Journal of Machine Learning Research 23 (2022) 1-55 Submitted 7/21; Revised 11/21; Published 2/22

LinCDE: Conditional Density Estimation via Lindsey's Method

Zijun Gao ZIJUNGAO@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Trevor Hastie HASTIE@STANFORD.EDU
Department of Statistics and Department of Biomedical Data Science
Stanford University
Stanford, CA 94305, USA

Editor: Victor Chernozhukov

Abstract

Conditional density estimation is a fundamental problem in statistics, with scientific and practical applications in biology, economics, finance and environmental studies, to name a few. In this paper, we propose a conditional density estimator based on gradient boosting and Lindsey's method (LinCDE). LinCDE admits flexible modeling of the density family and can capture distributional characteristics like modality and shape. In particular, when suitably parametrized, LinCDE will produce smooth and non-negative density estimates. Furthermore, like boosted regression trees, LinCDE does automatic feature selection. We demonstrate LinCDE's efficacy through extensive simulations and three real data examples.

Keywords: Conditional Density Estimation, Gradient Boosting, Lindsey's Method

1. Introduction

In statistics, a fundamental problem is characterizing how a response depends on a set of covariates. Numerous methods have been developed for estimating the mean response conditioning on the covariates—the so-called regression problem. However, the conditional mean may not always be sufficient in practice, and various distributional characteristics or even the full conditional distribution are called for, such as the mean-variance analysis of portfolios (Markowitz, 1959), the bimodality of gene expression distributions (DeSantis et al., 2014; Moody et al., 2019), and the peak patterns of galaxy redshift densities (Ball et al., 2008). Conditional distributions can be used for constructing prediction intervals, downstream analysis, visualization, and interpretation (Arnold et al., 1999). Therefore, it is worthwhile to take a step forward from the conditional mean to the conditional distribution.

There are several difficulties in estimating conditional distributions. First, distribution estimation is more complicated than mean estimation regardless of the conditioning. Second, as with conditional mean estimation, conditioning on a potentially large number of covariates suffers from the curse of dimensionality. When only a small subset of the covariates are relevant, proper variable selection is necessary to mitigate overfitting, reduce computational burden, and identify covariates that may be of interest to the practitioners.

In this paper, we develop a tree-boosted conditional density estimator based on Lindsey's method, which we call LinCDE (pronounced "linseed") boosting. LinCDE boosting is built on the base learner LinCDE tree. A LinCDE tree partitions the covariate space into subregions with homogeneous conditional distributions, estimates a local unconditional density in each subregion, and aggregates the unconditional densities to form the final conditional estimator. LinCDE boosting combines LinCDE trees to form a strong ensemble learner.

LinCDE boosting possesses several desirable properties. LinCDE boosting provides a flexible modeling paradigm and is capable of capturing distributional properties such as heteroscedasticity and multimodality. LinCDE boosting also inherits the advantages of tree and boosting methods, and in particular, LinCDE boosting is able to detect influential covariates. Furthermore, the conditional density estimates are automatically non-negative and smooth, and other useful statistics such as conditional quantiles or conditional cumulant distribution functions (CDFs) can be obtained in a straightforward way from the LinCDE boosting estimates.

The organization of the paper is as follows. In Section 2, we formulate the problem and discuss related work. We develop LinCDE in three steps:

- In Section 3, we describe Lindsey's method for (marginal) density estimation.
- In Section 4, we introduce LinCDE trees for conditional density estimation, which combine Lindsey's method with recursive partitioning.
- In Section 5, we develop a boosted ensemble model using LinCDE trees.

In Section 6, we discuss two optional but helpful preprocessing steps—response transformation and conditional mean centering. In Section 7, we evaluate the performance of LinCDE boosting on simulated data sets. In Section 8, we apply LinCDE boosting to three real data sets. We conclude the paper with discussions in Section 9 and provide links to the R software and instructions in Section 10. All proofs are deferred to the appendix.

2. Background

In this section, we formulate the conditional density estimation problem and discuss related work.

2.1 Problem Formulation

Let $y \in \mathbb{R}$ be a continuous response.[1] Let x be a d-dimensional covariate vector and $x^{(j)}$ be its j-th coordinate. We assume the covariates are generated from an unknown underlying distribution $f_x(x)$, and the response given the covariates are sampled from an unknown conditional density $f_{y|x}(y \mid x)$. The model is summarized as

$$\begin{aligned} x_i &\overset{i.i.d.}{\sim} f_x, \\ y_i \mid x_i &\overset{ind.}{\sim} f_{y|x}. \end{aligned} \tag{1}$$

1. The paper will focus on univariate responses, and the generalization to multivariate responses is straightforward and discussed in Section 9.

We observe n data pairs $\{(x_i, y_i)\}$ and aim to estimate the conditional density $f_{y|x}(y \mid x)$.

2.2 Literature

In general, there are three ways to characterize a conditional distribution: conditional density, conditional quantile, and conditional CDF. We categorize related works on conditional density, quantile, and CDF estimation according to the methodology and provide a review below. We conclude by discussing two desired properties of conditional density estimation and how our work complements the existing literature on the two points.

A line of study estimates the conditional distribution by localizing unconditional distribution estimators. Localization methods weight observations according to the distances between their covariates and those at the target point, and solve the unconditional distribution estimation problem based on the weighted sample. For conditional density, Fan et al. (1996) obtain conditional density estimates by local polynomial regression. For conditional quantile, Chaudhuri (1991a,b) partitions the covariate space into bins and fit a quantile model in each bin separately, and Yu and Jones (1998) tackle the conditional quantile estimation via local quantile loss minimization. For conditional CDF, Stone (1977) proposes a weighted sum of indicator functions, and Hall et al. (1999) consider a local logistic regression and a locally adjusted Nadaraya-Watson estimator. Localization methods enable systematic extensions of any unconditional estimator. Nevertheless, the weights usually treat covariates as equally important, and variable selection is typically not accommodated. This leaves the methods vulnerable to the curse of dimensionality.

As discussed by Stone (1991a,b), another approach making use of unconditional methods, first obtains the joint distribution estimate $\hat{f}_{y,x}(y, x)$ and the covariate distribution estimate $\hat{f}_x(x)$, and then follows

$$\hat{f}_{y|x}(y \mid x) = \hat{f}_{y,x}(y, x)/\hat{f}_x(x) \tag{2}$$

to derive the conditional density. Nevertheless, the joint distribution estimation is also challenging, if not more. Arnold et al. (1999) point out except for special cases like multivariate Gaussian, the estimation of a bivariate joint distribution in a certain exponential form is onerous due to the normalizing constant. Moreover, the approach is inefficient both statistically and computationally if the conditional distribution is comparatively simpler than the joint and covariate distributions; as an example, when the response is independent of the covariates.

A different thread directly models the dependence of the response's distribution on the covariates by a linear combination of a finite or infinite number of bases.

- For conditional density, Kooperberg and Stone (1991, 1992), Stone (1991b), Mâacsse and Truong (1999), and Barron and Sheu (1991) study the conditional logspline density model: modeling $\log(f_{y|x}(y \mid x))$ by tensor products of splines or trigonometric series and maximizing the conditional log-likelihood to estimate the parameters. The method is also known as entropy maximization subject to empirical constraints. In addition, Sugiyama et al. (2010) model the conditional density in reproducing kernel Hilbert spaces and estimate the loadings by the unconstrained least-squares importance fitting, and Izbicki and Lee (2016) expand the conditional density in the eigen-

functions of a kernel-based operator which adapts to the intrinsic dimension of the covariates.[2]

- For conditional quantiles, Koenker and Bassett Jr (1978) formulate conditional quantiles as linear functions of covariates and minimize quantile losses to estimate parameters. Koenker et al. (1994) explore quantile smoothing splines minimizing for a single covariate, and He et al. (1998) extend the approach to the bi-variate setting, i.e., two covariates. Li et al. (2007) propose kernel quantile regression (KQR) considering quantile regression in reproducing Hilbert kernel spaces. Belloni et al. (2019) approximate the conditional quantile function by a growing number of bases as the sample size increases.

- For conditional CDF, Foresi and Peracchi (1995) and Chernozhukov et al. (2013) consider distribution regression: estimating a sequence of conditional logit models over a grid of values of the response variable. Belloni et al. (2019) further extend the method to the high-dimensional-sparse-model setting.

The performance of parametric models depends on the selected bases or kernels. Covariate-specific bases or kernels require a lot of tuning, and the bases or kernels treat covariates equally, making the approaches less powerful in the presence of many nuisance covariates.

More recently, tree-based estimators arose in conditional distribution estimation. The overall idea is partitioning the covariate space recursively and fitting an unconditional model at each terminal node. For conditional quantiles, Chaudhuri and Loh (2002) investigate a tree-structured quantile regression. Nevertheless, the estimation of different quantiles requires separate quantile loss minimization, which complicates the full conditional distribution calculation. Meinshausen (2006) proposes the quantile regression forest (QRF) that computes all quantiles simultaneously. QRF first builds a standard random forest, then estimates the conditional CDF by a weighted distribution of the observed responses, and finally inverts the CDF to quantiles. For conditional CDF, Friedman (2019) proposes distribution boosting (DB). DB relies on Friedman's contrast trees—a method to detect the lack-of-fit regions of any conditional distribution estimator. DB estimates the conditional distribution by iteratively transforming the conditional distribution estimator and correcting the errors uncovered by contrast trees.

Beyond trees, neural network-based conditional distribution estimators have also been developed. For conditional quantile, standard neural networks with the quantile losses serve the goal. For conditional density, Bishop (1994, 2006) introduces the mixture density networks (MDNs) which model the conditional density as a mixture of Gaussian distributions. Neural network-based methods can theoretically approximate any conditional distributions well but are computationally heavy and lack interpretation.

In this paper, we propose LinCDE boosting which complements existing works by producing smooth density curves and performing feature selection.

- Compared to quantiles and CDF, smooth density curves are more suitable for the following reasons.

2. We remark that the approach in (Izbicki and Lee, 2016) adapts to the low intrinsic dimensionality of covariates but not the sparse dependency of the response on the covariates.

  - Smooth density curves can give valuable indications of distributional characteristics such as skewness and multimodality. Visualization of smooth density curves is comprehensible to practitioners and can yield self-evident conclusions or point the way to further analysis.
  - Conditional densities can be used to compute class-posterior probabilities using the Bayes' rule. The class-posterior probabilities can be further used for classification.
  - Densities can be used to detect outliers: if an observation lies in a very low-density region, the data point is likely to be an outlier.

  LinCDE boosting generates smooth density curves, while directly transforming estimates of conditional quantiles or CDF to conditional density estimates usually produces bumpy results.

- LinCDE trees and LinCDE boosting calculate feature importances and automatically focus on influential features in estimation. In contrast, methods like localization and modeling conditional densities in a linear space or RKHS often treat covariates equally, and covariate-specific neighborhoods, bases, and kernels may require a lot of tuning. Therefore, LinCDE boosting is more scalable to a large number of features and less sensitive to the presence of nuisance covariates.

