(\cdot)$
Real	Exponential	$\mathbb{R}0+$	1	Yes	1
	Gamma	"	2	Yes	1
	Inverse Gamma	$\mathbb{R}+$	2	Yes	1
	Inverse Gaussian	"	2	Yes	h
	Pareto	"	2	No	—
	Unit Scale Pareto	$(1, \infty)$	1	Yes	1
	Uniform	$\mathbb{R}$	2	No	—
Directional	Von Mises	$[0, 2\pi)$	2	Yes	1
	Von Mises-Fisher	$\mathcal{S}^d$ (sphere)	d	Yes	1
Simplicial	Beta	$(0, 1)$	2	Yes	$\mathsf{f} \circ \mathsf{w}$
	Dirichlet	Δ^d (simplex)	d	Yes	$\mathsf{f} \circ \mathsf{w}$
Multivariate Real	Gaussian	$\mathbb{R}^d$	$2d + \binom{d}{2}$	Yes	1
	Gaussian uncorrelated	"	$2d$	Yes	1
	Gaussian symmetric	"	$d + 1$	Yes	1
	Log-Gaussian	$\mathbb{R}_+^d$	$2d + \binom{d}{2}$	Yes	$\mathsf{f} \circ \mathsf{w}$
	Log-Gaussian uncorrelated	"	$2d$	Yes	$\mathsf{f} \circ \mathsf{w}$
	Log-Gaussian symmetric	"	$d + 1$	Yes	$\mathsf{f} \circ \mathsf{w}$
Integral	Geometric	$\mathbb{N}_0$	1	Yes	1
	Poisson	"	1	Yes	h
	Log Series	"	1	Yes	h
	Uniform	$\mathbb{Z}$	2	No	—
Nominal	Bernoulli	$\{0, 1\}$	1	Yes	1
	Categorical	$\{1, 2, \ldots, d\}$	$d - 1$	Yes	1

Gaussian uncorrelated refers to Gaussians with 0 nondiagonal covariance, *Gaussian symmetric* refers to Gaussians with scaled identity covariance matrices, and each log Gaussian variant refers to a logarithmically-transformed Gaussian family

split that minimizes impurity over all univariate threshold functions such that at least some user-specified number of training points are assigned to each child. We call this parameter the *Minimum Samples per Leaf* (MSL). Forests do not search all univariate threshold functions in all features, but instead consider only univariate thresholds on $\lfloor \sqrt{\dim(\mathcal{Y})} + 1/2 \rfloor$ features, drawn uniformly without replacement at each node.

Baseline We compare the accuracy and interpretability of various CADET models to RFCDE models (Pospisil and Lee 2018) on many multivariate CDE tasks. RFCDE was experimentally shown to be superior to other tree-based techniques such as QRFs (Meinshausen 2006) and TFs (Hothorn and Zeileis 2017), both of which only operate over univariate $\mathcal{Y}$. We use the RFCDE implementation provided by the authors, with Gaussian KDE, the *normal reference* dynamic kernel-width selection strategy, and a 7-term tensor-cosine basis. This implementation does not allow log-density queries, and thus can output conditional density 0 (due to limited floating-point precision), to which we assign log-density -1000. This choice does not artificially disadvantage RFCDE, as our CADET implementation permits floating-point log-density queries, which can attain values far below -1000.

Table 2 Comparison between Gaussian-CADET with MSE impurity, Gaussian-CADET, and AIC-regularized union-CADET over UCI datasets

Dataset	n	$\dim(\mathcal{X})$	$\dim(\mathcal{Y})$	Gaussian MSE		Gaussian		Union		RFCDE	
				ACLL	Size	ACLL	Size	ACLL	Size	ACLL	Size
air-quality	6287	11	4	−13.159	2450	−12.493	2338	**−12.241**	**1990**	−15.078	25144
anneal-U	763	26	6	−24.573	513	−21.718	459	**−20.342**	**156**	−29.114	4578
australian	586	12	3	−8.703	135	−8.640	135	**−7.630**	**66**	−9.516	1758
auto	174	21	5	−28.449	100	−28.490	100	**−28.163**	**50**	−29.073	870
balance-scale	531	1	4	**−6.168**	42	**−6.168**	42	**−6.168**	**36**	−6.704	2124
breast	594	6	4	−6.001	210	−5.622	238	**−5.408**	**120**	−7.070	2376
breast-cancer	243	7	3	−9.646	63	−9.646	63	**−9.176**	**30**	−9.593	729
cars	333	7	2	−5.278	35	−5.277	35	**−5.250**	**29**	−5.700	666
cleve	258	9	5	−19.174	100	−19.040	100	**−18.682**	**50**	−19.518	1285
crx	586	11	5	−24.655	260	−24.627	300	**−21.487**	**130**	−24.232	2930
diabetes	653	2	7	−25.882	595	−25.802	595	−25.905	**343**	**−25.766**	4564
german	850	17	4	−17.621	350	−18.434	294	−17.856	**176**	**−17.561**	3400
german-org	850	22	3	−11.870	189	−12.167	207	**−11.821**	**114**	−11.832	2550
heart	230	8	6	−21.626	135	−21.764	135	**−21.304**	**60**	−21.603	1374
hypothyroid	2689	22	4	−13.569	994	−12.765	1078	**−12.759**	**672**	−13.190	10752
iris	128	1	4	**−3.025**	**14**	**−3.025**	**14**	**−3.025**	**14**	−3.650	508
winequality	5222	5	8	−6.821	6820	**−6.691**	6468	−6.867	**4368**	−8.340	44176
Mean				−14.484	765	−14.257	741	**−13.766**	**494**	−15.149	6458
# Optimal				2	1	3	1	**14**	**17**	2	0

For each dataset, we provide training set size n, feature count $\dim(\mathcal{X})$, and label count $\dim(\mathcal{Y})$. Test set ACLL and model size are presented for each. Trees are trained with MSL 55, and data are divided 85:15 train:test. Means for ACLL and model size of each algorithm, and the number of times each algorithm is optimal (w.r.t. ACLL or model size) follow tabulated results

Bold values indicate the best performing method

Datasets, tasks, and families We used datasets with different associated prediction tasks, requiring different choices of the parametric family $\mathcal{F}$:

- The AIR QUALITY dataset (De Vito et al. 2008) tasks us with estimating (multivariate) conditional *probability densities* of the *concentrations* (particle count or unit mass per unit volume) of four pollutants, given time, temperature, humidity, and air quality sensor readings. We randomly split the dataset into training and test sets of 3,698 samples each. Concentrations must be non-negative, so we use CADET with the unconstrained, uncorrelated, and symmetric log-Gaussian families. We compare these models to logarithmically-transformed RFCDE.
- The BATTING dataset (Lahman 2018) is our largest dataset, with 88,461 samples, each representing a professional baseball player, with height, weight, age, handedness, birthplace, league, and team features. BATTING tasks us with estimating the probabilities of a player attaining each of five batting outcomes (base 1–3, home, or strikeout), thus outcome distributions are members of $\mathcal{Y} = \Delta^4 = \{y \in (0,1)^5 : \|y\|_1 = 1\}$. We also define a 2-way variant, where the task is to estimate the probability of striking out, in which case $\mathcal{Y} = [0,1]$.
  Dirichlet-CADET and beta-CADET produce densities w.r.t. the Lebesgue measure over Δ^4 and $[0,1]$, respectively, and thus are appropriate for the task. We compare these models to three Gaussian-CADET models and to RFCDE, using the ALRT from (13) for the 5-way Gaussian-CADET and RFCDE models to convert to a problem over $\mathbb{R}^4$, letting the strikeout probability be the last (asymmetric) variable of the transformation. In the 2-way task, we compare beta-CADET to Gaussian-CADET and RFCDE without transformation, although a density estimate ρ from the non-beta models has $\int_0^1 \rho(y)\ \mathrm{d}y < 1$ (see Sect. 4.4). This disadvantage is intrinsic to these approaches and highlights the flexibility of CADET with transformation functions.
  Interpretability is particularly difficult in the 5-way task. Here the Dirichlet estimates of Dirichlet-CADET should be understood by any analyst familiar with compositional statistics, making this model the most interpretable. The Gaussian models are also quite simple, but although standard in compositional statistics, some effort is required to interpret the ALRT (see (13)), which makes the Gaussian models behave roughly like log-Gaussian models. With covariance matrices, Gaussian-CADET can model *correlations* between (approximate) log-frequencies, e.g., between the probabilities of reaching first-base and reaching second-base, unlike the Dirichlet distribution, which has only 5 parameters. Thus, although inherently more complicated, Gaussian-CADET remains interpretable, and can even yield insights that would be impossible for Dirichlet-CADET.
- The task on the SML2010 dataset (Zamora-Martínez et al. 2014) is to estimate the time of day, represented as a value in $[0,1)$, where $1/2$ is noon. This task is interesting for its *cyclic* nature. Classical regression struggles around midnight, as training points immediately before and after midnight average to noon (maximally incorrect), and non-parametric methods fail to enforce the constraint that predicted times be on the interval $[0,1)$, nor do they leverage the cyclic nature of the label space. We use the *Von-Mises* distribution family, scaled to have support $[0,1)$, as well as Gaussian-CADET and beta-CADET, for our parametric models, and compare to RFCDE.
- We use many UCI datasets (Table 2) to evaluate the efficacy of the impurity criterion used by CADET, the use of union families, and the competitiveness of CADET against RFCDE on real-world learning tasks. Each task is a multivariate conditional density estimation task, with each label in $\mathcal{Y} = \mathbb{R}^{\dim(\mathcal{Y})}$. Most of these datasets are intended for univariate classification or regression, but we instead predict the continuous variables from the