3. Lindsey's Method

In this section, we first introduce the density estimation problem — an intermediate step towards the conditional density estimation. We then discuss how to solve the density estimation problem by Lindsey's method (Lindsey, 1974)—a stepping stone of LinCDE. Lindsey's method cleverly avoids the normalizing issue by discretization and solves the problem by fitting a simple Poisson regression. It can be thought of as a method for fitting a smooth histogram with a large number of bins.

We consider the density family

$$f(y) = \kappa(y)e^{g(y)}, \tag{3}$$

where $\kappa : \mathbb{R} \to \mathbb{R}$ is some carrying density, and $g(y)$ is known as a *tilting* function. The idea is that $\kappa(y)$ is known or assumed (such as Gaussian or uniform), and $g(y)$ is represented by a model. We represent g as a linear expansion

$$g(y) = z(y)^\top \beta + \beta_0, \tag{4}$$

where $z(y)$ is a basis of k smooth functions. As a simple example, if we use standard Gaussian as the carrying measure and choose $z(y)^\top = (y, y^2)$, the resulting density family corresponds to all possible Gaussian distributions. More generally we use a basis of natural cubic splines in $z(y)$ with knots spread over the domain of y, to achieve a flexible representation (Wahba, 1975, 1990).

Our goal is to find the density that maximizes the log-likelihood

$$\max_{\beta,\beta_0} \frac{1}{n}\sum_{i=1}^{n} \log(\kappa(y_i)) + z(y_i)^\top \beta + \beta_0, \quad \text{s.t.} \int \kappa(y)e^{z(y)^\top \beta+\beta_0} dy = 1. \tag{5}$$

The constrained optimization problem (5) can be simplified to the unconstrained counterpart below by the method of Lagrange multipliers (Silverman, 1986),

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\log(\kappa(y_i)) + z(y_i)^\top\beta + \beta_0\Big) - \int \kappa(y)e^{z(y)^\top\beta+\beta_0}dy, \tag{6}$$

where the multiplier turns out to be one.

The optimization problem (6) is difficult since the integral $\int \kappa(y)e^{z(y)^\top\beta+\beta_0}dy$ is generally unavailable in closed form. One way to avoid the integral is by discretization—the key idea underlying Lindsey's method. One divides the response range into B equal bins of width Δ with mid-points y_b. The integral is approximated by the finite sum

$$\int \kappa(y)e^{z(y)^\top\beta+\beta_0}dy \;\approx\; \sum_{b=1}^{B}\kappa(y_b)e^{z(y_b)^\top\beta+\beta_0}\Delta.$$

As for the first part of (6), one replaces y_i by its bin midpoint $y_{b(i)}$ and groups the observations,

$$\begin{aligned}\sum_{i=1}^{n}\log(\kappa(y_i)) + z(y_i)^\top\beta + \beta_0 &\approx \sum_{i=1}^{n}\log(\kappa(y_{b(i)})) + z(y_{b(i)})^\top\beta + \beta_0\\ &= \sum_{b=1}^{B} n_b\Big(\log(\kappa(y_b)) + z(y_b)^\top\beta + \beta_0\Big),\end{aligned}$$

where $b(i)$ denotes the bin that the i-th response falls in, and n_b represents the number of samples in bin b. Combining the above two parts, the Lagrangian function with response discretization takes the form

$$\frac{1}{n}\sum_{b=1}^{B} n_b\Big(\log(\kappa(y_b)) + z(y_b)^\top\beta + \beta_0\Big) - \sum_{b=1}^{B}\kappa(y_b)e^{z(y_b)^\top\beta+\beta_0}\Delta. \tag{7}$$

The objective function (7) is equivalent to that of a Poisson regression with B observations $\{n_b\}_{1\le b\le B}$ and mean parameters $\mu_b \propto \kappa(y_b)e^{z(y_b)^\top\beta}$. Therefore, Lindsey's method estimates the coefficient β by fitting the Poisson regression with predictors $z(y)$ and offset $\log(\kappa(y_b))$. The normalizing constant β_0 in (7) is absorbed in the Poisson regression's intercept. Despite the discretization error, Lindsey's estimates are consistent, asymptotically normal, and remarkably efficient (Moschopoulos and Staniswalis, 1994; Efron, 2004). We will demonstrate the efficacy of Lindsey's method in two examples at the end of this section.

The number of bins B balances the statistical performance and the computational complexity of Lindsey's method: as B increases, the discretized objective (7) approaches the original target (6), and the resulting estimator converges to the true likelihood maximizer; on the other hand, the computations increase linearly in B. This can become a factor later when we fit many of these Poisson models repeatedly.

The relationship with a histogram becomes clear now, as well. We could use the counts in the B bins to form a density estimate, but this would be very jumpy. Typically we would control this by reducing the number of bins. Lindsey's method finesses this by having B

large, but controlling the smoothness of the bin means via the $k \ll B$ basis functions and associated coefficients.

To control the model complexity further and avoid numeric instability, we add a regularization term to (6). For example, we can penalize deviations from Gaussian distributions via the regularizer (Wahba, 1977; Silverman, 1982, 1986)[3]

$$\int \left(\frac{d^3}{dy^3} \left(z(y)^\top \beta + \beta_0 \right) \right)^2 dy. \tag{8}$$

The penalty measures the roughness of the tilting function and is zero if and only if the tilting function's exponent is a quadratic polynomial, i.e., a Gaussian distribution. We also attach a hyper-parameter λ to trade-off the objective (6) and the penalty (8), and tune λ to achieve the best performance on validation data sets.[4]

It is convenient to tailor the spline basis functions to the penalty (8). Note that, for arbitrary bases $z(y)$, the penalty (8) is a quadratic form in β

$$\int \left(\frac{d^3}{dy^3} \left(z(y)^\top \beta + \beta_0 \right) \right)^2 dy = \sum_{j,l=1}^{k} \beta_j \beta_l \int z_j'''(y) z_l'''(y) dy =: \beta^\top \Omega \beta, \tag{9}$$

where $\Omega_{jl} = \int z_j'''(y) z_l'''(y) dy$. We transform our splines so that the associated $\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_k)$ is diagonal and the penalty reduces to a weighted ridge penalty

$$\sum_{j=1}^{k} \omega_j \beta_j^2. \tag{10}$$

(details in Appendix B). Figure 1 depicts an example of the transformed spline bases (in increasing order of ω_j) and the corresponding smoothed versions. Among the transformed bases, the linear and quadratic components (the first and the second bases in Figure 1) are not shrunk by the roughness penalty (Claim 1), and higher-complexity splines are more heavily penalized (Hastie et al., 2009, chap. 5, for example).

Claim 1 *Assume $u \in \mathbb{R}^k$ and $\Omega u = 0$. Then $z(y)^\top u$ is a linear or quadratic function of y.*

We provide examples of Lindsey's performance in estimating bimodal and skewed distributions in Appendix G.

4. LinCDE Trees

In this section, we extend the density estimation problem to the conditional density estimation problem. We introduce LinCDE trees combining Lindsey's method and recursive partitioning: LinCDE trees partition the covariate space and estimate a local unconditional density via Lindsey's method in each subregion.

3. We regard exponential distribution as a special case of Gaussian distribution with $\sigma^2 = \infty$.
4. The hyper-parameter is a function of the penalized Poisson regression's degrees of freedom (see Appendix A for more details), and we tune the degrees of freedom to achieve the best performance on validation data sets.

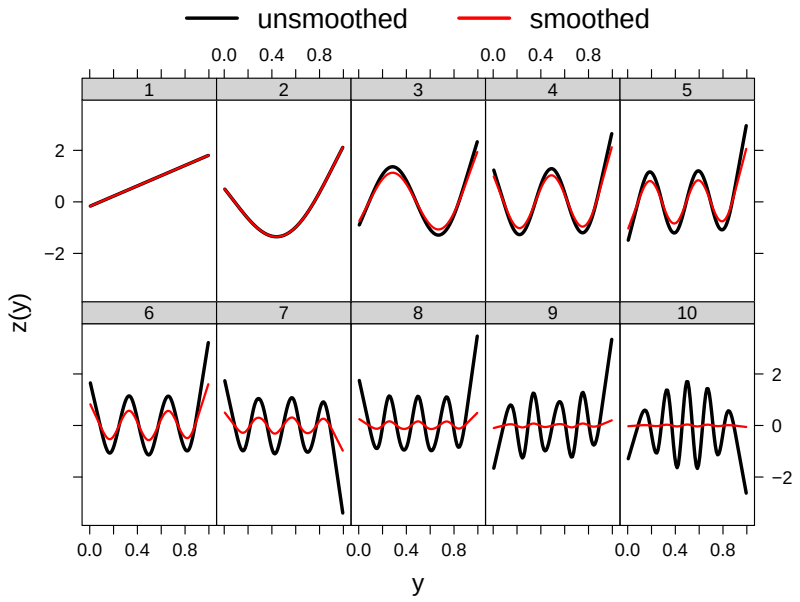


Figure 1: Transformed natural cubic spline bases. The transformed basis functions before smoothing (in black) are normalized such that $\int (z_j(y))^2 dy = 1$ and ordered by increasing penalty factors. The damped functions (in red) represent smoothed basis functions (with 6 degrees of freedom).

To begin, we restate the target density family (3) in the language of exponential families. We call $z(y)$ the sufficient statistics and $g(y)$ the natural parameter. We replace the normalizing constant β_0 by the negative cumulant generating function $\psi(\beta)$ defined as

$$e^{\psi(\beta)} = \int \kappa(y) e^{z(y)^\top \beta} dy.$$

As a result, the density normalizing constraint of $\kappa(y)e^{z(y)^\top \beta - \psi(\beta)}$ is automatically satisfied.

In the conditional density estimation problem, we consider the target family generalized from (3)

$$f_{y|x}(y \mid x) = \kappa(y) e^{z(y)^\top \beta(x) - \psi(\beta(x))}. \tag{11}$$

The dependence of the response on the covariates is encoded in the parameter function $\beta(x)$.

- If there is only $k = 1$ basis function $z(y)$ and the parameter function satisfies $\beta(x) = x^\top \theta$, then the conditional density family (11) reduces to a generalized linear model.
- If we consider a standard Gaussian prior, basis function $z(y) = y$, and tree-structured $\beta(x)$, then the conditional density family (11) reduces to a regression tree model. In fact, the optimization problem based on (11) is equivalent to finding the tree $\beta(x)$ that minimizes the sum of squares $\sum_{i=1}^n (y_i - \beta(x_i))^2$.

Similar to the density estimation problem, we aim to find the member in the family (11) that maximizes the conditional log-likelihood with the ridge penalty (10)

$$\begin{aligned}\ell(\mathcal{R}_0;\beta) :=& \sum_{i=1}^{n}\left(\log(\kappa(y_i)) + z(y_i)^\top\beta(x_i) - \psi(\beta(x_i)) - \lambda\sum_{j=1}^{k}\omega_j\beta_j^2(x_i)\right)\\ =& \sum_{i=1}^{n}\Big(\log(\kappa(y_i)) + z(y_i)^\top\beta(x_i) - \psi(\beta(x_i))\Big) - \lambda\sum_{i=1}^{n}\sum_{j=1}^{k}\omega_j\beta_j^2(x_i),\end{aligned} \tag{12}$$

where $\mathcal{R}_0$ denotes the full covariate space. If $\beta(x)$ is constant, the problem (12) simplifies to the unconditional problem (5).