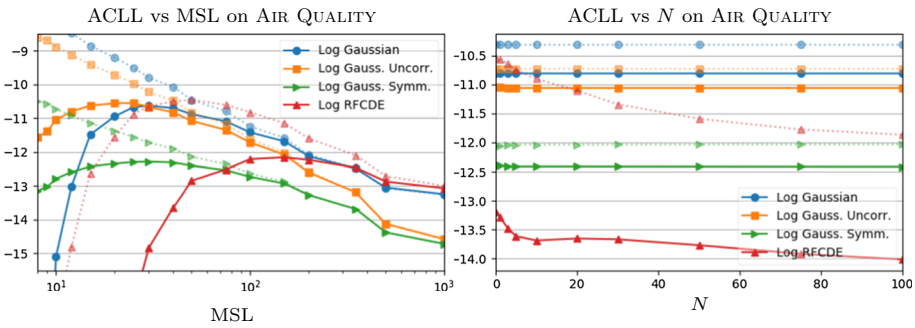

Fig. 1 ACLL as a function of MSL (left) and as a function of the number of noise features N (right) on the Air Quality dataset. Test ACLL plotted with solid lines, and training ACLL with dotted lines

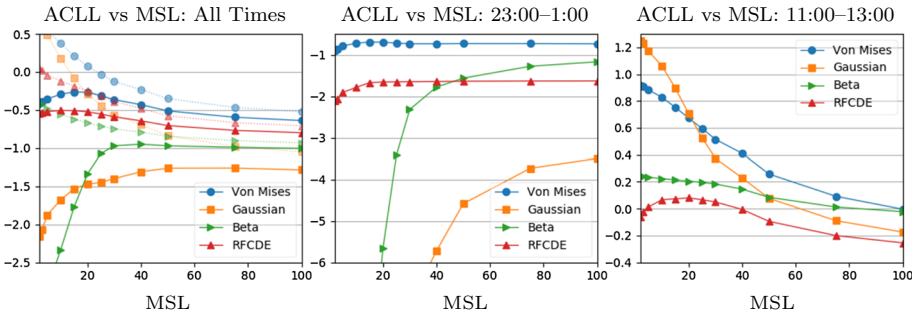

Fig. 2 Experiments with Sml2010. Dotted lines denote training ACLL, and solid lines test ACLL. Test ACLL on all times, 23:00–1:00, and 11:00–13:00 are plotted seperately

categorical or integer-valued variables. In some cases, due to many missing values or lack of features, we leave some continuous values as features; the details are presented in the supplementary material. We compare several CaDET variants and RFCDE on these datasets.

6.1 Results

Impact of minimum samples per leaf on overfitting We first study how the (MSL) parameter, which controls the *minimum number of training samples per leaf* (enforced by the learning procedure), impacts overfitting in CDE trees. We plot MSL versus training and test ACLL on the Air Quality dataset in Fig. 1 (left). Here we consider only single trees, as diversity in random forests tends to obscure overfitting in individual trees.

We see classic *bias-variance trade-off* curves for all models, with training ACLL monotonically decreasing with the MSL, and test ACLL first increasing, then decreasing. The training-test ACLL difference is a measure of overfitting, and here it decreases as MSL increases, and also as the number of *parameters* in each parametric family (see Table 1) decreases. The CaDET trees all perform optimally at MSL ≈ 25, whereas RFCDE reaches optimal performance with MSL ≈ 100, illustrating the lower sample complexity of parametric methods (see Sect. 4.4).

Figure 2 shows ACLL as function of MSL on the SML2010 task. In Fig. 2 (left), we see that the Von-Mises-CADET outperforms all competitors, which is unsurprising, as Von-Mises density estimates are best able to represent uncertainty across the midnight boundary. Indeed in Fig. 2 (center), we see that the Von-Mises-CADET ACLL decreases least when considering only test samples on the 23:00–1:00 interval, while the Gaussian-CADET (which is least able to split mass between late night and early morning) ACLL decreases the most. On the 11:00–13:00 interval, Fig. 2 (right), Gaussian-CADET outperforms the remaining models for small MSL (i.e., large trees with many leaves), but for higher MSL values, this advantage disappears. These results support our claim that parametric models that leverage domain-specific knowledge (in this case the cyclic nature of time) are superior to generic models that do not.

Susceptibility to irrelevant features We now examine the response of the models to irrelevant features: a well-designed impurity criterion should be imperturbable to such features and choose splits independently from them. We augment the AIR QUALITY dataset by generating N additional noise features, where each feature value for each sample was drawn i.i.d. from the standard Gaussian distribution. We train models using MSL 50 on this augmented dataset. We plot training and test ACLL as a function of N in Fig. 1 (right). As N increases, CADET ACLL decreases almost imperceptibly, whereas RFCDE ACLL drops sharply and significantly. RFCDE's performance drop is not due to overfitting to the noise features, as both test and *training* ACLL rapidly decrease. Rather, we attribute it to the *approximation error* of its learning algorithm, which is a consequence of the chosen impurity criterion. The heuristic $\mathsf{m}_{\mathsf{mise}}$ estimate used by RFCDE inadequately assesses the quality of splits, thus RFCDE often splits on noise features, degrading model accuracy. The $\mathsf{m}_{\mathsf{ece}}$ used by CADET strongly disincentivizes splits on noise features, resulting in similar models regardless of N.

Effectiveness of the chosen impurity criterion We now examine the importance of the impurity criterion via ablation. We train Gaussian-CADET trees with the MSE impurity from (1) instead of the ECE from (10), and train "vanilla" CADET trees as a control. The results are shown in the two leftmost columns of Table 2 (we discuss the other columns later). We report ACLL to measure model accuracy, and we quantify interpretability using *model size*, defined as the total number of continuous parameters required to represent the distributions at each leaf of the tree. Vanilla CADET yields on average, and more often than not, higher (better) ACLL scores, with much smaller, thus more interpretable, models.

Dependence on training set size We now evaluate the behavior of the ACLL as we increase the training set size n on the BATTING dataset. The MSL is fixed to $\lfloor \sqrt{n} + \frac{1}{2} \rfloor$, and test ACLL is computed on all samples not in the training set. We plot tree and forest experiments in Fig. 3, though overfitting is clearer in the trees.

In the 2-way task, when using trees, beta-CADET performs the best, though with sufficiently large samples, all models are comparable. In particular, each CADET model uses a 2-parameter distribution, so we expect a similar amount of overfitting in each, and indeed we see similar rates of improvement as the training size increases. RFCDE, as expected, overfits more due to its KDE estimates: its rate of improvement levels off more slowly than the CADET models.