The conditional density estimation problem (12) is more complicated than the unconditional version due to the covariates x:

1. Given a covariate configuration x, there is often at most one observation whose covariates take the value x, and it is infeasible to estimate the multi-dimensional natural parameter $\beta(x)$ based on a single observation;

2. There may be a multitude of covariates, and only a few are influential. Proper variable selection or shrinkage is necessary to avoid serious overfitting.

One way to finesse these difficulties is to use trees (Breiman et al., 1984). We divide the covariate space into subregions with approximately homogeneous conditional distributions, and in each subregion, we estimate a density independent of the covariate values. We name the method "LinCDE trees". In response to the first difficulty, by conditioning on a subregion instead of a specific covariate value, we have more samples for local density estimation. In response to the second difficulty, trees perform internal feature selection, and are thus resistant to the inclusion of many irrelevant covariates. Moreover, the advantages of tree-based methods are automatically inherited, such as being tolerant of all types of covariates, computationally efficient, and easy to interpret.

Before we delve into the details, we again draw a connection between LinCDE trees and a naive binning approach—fitting a multinomial model using trees. The naive approach discretizes the response into multiple bins and predicts conditional cell probabilities through recursive partitioning. The normalized conditional cell probabilities serve as an approximation of the conditional densities, and the more bins used, the higher resolution the approximation is. The naive approach is able to detect subregions with homogeneous multinomial distributions. However, the estimates are bumpy, especially with a large number of bins. To stabilize the method, restrictions enforcing smoothness are required, and LinCDE trees realize the goal by modeling the density exponent using splines.

We now explain how LinCDE trees work. In standard tree algorithms, there are two major steps:

- *Splitting*: partitioning the covariate space into subregions;

- *Fitting*: performing estimation in each subregion. The estimator is usually obtained by maximizing a specific objective function. For example, in a regression tree with ℓ_2 loss, the estimator is the sample average; in a classification tree with misclassification error, the estimator is the majority's label.

The fitting step is a direct application of Lindsey's method in Section 3. In a subregion $\mathcal{R}$, we treat the natural parameter functions as a constant vector and solve the density estimation problem via Lindsey's method. We denote the objective function value in region $\mathcal{R}$ with parameter β by $\ell(\mathcal{R};\beta)$, and let $\hat{\beta}_{\mathcal{R}} := \arg\max_\beta \ell(\mathcal{R};\beta)$.

Now for the splitting step. Similar to standard regression and classification trees, we proceed with a greedy algorithm and select the candidate split that improves the objective the most. Mathematically, starting from a region $\mathcal{R}$, we maximize the improvement statistic

$$\Delta\ell(\mathcal{R}, s) := \ell(\mathcal{R}_{s,L}; \hat{\beta}_{\mathcal{R}_{s,L}}) + \ell(\mathcal{R}_{s,R}; \hat{\beta}_{\mathcal{R}_{s,R}}) - \ell(\mathcal{R}; \hat{\beta}_{\mathcal{R}}), \tag{13}$$

where $\mathcal{R}_{s,L}$ and $\mathcal{R}_{s,R}$ are the regions on the left and right of the candidate split, respectively. Direct computation of the difference (13) requires running Lindsey's method twice for *each* candidate split s to obtain $\hat{\beta}_{\mathcal{R}_{s,L}}$, $\hat{\beta}_{\mathcal{R}_{s,R}}$, and the total computation time is prohibitive. Instead, we approximate the difference (13) by a simple quadratic term in Proposition 2, which can be computed much faster.

Proposition 2 (Improvement approximation for LinCDE trees) *Let $n_{\mathcal{R}}$, $\bar{z}_{\mathcal{R}}$ be the sample size and average sufficient statistics in a region $\mathcal{R}$. Assume that $\nabla^2\psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega$ is invertible, then for a candidate split s,*

$$\frac{1}{n_{\mathcal{R}}}\Delta\ell(\mathcal{R}, s) = \frac{n_{\mathcal{R}_{s,L}} n_{\mathcal{R}_{s,R}}}{2n_{\mathcal{R}}^2}(\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}_{s,R}})^\top \left(\nabla^2\psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} (\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}_{s,R}}) + r_s,$$

where the remainder term satisfies $r_s = O\left(\|\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}}\|_2^3 + \|\bar{z}_{\mathcal{R}_{s,R}} - \bar{z}_{\mathcal{R}}\|_2^3\right)$.

Proposition 2 writes the difference (13) as a quadratic form plus a higher-order residual term. If $z(y) = y$, the model amounts to a regression tree, and the residual term is zero. For general $z(y)$, when the average sufficient statistics $\bar{z}_{\mathcal{R}_{s,L}}$, $\bar{z}_{\mathcal{R}_{s,R}}$ are similar, the residual term is of smaller order than the quadratic form and can thus be dropped; when $\bar{z}_{\mathcal{R}_{s,L}}$, $\bar{z}_{\mathcal{R}_{s,R}}$ are considerably different, the residual term is not guaranteed to be small theoretically. However, we empirically demonstrate in Appendix C that at such splits, the quadratic form is still sufficiently close to the true log-likelihood difference. Based on this empirical evidence, we use the quadratic approximation to determine the optimal splits.

The quadratic approximation suggested by Proposition 2 is the product of the squared difference between the average sufficient statistics in $\mathcal{R}_L$ and $\mathcal{R}_R$ normalized by $\nabla^2\psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega$, further multiplied by the sample proportions in $\mathcal{R}_L$ and $\mathcal{R}_R$. By selecting the candidate split that maximizes the quadratic term, we will end up with two subregions different in the sufficient statistics means and reasonably balanced in sample sizes.

To compute the quadratic approximation, we need subsample proportions $n_{\mathcal{R}_{s,L}}/n_{\mathcal{R}}$, $n_{\mathcal{R}_{s,R}}/n_{\mathcal{R}}$, average sufficient statistics $\bar{z}_{\mathcal{R}_{s,L}}$, $\bar{z}_{\mathcal{R}_{s,R}}$, and the inverse matrix of $\nabla^2\psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega$. For the candidate splits based on the same covariate, $\{n_{\mathcal{R}_{s,L}}, n_{\mathcal{R}_{s,R}}, \bar{z}_{\mathcal{R}_{s,L}}, \bar{z}_{\mathcal{R}_{s,R}}\}$ can be computed efficiently for all split points by scanning through the samples in $\mathcal{R}$ once. For all candidate splits, this takes $O(dn_{\mathcal{R}}k)$ operations in total. The matrix $\nabla^2\psi(\hat{\beta}_{\mathcal{R}})$ is shared by all candidate splits and needs to be computed only once. The difficulty is that $\nabla^2\psi(\beta)$ is often unavailable in closed form. However, since $\nabla^2\psi(\beta_{\mathcal{R}})$ is the covariance matrix of the sufficient statistics $z(y)$ if the responses y are generated from the model parameterized

by $\beta_{\mathcal{R}}$, we apply Lindsey's method to estimate $\beta_{\mathcal{R}}$ and compute the covariance matrix of the sufficient statistics based on the multinomial cell probabilities, which takes $O(k^2B)$. Appendix F shows the resulting covariance matrix approximates $\nabla^2\psi(\hat{\beta}_{\mathcal{R}})$ with a fine discretization. The total time complexity of the above splitting procedure is summarized in the following Proposition 3.

Proposition 3 *Assume that there are S candidate splits, k basis functions, d covariates, B discretization bins, and $n_{\mathcal{R}}$ observations in the current region. Then the splitting step for LinCDE trees is of time complexity $O(dn_{\mathcal{R}}k + k^2B + k^3 + Sk^2)$.*

According to Proposition 3, the computation time based on the quadratic approximation is significantly reduced compared to running Lindsey's methods in $\mathcal{R}_{s,L}$ and $\mathcal{R}_{s,R}$ for all candidate splits, which takes $\tilde{O}(S(n_{\mathcal{R}}k + k^2B + k^3))$.

Having found the best split $s_{\max}$, we partition $\mathcal{R}$ into two subregions $\mathcal{R}_{s_{\max},L}$ and $\mathcal{R}_{s_{\max},R}$, and repeat the splitting procedure in the two subregions. Along the recursively partitioning, the response distribution's heterogeneity is reduced. The fitting and the splitting steps of LinCDE trees are summarized below, the complete algorithm is given in Algorithm 1, and implementation details are displayed in Section 10. Stopping criteria for LinCDE trees are discussed in Appendix D.

- *Fitting (LinCDE tree).* At a region $\mathcal{R}$:
  1. Count the number of observations $\{n_{\mathcal{R},b}\}$ in each bin.
  2. Fit a Poisson regression model with the response variable $\{n_{\mathcal{R},b}\}$, regressors $z(y_b)$, the offset $\log(\kappa(y_b))$, and a weighted ridge penalty.[5] Denote the estimated coefficients by $\hat{\beta}_{\mathcal{R}}$.
- *Splitting (LinCDE tree).* At a region $\mathcal{R}$:
  1. Compute $\{n_{\mathcal{R}_{s,L}}, n_{\mathcal{R}_{s,R}}, \bar{z}_{\mathcal{R}_{s,L}}, \bar{z}_{\mathcal{R}_{s,R}}\}$ for each candidate split, and approximate $\nabla^2\psi(\hat{\beta}_{\mathcal{R}})$ by $z(y)$'s covariance matrix in $\mathcal{R}$ using $\hat{\beta}_{\mathcal{R}}$.
  2. For each candidate split $s \in \mathcal{S}$, compute the quadratic approximation $\widehat{\Delta}\ell(\mathcal{R}, s)$ by Proposition 2, and choose the split $s_{\max} = \arg\max_{s\in\mathcal{S}} \hat{\Delta}\ell(\mathcal{R}, s)$.

Algorithm 1: LinCDE tree

Start at the full covariate space.
1. Apply *Fitting (LinCDE tree)* and obtain the natural parameter estimator $\hat{\beta}$.
2. Apply *Splitting (LinCDE tree)* and obtain the optimal split $s_{\max}$.
3. Repeat steps 1 and 2 to the left and right children of $s_{\max}$ until the stopping rule is satisfied, e.g., the maximal tree depth is reached. Output the natural parameter estimator $\hat{\beta}$ in each subregion.

5. Weights (penalty factors) of the ridge penalty are $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_k]^\top$.

To conclude the section, we demonstrate the effectiveness of LinCDE trees in three toy examples. We generate 10 covariates randomly uniformly on $[-1, 1]$. The response follows

$$f_{y|x}(y \mid x) = \begin{cases} f_1(y), & x^{(1)} < -0.2, \\ f_2(y), & x^{(1)} \geq -0.2, x^{(2)} \geq 0, \\ f_3(y), & x^{(1)} \geq -0.2, x^{(2)} < 0, \end{cases} \tag{14}$$

with three different local densities $f_l(x)$, $1 \leq l \leq 3$, varying in variance, number of modes, and skewness. The response distribution is determined by the first two covariates and independent of the rest. In Figure 2, we plot the average conditional density estimates in the three subregions. LinCDE trees are able to distinguish the densities differing in the above characteristic properties and produce good fits. We also compute the normalized importance score—the proportion of overall improvement in the split-criterion attributed to each splitting variable. In all settings, the first two covariates contribute over 99% importance. In other words, LinCDE trees focus on the first two influential covariates and avoid splitting at nuisance covariates.

5. LinCDE Boosting

Although LinCDE trees are useful as stand-alone tools, our ultimate goal is to use them as *weak learners* in a boosting paradigm. Standard tree boosting (Friedman, 2001) builds an additive model of shallow trees in a forward stagewise manner. Though a single shallow tree is high in bias, tree boosting manages to reduce the bias by successively making small modifications to the current estimate.[6]

We proceed with the boosting idea and propose LinCDE boosting. Starting from a null estimate, we iteratively modify the current estimate by modifying the natural parameter functions via a LinCDE tree. In particular, at the t-th iteration,

$$\begin{aligned} \gamma^t(x) &= \underset{\text{LinCDE tree } \gamma(x)}{\arg\max} \ell(\mathcal{R}_0; \beta^t(x) + \gamma(x)), \\ \beta^{t+1}(x) &\leftarrow \beta^t(x) + \gamma^t(x). \end{aligned} \tag{15}$$

Section 5.1 gives details. We remark that the LinCDE tree modifier $\gamma^t(x)$ for boosting is an expanded version of that in Section 4: in previous LinCDE trees, all samples share the same carrying density $\kappa(y)$, while in LinCDE trees for boosting, the carrying densities $\kappa(y)e^{z(y)^\top \beta^t(x) - \psi(\beta^t(x))}$ differ across units. We elaborate on LinCDE trees with heterogeneous carrying densities in Sections 5.2 and 5.3.

Before discussing the details of LinCDE boosting, we compare LinCDE boosting and LinCDE trees on a toy example in Figure 3. We consider a locally Gaussian distribution with heterogeneous mean and variance

$$y = x^{(1)} + 0.5x^{(2)}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1). \tag{16}$$

6. We remark that another successful ensemble method —random forests (Breiman, 2001)—are not appropriate for LinCDE trees. Random forests construct a large number of trees with low correlation and average the predictions. Deep trees are grown to ensure low-bias estimates, which is, however, unsatisfactory here because deep LinCDE trees will have leaves with too few observations for density estimation.

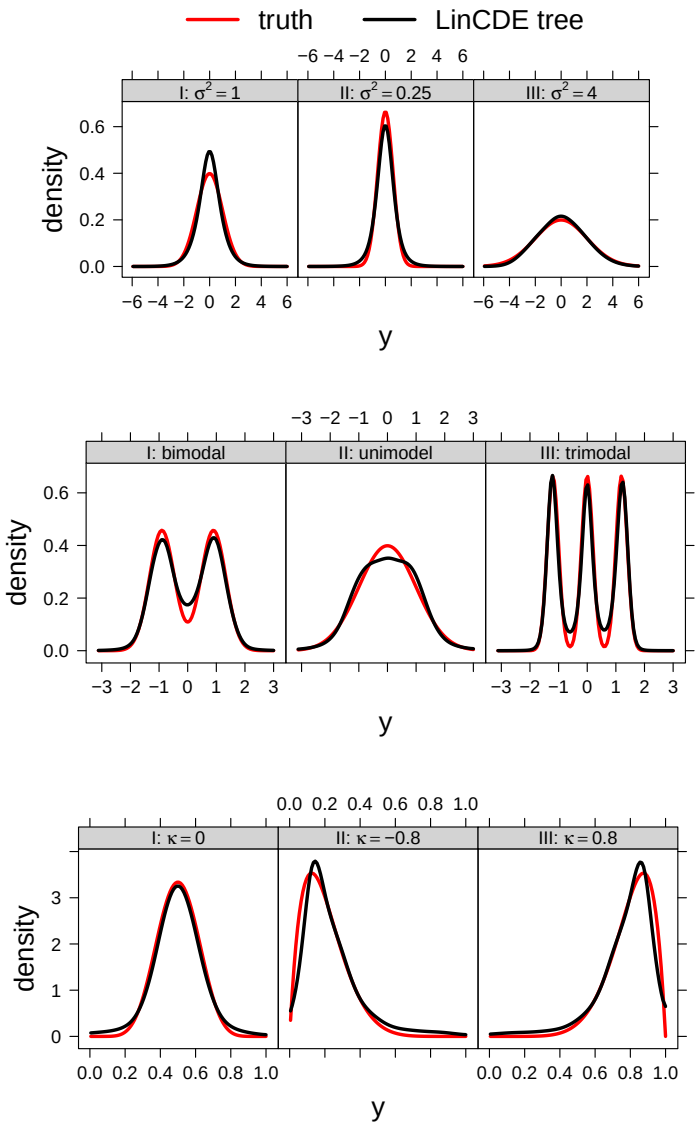


Figure 2: LinCDE trees' conditional density estimates of heteroscedastic, multimodal and skewed distributions. The responses are generated from the model (14). The local densities are Gaussian with variances $\sigma^2 \in \{0.25, 1, 4\}$ (first row), Gaussian mixtures with 1, 2, 3 components (second row), and Beta distributions with skewness $\kappa \in \{-0.8, 0, 0.8\}$ (third row). In each trial, we sample 400 observations. We use 10 natural cubic splines , 5 degrees of freedom, and a maximal tree depth at 2. We repeat 100 times and plot the average fits against the true densities.

The covariate generating mechanism is the same as in Figure 2. We plot the average estimated conditional densities plus and minus one standard deviation. Though both LinCDE trees and LinCDE boosting produce good fits in all settings, the estimation bands of LinCDE

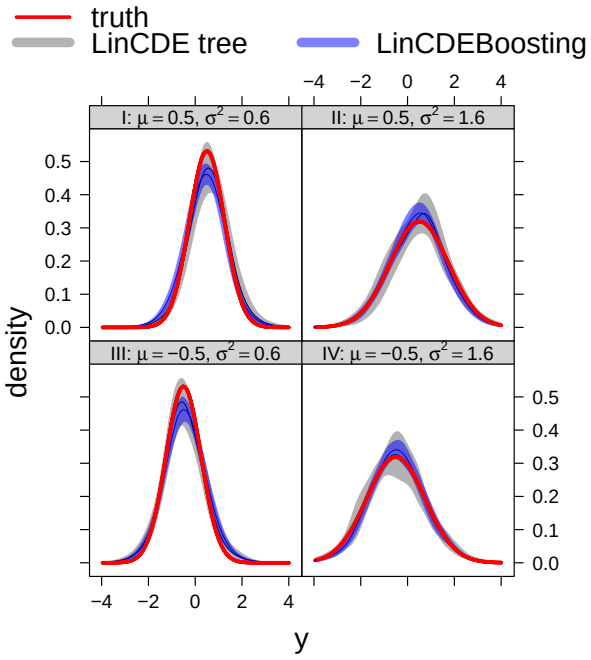


Figure 3: Comparison of LinCDE trees and LinCDE boosting. The responses are generated from (16). We pick 4 landmarks corresponding to different conditional means and variances. From top to bottom, the conditional means decrease from 0.5 to -0.5. From left to right, the conditional variances increase from 0.6 to 1.6. In each trial, we sample 400 observations from the target distribution. We repeat each setting 100 times, and plot the average estimated conditional densities plus and minus one standard deviation against the true densities.

boosting are always narrower by around a half. The observation implies LinCDE boosting is more stable than LinCDE trees.

5.1 Additive Model in the Natural Parameter Scale

In LinCDE boosting, we build an additive model in the natural parameter scale of the density (11). We find a sequence of LinCDE tree-based learners with parameter functions $\{\gamma^t(x)\}_{0\le t\le T-1}$, and aggregate those "basis" functions to obtain the final estimate[7]

$$\beta^T(x) = \sum_{t=0}^{T-1} \gamma^t(x). \tag{17}$$

In other words, at the t-th iteration, we tilt the current conditional density estimate

$$f_{y|x}^{t+1}(y \mid x) = f_{y|x}^{t}(y \mid x) \cdot e^{z(y)^\top \gamma^t(x) - \phi_{\beta^t(x)}(\gamma^t(x))} \tag{18}$$

7. To stabilize the performance, we may shrink $\gamma^t(x)$ by some learning rate $\eta \in (0, 1]$, and let $\beta^T(x) = \sum_{t=0}^{T-1} \eta\gamma^t(x)$.

based on knowledge $\gamma^t(x)$ learned by the new tree. Here $\phi_{\beta^t(x)}(\gamma^t(x)) = \psi(\beta^t(x)+\gamma^t(x)) - \psi(\beta^t(x))$ is the updated normalizing function (depending on x). Appendix E shows that if the true conditional density is smooth, the approximation error of LinCDE boosting's function class (18) with splines $z(y)$ will vanish as the number of splines k increases.

We determine the LinCDE tree modifiers in (17) by log-likelihood maximization. We aim to find the modifier that produces the largest improvement in the objective $\ell(\mathcal{R}_0; \beta(x)+\gamma(x))$ defined as

$$\sum_{i=1}^{n} \Big(\log(f^t_{y|x}(y_i \mid x_i)) + z(y_i)^\top \gamma(x_i) - \phi_{\beta^t(x_i)}(\gamma(x_i)) \Big) + \lambda \sum_{i=1}^{n} \sum_{j=1}^{k} \omega_j \gamma_j^2(x_i). \tag{19}$$

Compared to the objective (12) of LinCDE trees, the only difference in (19) is the normalizing function $\phi_{\beta^t(x)}(\gamma^t(x))$. When $\beta^t(x)$ is a constant function, the normalizing functions of LinCDE trees and LinCDE boosting coincide. In Section 5.2 and 5.3, we demonstrate how LinCDE boosting's heterogeneous normalizing function complicates the fitting and splitting steps and propose corresponding solutions.

5.2 Fitting Step

For fitting, given a subregion $\mathcal{R}$, the problem (19) can not be solved by Lindsey's method as in LinCDE trees, because $\beta^t(x)$ could be non-constant in the subregion $\mathcal{R}$. Explicitly, instead of the single constraint in (5), we could have up to $n_{\mathcal{R}}$ constraints

$$\int \kappa(y) e^{z(y)^\top (\beta^t(x_i)+\gamma^t(x_i))+\beta_0(x_i)} dy = 1, \quad x_i \in \mathcal{R}. \tag{20}$$

As a result, the Lagrangian function (6) as well as subsequent discrete approximations for Lindsey's method are invalid.

Fortunately, we can solve the fitting problem iteratively (*Fitting (LinCDE boosting)* below). Define bin probabilities

$$\begin{aligned} p_b(\beta^t(x)) &:= \frac{\kappa(y_b) e^{z^\top(y_b)\beta^t(x)}}{\sum_{b'=1}^{B} \kappa(y_{b'}) e^{z^\top(y_{b'})\beta^t(x)}}, \\ \bar{p}_b(\mathcal{R}; \beta^t(x)) &:= \frac{1}{n_{\mathcal{R}}} \sum_{x_i \in \mathcal{R}} \bar{p}_b(\beta^t(x_i)). \end{aligned} \tag{21}$$

We feed the marginal cell probabilities $\bar{p}_b(\mathcal{R}; \beta^t(x))$ to the fitting step as the baseline for modification. In Step 1, Lindsey's method produces a natural parameter modifier and a universal intercept for all samples in $\mathcal{R}$. The intercept produced by Lindsey's method guarantees that the marginal cell probabilities to sum to unity, but not for every individual x_i. In Step 2, we update the individual normalizing constants to ensure all constraints (20) are satisfied. In Proposition 4, we show that the fitting step of LinCDE boosting converges to the maximizer of the objective (19).