In the 5-way tree task, $\dim(\Theta)$, which varies between 5 for the Dirichlet and symmetric Gaussian families, and 14 for the Gaussian family, strongly influences model performance. Each model outperforms all others for a *contiguous range* of n, and these ranges occur in order of $\dim(\Theta)$, with RFCDE beating the CADET models only for the highest n we examined. The fact that RFCDE beats all other models with sufficient data is unsurprising, as its KDE

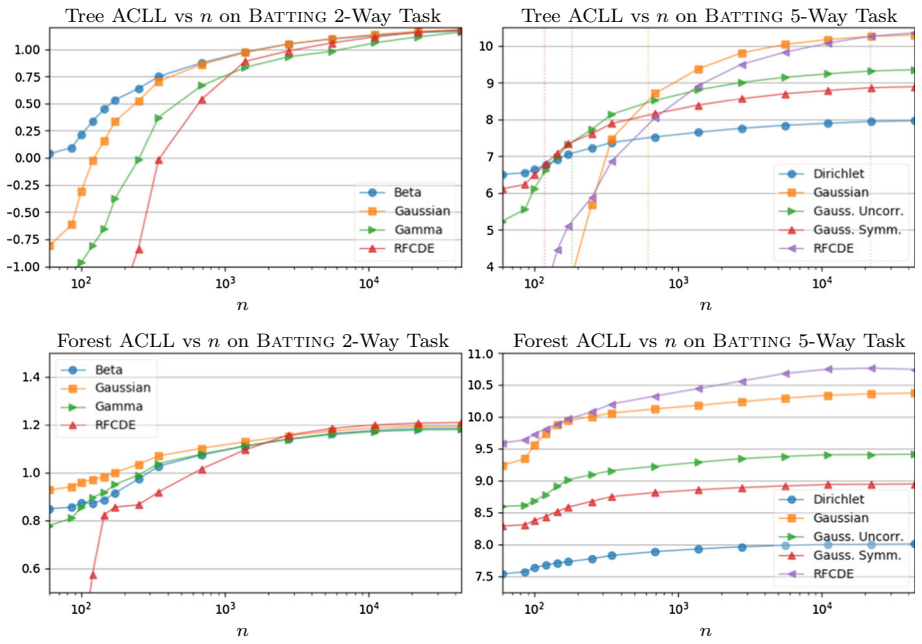

Fig. 3 Test ACLL versus training size n on the 2-way and 5-way BATTING tasks. Vertical lines mark where one model overtakes another in the 5-way experiment

estimates are *consistent*, thus with enough data should outperform the parametric estimates of CADET. The BATTING dataset contains 88,461 samples, and with only 7 features, we would expect simple *conditional density* relationships between $\mathcal{X}$ and $\mathcal{Y}$, thus this task, relative to the others, should measure a model's capacity to fit *unconditional densities* (at leaves) more than its ability to model *conditional structure* via splits.

These experiments highlight that CADET is not only particularly well-suited to small-sample settings, but also that non-parametric methods overtake CADET only when an enormous amount of data is available, even on very simple datasets. The case for CADET is even stronger when interpretability is considered: CADET trees have $\mathbf{O}\big(\dim(\Theta)\sqrt{n}\big)$ total parameters, while RFCDE trees have $\mathbf{\Theta}\big(\dim(\mathcal{Y})n\big)$, as they must store *all training labels* at tree leaves.

Forests improve over trees for every model examined in this experiment. The most significant improvement is in *small-sample* performance, which is unsurprising, as forests combine estimates across trees, thus prediction are based on larger numbers of training samples. The effect is most pronounced with RFCDE, as while its small-sample performance is still worse than all CADET forests, with enough data, it eventually outperforms them. Again we conjecture that this is because the BATTING tasks primarily assess *unconditional density estimation* (at leaves), and the bagging in forests reduces KDE overfitting in RFCDE.