- *Fitting (LinCDE boosting).* In a region $\mathcal{R}$, initialize $\gamma = 0 \in \mathbb{R}^k$, $\gamma_0 = 0 \in \mathbb{R}^{n_{\mathcal{R}}}$. Count the number of observations $\{n_{\mathcal{R},b}\}$ in each bin.

1. *Updating* γ. Compute $\bar{p}_b(\mathcal{R}; \beta^t(x)+\gamma)$ in (21), fit a Poisson regression model with the response variable $\{n_{\mathcal{R},b}\}$, regressors $z(y_b)$, the offset $\log(\bar{p}_b(\mathcal{R}; \beta^t(x) + \gamma))$, and a weighted ridge penalty. Denote the estimated coefficients by $\Delta\gamma$. Update $\gamma \leftarrow \gamma + \Delta\gamma$.
2. *Updating* γ_0 *(normalization).* Compute the normalizing constants for all samples in $\mathcal{R}$

$$\gamma_{0,i} = -\log\left(\sum_b p_b(\beta^t(x_i))e^{z(y_b)^\top\gamma}\right).$$

3. Repeat steps 1 and 2 until $\|\Delta\gamma\|_2 \le \varepsilon$. Output γ, γ_0.

Proposition 4 *Assume that* $\lambda = 0$ *and* Y *is supported on the midpoints* $\{y_b\}$*, then the fitting step of LinCDE boosting converges, and the output* $\gamma^t_{\mathcal{R}}$ *satisfies*

$$\gamma^t_{\mathcal{R}} = \arg\max_\gamma \ell(\mathcal{R}; \beta^t(x) + \gamma).$$

We offer some intuition behind Proposition 4; i.e., why the fitting step of LinCDE boosting will stop at the likelihood maximizer. If $\beta^t(x) + \gamma$ is already optimal, then the average sufficient statistics $z(y)$ under marginal probabilities (21) should match the observations, which yields the KKT condition of the Poisson regression in Lindsey's method. As a result, Lindsey's method will produce zero updates, and the algorithm converges.

5.3 Splitting Step

Reminiscent of the splitting step for LinCDE trees, we seek the split that produces the largest improvement in the objective (19). Proposition 2 is not valid due to the heterogeneity in $\beta^t(x)$, and we propose an expanded version.

Proposition 5 *In a region* $\mathcal{R}$*, let* $n_{\mathcal{R}}$ *be the sample size and* $\gamma^t_{\mathcal{R}}$ *be the optimal update. Define the average sufficient statistics residuals as*

$$\bar{r}^t_{\mathcal{R}} := \frac{1}{n_{\mathcal{R}}}\sum_{x_i\in\mathcal{R}}\left(z_i - \nabla\psi(\beta^t(x_i) + \gamma^t_{\mathcal{R}})\right).$$

Given a candidate split s*, define*

$$\Psi^t_s(\gamma^t_{\mathcal{R}}) := \frac{n_{\mathcal{R}_{s,R}}}{n_{\mathcal{R}}}\left(\frac{1}{n_{\mathcal{R}_{s,L}}}\sum_{x_i\in\mathcal{R}_{s,L}}\nabla^2\psi\left(\beta^t(x_i) + \gamma^t_{\mathcal{R}}\right)\right)^{-1} + \frac{n_{\mathcal{R}_{s,L}}}{n_{\mathcal{R}}}\left(\frac{1}{n_{\mathcal{R}_{s,R}}}\sum_{x_i\in\mathcal{R}_{s,R}}\nabla^2\psi\left(\beta^t(x_i) + \gamma^t_{\mathcal{R}}\right)\right)^{-1}.$$

Then the improvement of the unpenalized conditional log-likelihood satisfies

$$\frac{1}{n_{\mathcal{R}}}\Delta\ell^t(\mathcal{R}, s) = \frac{n_{\mathcal{R}_{s,L}}n_{\mathcal{R}_{s,R}}}{2n^2_{\mathcal{R}}}\left(\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}_{s,R}}\right)^\top\Psi^t_s(\gamma^t_{\mathcal{R}})\left(\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}_{s,R}}\right) + r_s,$$

where $r_s = O(\|\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}}\|_2^3 + \|\bar{r}^t_{\mathcal{R}_{s,R}} - \bar{r}^t_{\mathcal{R}}\|_2^3)$.

The average sufficient-statistic residuals $\bar{r}^t_{\mathcal{R}}$ measures the deviation of the current estimator from the observations. By maximizing the quadratic approximation in Proposition 5, we find the candidate split s such that $\bar{r}^t_{\mathcal{R}_{s,L}}$ and $\bar{r}^t_{\mathcal{R}_{s,R}}$ are far apart, and modify the current estimator differently in the left and right children determined by the selected split. The updated splitting procedure is summarized in *Splitting (LinCDE boosting)*.

- *Splitting (LinCDE boosting).* In a region $\mathcal{R}$:
  1. Compute $\{n_{\mathcal{R}_{s,L}}, n_{\mathcal{R}_{s,R}}, \bar{z}_{\mathcal{R}_{s,L}}, \bar{z}_{\mathcal{R}_{s,R}}\}$ for each candidate split s, and approximate $\tilde{\Psi}^t(\gamma^t_{\mathcal{R}})$ in (22) by the average covariance matrix of $z(y)$ in $\mathcal{R}$.
  2. For each candidate split $s \in \mathcal{S}$, compute the quadratic approximation $\widehat{\Delta}\ell(\mathcal{R}, s)$ by Proposition 5, and choose the split $s_{\max} = \arg\max_{s\in\mathcal{S}} \hat{\Delta}\ell(\mathcal{R}, s)$.

The computation of the quadratic approximation is largely the same, except that the normalization matrix $\Psi^t_s(\gamma^t_{\mathcal{R}})$ varies across candidate splits and requires separate computation. To relieve the computational burden, we propose the following surrogate independent of candidate splits[8]

$$\tilde{\Psi}^t(\gamma^t_{\mathcal{R}}) = \left(\frac{1}{n_{\mathcal{R}}} \sum_{x_i \in \mathcal{R}} \nabla^2 \psi \left(\beta(x_i) + \gamma^t_{\mathcal{R}} \right) \right)^{-1}. \tag{22}$$

The surrogate $\tilde{\Psi}^t(\gamma^t_{\mathcal{R}})$ coincides with $\Psi^t_s(\gamma^t_{\mathcal{R}})$ if $\beta^t(x)$ is a constant vector, or the normalizing function $\psi(\beta)$ is quadratic.

Proposition 6 gives the computational time complexity of the splitting procedure (*Splitting (LinCDE boosting)*). The computation time scales linearly with regard to the sample size multiplied by dimension and the number of candidate splits. The extra computation compared to LinCDE trees comes from residual calculations and individual normalizations.

Proposition 6 *Assume that there are S candidate splits, then the splitting step for LinCDE boosting is of computational time complexity $\tilde{O}(dn_{\mathcal{R}}kB + n_{\mathcal{R}}k^2B + k^3 + Sk^2)$.*

6. Pretreatment

In this section, we discuss two pretreatments: response transformation and centering. The pretreatments are helpful when the response is heavy-tailed and when the conditional distributions $f_{y|x}(y \mid x)$ vary wildly in location.

6.1 Response Transformation

Heavy-tailed response distributions are common in practice, such as income and waiting time. If the response is heavy-tailed, then in Lindsey's method, most bins will be approximately empty. As a result the model tends to be over-parameterized and the estimates tend to overfit.

8. In practice, we add a universal diagonal matrix to $\frac{1}{n_{\mathcal{R}}} \sum_{x_i \in \mathcal{R}} \nabla^2 \psi(\beta(x_i) + \gamma^t_{\mathcal{R}})$ to stabilize the matrix inversion.
9. The initialization $\beta^0(x)$ is usually a constant vector, e.g., zero vector, independent of the covariates.

Algorithm 2: LinCDE boosting

Initialize the natural parameter function $\beta^0(x)$.[9]

for $t = 1{:}T$ **do**

1. Apply Algorithm 1 with *Fitting (LinCDE boosting)*, *Splitting (LinCDE boosting)*, and obtain the optimal LinCDE tree modifier $\hat{\gamma}^{t-1}(x)$.
2. Update

$$\hat{\beta}^t(x) \leftarrow \hat{\beta}^{t-1}(x) + \hat{\gamma}^{t-1}(x).$$

end

Output $\hat{\beta}^T(x)$.

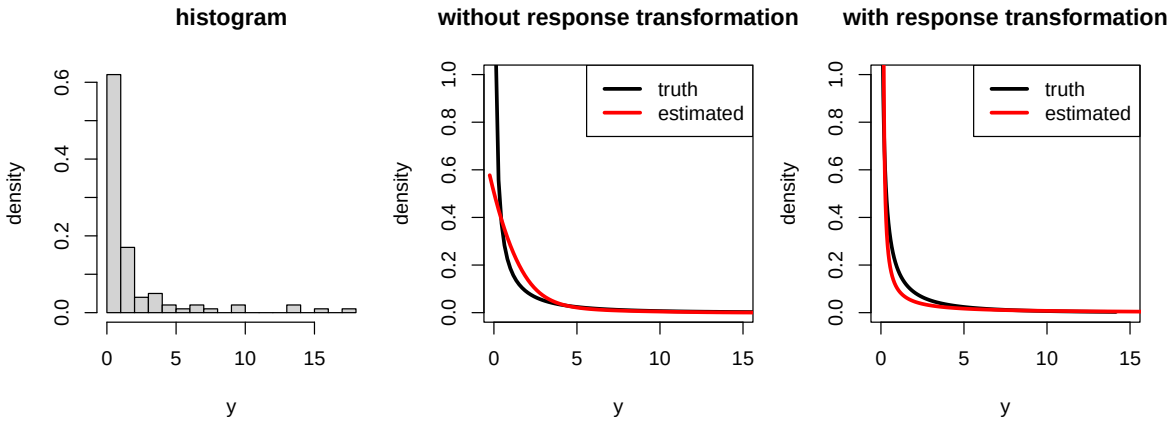


Figure 4: Response transformation and Lindsey's method. We generate 100 responses $y = w^2$, $w \sim \exp(1)$, and apply Lindsey's method. The left panel plots the histogram of the responses, and we observe several extreme values (≥ 10). The middle panel plots the true density and Lindsey's estimate. Due to a limited number (20) of bins and a wide response range caused by outliers, Lindsey's estimate is inaccurate (not sharp enough) around zero. The right panel demonstrates Lindsey's estimate using log-transformed responses. The log-transformation helps the estimation around zero and does not sacrifice the fit at the tail.

In response to the heavy-tailed responses, we recommend transforming the response first to be marginally normally distributed. Often log and cube-root transformations are useful. In a more principled way, we can apply the Box-Cox transformation to the responses and choose the optimal power parameter to produce the best approximation of a Gaussian distribution curve. Once the model is fit to the transformed data, we map the estimated conditional densities of the transformed responses back to those of the original observations. See Figure 4 for an example.

6.2 Centering

For a distribution whose conditional components differ wildly in location, LinCDE needs a large number of sufficient statistics to capture local distributional characteristics. For instance, Figure 5 displays a conditional Gaussian mixture with location shift

$$\begin{aligned} & y = 3x^{(1)} + wz^{(1)} + (1-w)z^{(2)}, \\ w \sim \text{Ber}(0.5), \quad & z^{(1)} \sim \mathcal{N}(-0.5, 0.06), \quad z^{(2)} \sim \mathcal{N}(0.5, 0.06), \quad z^{(1)} \perp\!\!\!\perp z^{(2)}. \end{aligned} \tag{23}$$

When we apply LinCDE boosting with $k = 10$ sufficient statistics, the estimates do not reproduce the bimodalities due to a lack of flexibility. We call this the "disjoint support" problem.

A straightforward solution to the disjoint support problem is to increase the number k so that the sufficient statistics $z(y)$ are adequately expressive. As a consequence, the number of components in the parameter function $\beta(x)$ goes up. This approach is prone to overfitting, especially when there are a small number (~ 20) samples in a terminal node. In addition, this approach will significantly slow down the splitting procedure, which scales $O(k^3)$ by Proposition 6.

Our solution is to *center* the response prior to fitting the LinCDE model. Since the difference in location causes the disjoint support problem, we suggest aligning the centers of the conditional densities in advance. Explicitly, we first estimate the locations via some conditional mean estimator and then subtract the estimates from the responses. The support of the residuals are less heterogeneous, and we apply LinCDE boosting to these residuals to capture additional distributional structures. Finally, we transform the resulting density estimates back to those of the responses. The procedure is summarized in Algorithm 3.

Algorithm 3: Centering

1. Estimate the conditional mean $\hat{h}(x)$ using the training data $\{(x_i, y_i)\}$. Compute the residuals $r_i = y_i - \hat{h}(x_i)$.
2. Apply LinCDE boosting to $\{(x_i, r_i)\}$, and obtain $\hat{f}_{R|X}(r \mid x)$.
3. Define $\hat{f}_{Y|X}(y \mid x) = \hat{f}_{R|X}(y - \hat{h}(x) \mid x)$ and output $\hat{f}_{Y|X}$.

Centering splits the task of conditional distribution estimation into conditional mean estimation and distributional property estimation. For centering we have available a variety of popular conditional mean estimators, such as the standard random forest, boosting, and neural networks. Once the data are centered, LinCDE boosting has a more manageable task. Figure 5 shows that with centering, LinCDE boosting is able to reproduce the bimodal structure in the above example with the same set of sufficient statistics.

7. Simulation

In this section, we demonstrate the efficacy of LinCDE boosting on simulated examples.

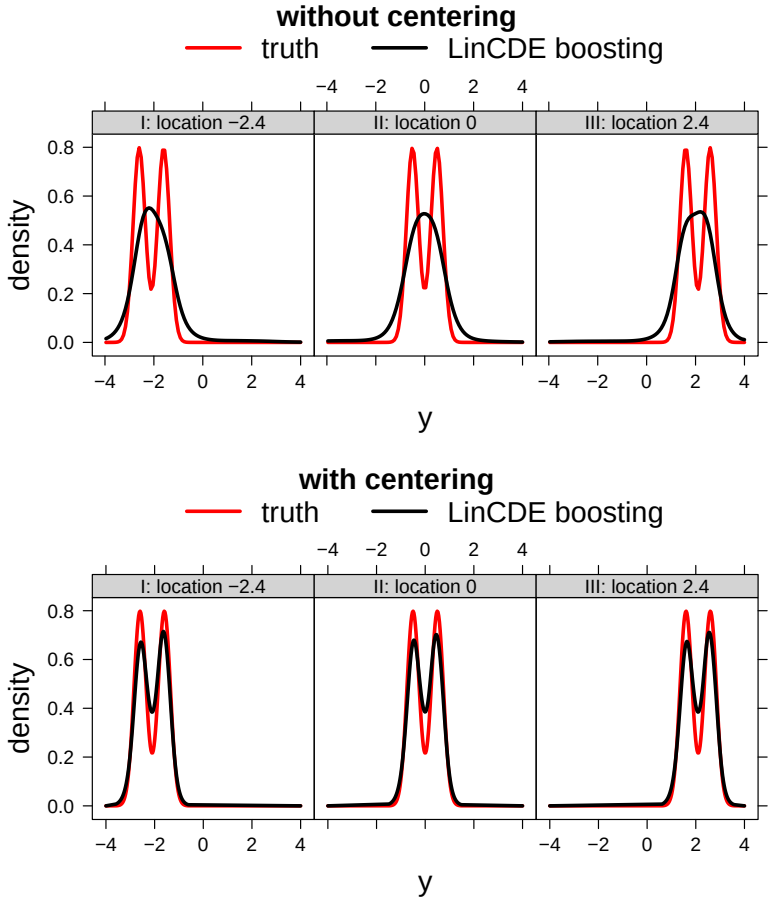


Figure 5: Conditional density estimation with and without centering. We consider the conditional density (23) and pick 3 landmarks corresponding to different locations. The first row plots LinCDE boosting's estimates without centering, and the second row plots the estimates augmented with true means. In each trial, we sample 1000 observations from the target distribution. We repeat each setting 100 times, and plot the average estimated conditional densities. In both settings, LinCDE boosting uses $k = 10$ sufficient statistics and 20 response bins.

7.1 Data and Methods

Consider $d = 20$ covariates randomly generated from uniform $[-1, 1]$. The responses given the covariates are sampled from the following distributions:

- *Locally Gaussian distribution (LGD)*:

$$Y \mid X = x \ \sim \ \mathcal{N}\left(0.5x^{(1)} + x^{(1)}x^{(2)}, \ \left(0.5 + 0.25x^{(2)}\right)^2\right).$$

At a covariate configuration, the response is Gaussian with the mean determined by $x^{(1)}$ and $x^{(2)}$, and the variance determined by $x^{(2)}$. Covariates $x^{(3)}$ to $x^{(20)}$ are nuisances variables;

- *Locally Gaussian or Gaussian mixture distribution (LGGMD)*:

$$Y \mid X = x \sim \begin{cases} 0.5\mathcal{N}\left(\mu(x^{(1)}) - 0.5,\ \sigma_+^2(x^{(3)})\right) & \\ \quad + 0.5\mathcal{N}\left(\mu(x^{(1)}) + 0.5,\ \sigma_-^2(x^{(3)})\right), & x^{(2)} \le 0.2, \\ \mathcal{N}\left(\mu(x^{(1)}),\ \sigma^2\right), & x^{(2)} > 0.2, \end{cases}$$

where the means and variances are

$$\mu\left(x^{(1)}\right) = 0.25x^{(1)},\ \sigma^2 = 0.3,$$
$$\sigma_+^2\left(x^{(3)}\right) = 0.25\left(0.25x^{(3)} + 0.5\right)^2,$$
$$\sigma_-^2\left(x^{(3)}\right) = 0.25\left(0.25x^{(3)} - 0.5\right)^2.$$

The mean is determined by $x^{(1)}$. The modality depends on $x^{(2)}$: in the subregion $x^{(2)} \ge 0.2$, the response follows a bimodal Gaussian mixture distribution, while in the complementary subregion, the response follows a unimodal Gaussian distribution. The skewness or symmetry is controlled by $x^{(3)}$ in the Gaussian mixture subregion: larger absolute values of $x^{(3)}$ imply higher asymmetry. Overall, the conditional distribution has location, shape, and symmetry dependent on the first three covariates. Covariates $x^{(4)}$ to $x^{(20)}$ are nuisance variables.

The training data set consists of 1000 i.i.d. samples. The performance is evaluated on an independent test data set of size 1000.

We compare LinCDE boosting with quantile regression forest and distribution boosting.[10] There are a number of tuning parameters in LinCDE boosting. The primary parameter is the number of trees (iteration number). Secondary tuning parameters include the tree size, the learning rate, and the ridge penalty parameter. On a separate validation data set, we experimented with a grid of secondary parameters, each associated with a sequence of iteration numbers, and select the best-performing configuration. By default, we use $k = 10$ transformed natural cubic splines and a Gaussian carrying density We use a small learning rate $\eta = 0.01$ to avoid overfitting. We use 40 discretization bins for training, and 20 or 50 for testing. The simulation examples do not have heavy-tail or disjoint support issues, and thus no pretreatments are needed.

7.2 Results of Conditional Density Estimation

Let the oracle be provided with the true density, and the null method estimates a marginal Gaussian distribution. We consider the following metric

$$\frac{\ell_{\text{method}} - \ell_{\text{null}}}{\ell_{\text{oracle}} - \ell_{\text{null}}}, \tag{24}$$

10. LinCDE boosting: `https://github.com/ZijunGao/LinCDE`. Quantile regression forest: R package *quantregForest* (Meinshausen, 2017). Distribution boosting: R package *conTree* (Friedman and Narasimhan, 2020).

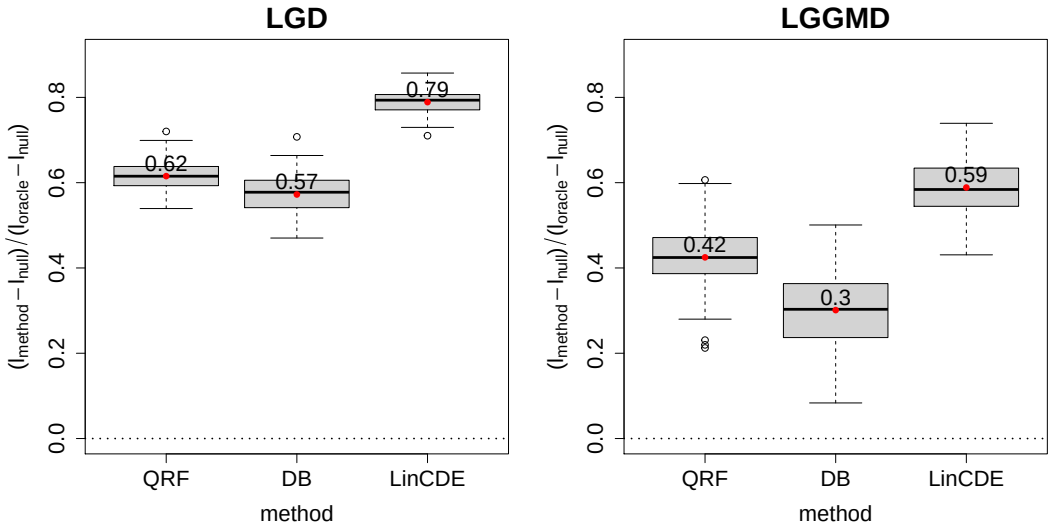


Figure 6: Box plots of goodness-of-fit measures (24) in the setting *LGD* (left panel) and the setting *LGGMD* (right panel). The goodness-of-fit measure is based on log-likelihoods, and a larger value indicates a better estimate. We compare quantile regression forests (QRF), distribution boosting (DB), and LinCDE boosting. Densities of quantile regression forests and distribution boosting are computed according to (25) with 20 bins.

where $\ell.$ denotes the test conditional log-likelihood of a specific method. The criterion is analogous to the goodness-of-fit measure R^2 of linear regression. It measures the performance of the method relative to the oracle; larger values indicate better fits, and the ideal value is one.