Effect of using union families We study CADET with a union family, containing the unconstrained, uncorrelated, and symmetric variants of the Gaussian and log-Gaussian families. As the three variants of the Gaussian and log-Gaussian families (each) are nested, the uncorrelated and symmetric families never uniquely maximize sample likelihood. We employ AIC regularization from (16) to incentivize the uncorrelated and symmetric families.

In the rightmost six columns of Table 2, we compare union-CADET to Gaussian-CADET and RFCDE. The union-CADET trees significantly outperform the Gaussian-CADET trees, as measured by ACLL, while maintaining significantly smaller model sizes. CADET produces smaller models than RFCDE, which averages 6458 distribution parameters per tree, while producing less accurate (as measured by ACLL) models. The average training-test ACLL gap for the CADET models is ≈ 0.9, but for RFCDE it is ≈ 0.1. We thus claim that RFCDE is underfitting, and that its low training-set ACLL is due to poor split selection, since if all else were equal, the KDE at RFCDE leaves should be able to overfit much more than CADET's parametric estimates.

7 Conclusion

We present CADET, a tree-based algorithm for parametric CDE. CADET learns interpretable models that produce interpretable estimates. CADET trees are built by minimizing the *Empirical Cross-Entropy* (ECE) impurity criterion. ECE is *specific to CDE*, thus creates better splits that lead to better estimates than *generic regression impurity criteria*. CADET is a natural generalization of both MSE regression trees and information-gain classification trees, and attains the same training time and space complexities, under mild conditions. Our experimental evaluation shows that CADET is less prone to overfitting than existing CDE tree-based algorithms, and can outperform them in both accuracy and interpretability.

References

Agarwal, R., Chen, Z., & Sarma, S. V. (2017). A novel nonparametric maximum likelihood estimator for probability density functions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *7*, 1294–1308.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B (Methodological)*, *44*(2), 139–177.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Davidson: Chapman and Hall/CRC.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Chaudhuri, P., Loh, W. Y., et al. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, *8*(5), 561–576.

De Vito, S., Massera, E., Piga, M., Martinotto, L., & Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, *129*(2), 750–757.

Di Mauro, N., Vergari, A., Basile, T. M., & Esposito, F. (2017). Fast and accurate density estimation with extremely randomized cutset networks. In: *Joint European conference on machine learning and knowledge discovery in databases* (pp. 203–219). Berlin: Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.

Halmos, P. R., Savage, L. J., et al. (1949). Application of the radon-nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, *20*(2), 225–241.

Hothorn, T., & Zeileis, A. (2017). *Transformation forests*. arXiv preprint arXiv:1701.02110.

Kanazawa, Y. (1993). Hellinger distance and Kullback–Leibler loss for the kernel density estimator. *Statistics & Probability Letters*, *18*(4), 315–321.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, *39*(3), 399–409.

Lahman, S. (2018). *Sean Lahman's baseball archive*. http://www.seanlahman.com/baseball-archive/statistics/.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.

Molina, A., Vergari, A., Di Mauro, N., Natarajan, S., Esposito, F., & Kersting, K. (2018). Mixed sum-product networks: A deep architecture for hybrid domains. In: *Thirty-second AAAI conference on artificial intelligence*.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, *135*(3), 370–384.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pospisil, T., & Lee, A. B. (2018). *RFCDE: Random forests for conditional density estimation*. arXiv preprint arXiv:1804.05753.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Rahman, T., Kothalkar, P., & Gogate, V. (2014). Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of Chow–Liu trees. In: *Joint European conference on machine learning and knowledge discovery in databases* (pp. 630–645).

Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis II*, *25*, 31.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shuford, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, *31*(2), 125–145.

Weisberg, S. (2005). Binomial regression. In S. Weisberg (Ed.), *Applied linear regression* (3rd ed., pp. 253–54). Hoboken, NJ: Wiley.

Zamora-Martínez, F., Romeu, P., Botella-Rocamora, P., & Pardo, J. (2014). On-line learning of indoor temperature forecasting models towards energy efficiency. *Energy and Buildings*, *83*, 162–172.

Zhu, J., & Hastie, T. (2002). Kernel logistic regression and the import vector machine. In *Advances in neural information processing systems* (pp. 1081–1088).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.