Quantile regression forests and distribution boosting estimate conditional quantiles instead of densities. To convert the quantile estimates to density estimates, we define a grid of bins with endpoints $y_{b,L}$ and $y_{b,R}$, and approximate the density in bin b by

$$\hat{f}_b = \frac{\hat{q}^{-1}(y_{b,R}) - \hat{q}^{-1}(y_{b,L})}{y_{b,R} - y_{b,L}}, \tag{25}$$

where $\hat{q}^{-1}(y)$ represents the inverse function of the quantile estimates. As the bin width shrinks, $\hat{f}_b$ is less biased but of larger variance. In simulations, we display the results with 20 bins and 50 bins (Appendix G). We observe that LinCDE boosting is robust to the bin size, while distribution boosting and quantile regression forests prefer 20 bins due to the smaller variances.

Figure 6 presents the goodness-of-fit measure (24) of the three methods under the *LGD* and *LGGMD* settings. In both settings, LinCDE boosting leads in performance, improving the null method by 60% to 80% of the oracle's improvements.

Figures 7 and 8 depict the estimated conditional densities of LinCDE boosting in different subregions. In both settings, LinCDE boosting identifies the roles of important covariates: in the *LGD* setting, the estimated conditional densities vary in location as $x^{(1)}$

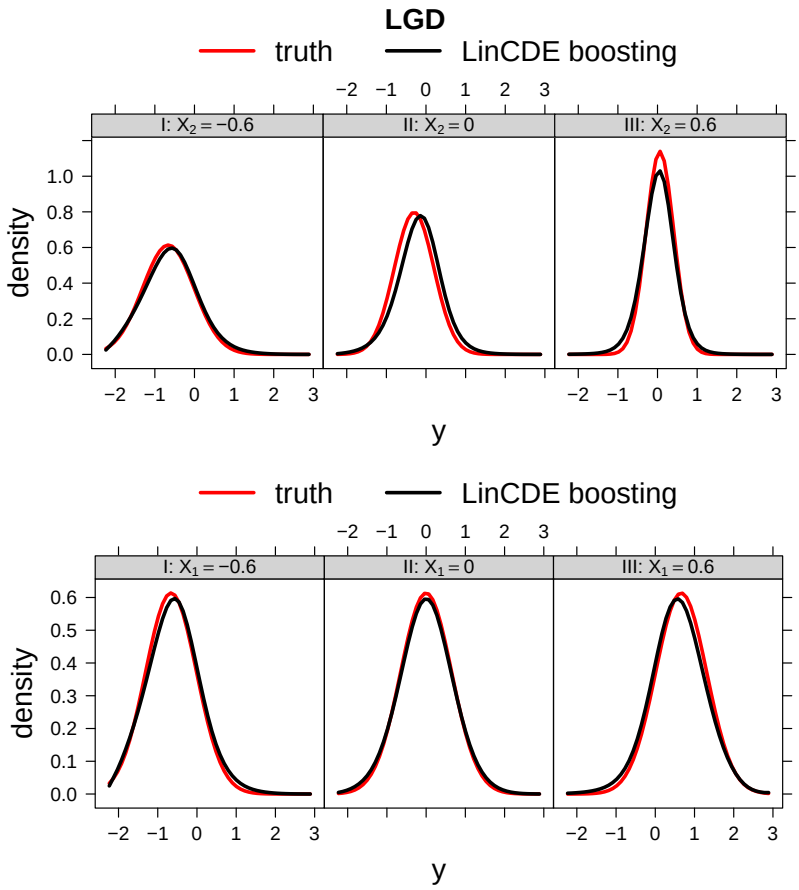


Figure 7: Conditional densities estimated by LinCDE boosting in the *LGD* setting. We take $x^{(2)} \in \{-0.6, 0, 0.6\}$ (upper panel) and $x^{(1)} \in \{-0.6, 0, 0.6\}$ (lower panel) and fix other covariates. The estimated conditional densities are close to the truth.

changes, and in scale as $x^{(2)}$ changes; in the *LGGMD* setting, the estimated conditional densities vary in location as $x^{(1)}$ changes, in shape as $x^{(2)}$ changes, and in symmetry as $x^{(3)}$ changes. To further illustrate the ability of LinCDE boosting to detect influential covariates, we present the importance scores in Figure 9. In the *LGD* setting, LinCDE boosting puts around 87% of the importance on $x^{(1)}$ and $x^{(2)}$, while quantile regression forest distributes more importance on the nuisances ($x^{(1)}$ and $x^{(2)}$ accounting for 40%). In the *LGGMD* setting, LinCDE boosting is able to detect all influential covariates $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, while the quantile regression forest only recognizes $x^{(1)}$.

7.3 Results of Conditional CDF Estimation

Here we evaluate the conditional CDF estimates of the three methods, bringing the comparisons closer to the home court of distribution boosting. We consider the average absolute

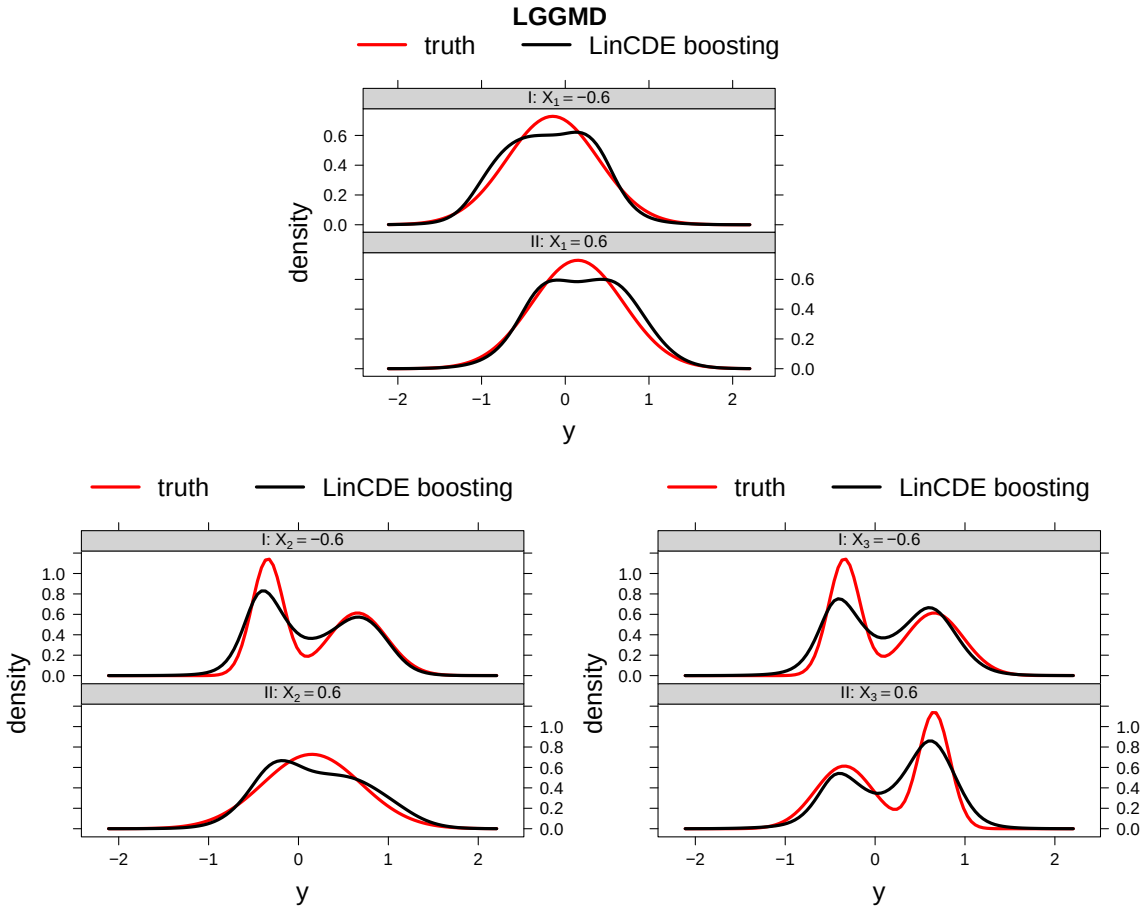


Figure 8: Conditional densities estimated by LinCDE boosting in the *LGGMD* setting. We take $x^{(1)}, x^{(2)}, x^{(3)} \in \{-0.6, 0.6\}$ respectively. The estimated conditional densities vary in location as $x^{(1)}$ changes, in shape as $x^{(2)}$ changes, and in symmetry as $x^{(3)}$ changes. The estimated conditional densities are close to the truth.

error (AAE) used by Friedman (2019)

$$\text{AAE} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left|\hat{F}(q(u_j \mid x_i) \mid x_i) - F(q(u_j \mid x_i) \mid x_i)\right|, \tag{26}$$

where $\{u_j\}$ is an evenly spaced grid on $[0, 1]$, and $q(u \mid x)$ denotes the u quantile at the covariate value x. To compute the CDF estimates, for distribution boosting and quantile regression forest, we directly invert the estimated quantiles to CDFs. For LinCDE boosting, we compute the multinomial cell probabilities with a fine grid (50 bins) and obtain the CDFs based on the cell probabilities.

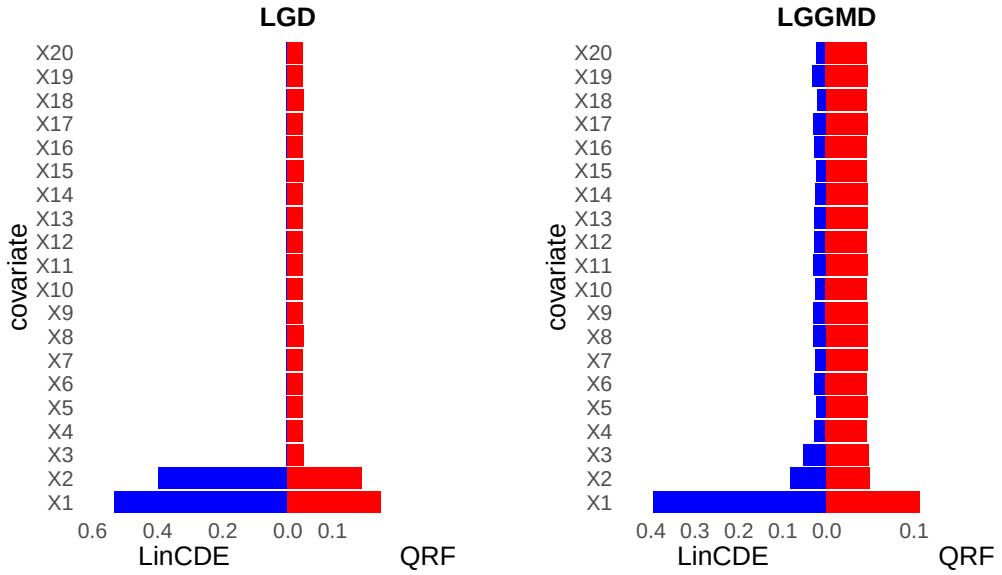


Figure 9: Importance scores of LinCDE boosting and quantile regression forest in the *LGD* (left panel) and *LGGMD* (right panel) settings. We normalize the importance scores to sum to one. In the *LGD* setting, both methods detect all the influential covariates $x^{(1)}$ and $x^{(2)}$. In the *LGGMD* setting, LinCDE boosting identifies all the influential covariates $x^{(1)}$, $x^{(2))}$, and $x^{(3)}$, while quantile regression forest only identifies $x^{(1)}$.

Figure 10 depicts the AAE metrics. In both settings, LinCDE boosting produces the smallest AAE. Notice that

$$\begin{aligned}\hat{F}(q(u_j) \mid x_i) - u_j &= \hat{F}(q(u_j \mid x_i) \mid x_i) - \hat{F}(\hat{q}(u_j \mid x_i) \mid x_i) \\ &\approx \hat{f}(q(u_j \mid x_i) \mid x_i) \cdot (q(u_j \mid x_i) - \hat{q}(u_j \mid x_i)).\end{aligned}$$

Though distribution boosting and quantile regression forest estimate the quantiles well, the CDF estimates can be harmed by the implicit density estimator multiplied. In Appendix G, we also compare the CDF estimates using Cramér-von Mises distance and observe consistent patterns to what we see here.

7.4 Results of Conditional Quantile Estimation

Here the comparisons are in the home court of quantile random forests. We evaluate the conditional quantile estimates of the three methods. We compute the quantile losses at $\{5\%, 25\%, 50\%, 75\%, 95\%\}$ levels (Table 1). For LinCDE boosting, we compute the multinomial cell probabilities (50 bins) and obtain the quantiles based on the cell probabilities. Despite the fact that quantile-based metrics should favor quantile-based methods, we observe that the performance of LinCDE boosting is similar.

We also construct 50% and 90% prediction intervals based on the quantiles. In Table 1, we display the coverages and widths of the 90% prediction intervals. All methods produce

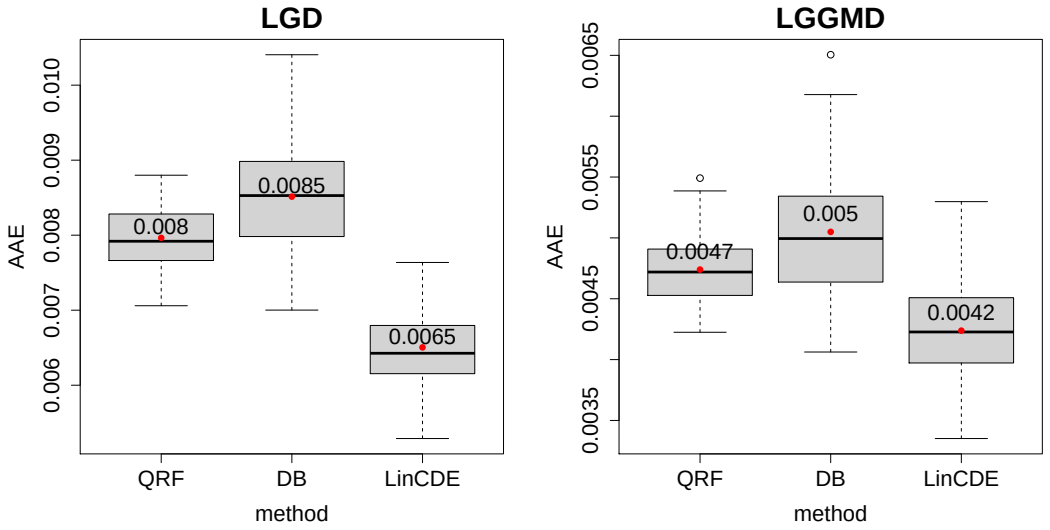


Figure 10: Box plots of AAE (26) in the LGD (left panel) and *LGGMD* (right panel) settings. AAE is a metric naturally defined for conditional CDF estimates (Friedman, 2019), and a smaller value indicates a better estimate.

data	method	quantile loss					coverage	width
		5 %	25 %	50 %	75 %	95 %	90% PI	90% PI
LGD	QRF	0.058 (0.001)	0.174 (0.02)	0.218 (0.002)	0.174 (0.02)	0.056 (0.001)	93.3% (0.9%)	2.02 (0.048)
	DB	0.058 (0.001)	0.176 (0.03)	0.218 (0.003)	0.174 (0.03)	0.057 (0.002)	92.5% (1.2%)	1.95 (0.051)
	LinCDE	0.055 (0.001)	0.168 (0.01)	0.212 (0.001)	0.169 (0.02)	0.054 (0.001)	91.9% (1.0%)	1.84 (0.045)
LGGMD	QRF	0.054 (0.001)	0.180 (0.001)	0.246 (0.002)	0.181 (0.001)	0.055 (0.001)	89.5% (0.7%)	1.76 (0.034)
	DB	0.054 (0.001)	0.182 (0.002)	0.246 (0.001)	0.182 (0.002)	0.055 (0.001)	91.2% (0.9%)	1.88 (0.038)
	LinCDE	0.053 (0.001)	0.181 (0.001)	0.246 (0.001)	0.181 (0.001)	0.055 (0.001)	90.5% (0.7%)	1.78 (0.032)

Table 1: Table of quantile losses (the smaller the better) at $\{5\%, 25\%, 50\%, 75\%, 95\%\}$ levels and 90% prediction interval coverages (ideally 90%), interval widths (the narrower the better) in the *LGD* and *LGGMD* settings. Standard deviations are in the parentheses. PI stands for prediction interval.

fairly good coverages. Results of 50% prediction intervals are consistent and can be found in Appendix G.

7.5 Computation Time

We compare the computation time of the three methods.[11] We normalize the training time by the number of trees. In Table 2, quantile random forest is the fastest, followed by LinCDE boosting. LinCDE boosting takes about 5 seconds for 100 iterations.

time(s)	QRF	DB	LinCDE
LGD	1.3 (0.018)	13 (0.21)	4.8 (0.36)
LGGMD	1.4 (0.15)	14 (1.0)	5.1 (0.93)

Table 2: Table of computation time in seconds. We present the training time for $n = 1000$ samples and $d = 20$ features per 100 trees. Standard deviations are in parentheses.

8. Real Data Analysis

In this section, we analyze real data sets with LinCDE boosting. The pipeline is as follows: first, we split the samples into training and test data sets; next, we perform 5-fold cross-validation on the training data set to select the hyper-parameters; finally, we apply the estimators with the selected hyper-parameters and evaluate multiple criteria on the test data set. We repeat the procedure 20 times and average the results. As for the centering, we use random forests as the conditional mean learner.

8.1 Old Faithful Geyser Data

The Old Faithful Geyser data records the eruptions from the "Old Faithful" geyser in the Yellowstone National Park (Azzalini and Bowman, 1990) and represents continuous measurement from August 1 to August 15, 1985. The data consists of 299 observations and 2 variables: eruption time and waiting time for the eruption. We estimate the conditional distribution of the eruption time given the waiting time.

In Figure 11, we plot the eruption time versus the waiting time. There is a clear cutoff at 70min: for any waiting time over 70min, the distribution of eruption time is bimodal, while for any waiting time below 70min, the distribution is unimodal. In Figure 11, we display the estimated conditional densities of LinCDE boosting at waiting time 85min and 60min. LinCDE boosting is capable of detecting the difference in modality. In Table 3, we summarize the comparison between LinCDE boosting, distribution boosting, and quantile regression forest regarding the negative log-likelihoods and quantile losses. LinCDE boosting outperforms in log-likelihood, and is competitive in quantile losses. We remark the AAE and Cramér-von Mises distance can not be computed since we do not have the true conditional distributions.

8.2 Human Height Data

The human height data is taken from the NHANES data set: a series of health and nutrition surveys collected by the US National Center for Health Statistics (NCHS). We estimate the

11. The experiments are run on a personal computer with a dual-core CPU and 8GB memory.

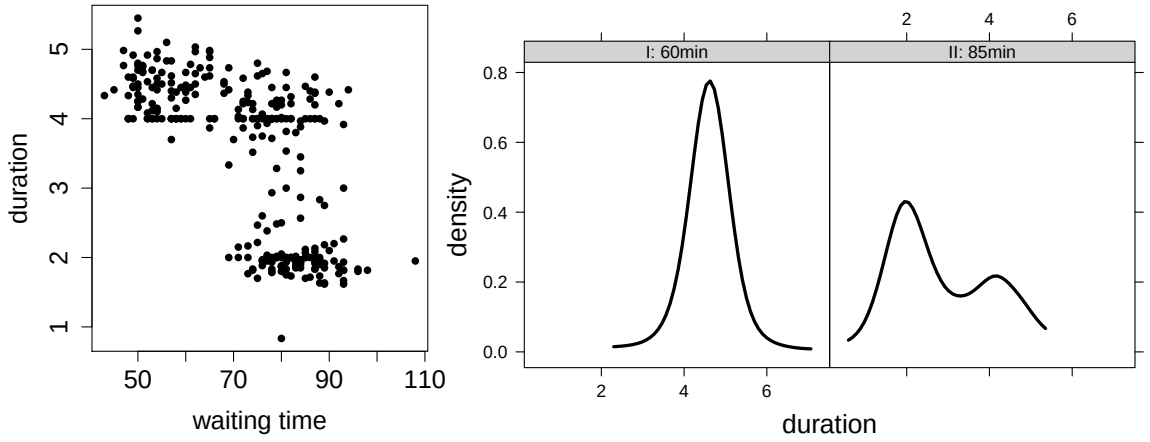


Figure 11: LinCDE boosting applied to the Old Faithful Geyser data. On the left, we plot the eruption time versus the waiting time. On the right, we display the estimated conditional densities by LinCDE boosting at waiting times 60min and 85min.

conditional distribution of the standing height. We consider two subsets: 542 samples in the age range 14 to 17, and 1956 samples in the age range 14 to 40. In the smaller subset, we only consider two covariates: age and poverty; in the larger subset, we consider 9 covariates, including age, poverty, race, gender, etc. In the smaller subset, we tune by cross-validation; in the larger subset, we split the data set for validation, training, and test (proportion $2:1:1$), and tune on the hold-out validation data.

The distribution of heights combining male and female is used as a typical illustration of bimodality (Devore and Peck, 2005). However, Schilling et al. (2002) point out the separation between the heights of men and women is not large enough to produce the bimodality. In Figure 12, we demonstrate the histogram of heights of white teenagers in the age range 15-19. The distribution of the combined data is sightly bimodal. Boys' heights are larger and more concentrated. We also provide LinCDE boosting's conditional density estimates obtained from the larger data set. Overall, the estimates accord with the histograms. The estimates without the covariate gender is on the borderline of unimodal and bimodal. The estimates with the gender explains the formation of the quasi-bimodality.

The comparisons of log-likelihood and quantile losses are summarized in Table 3. In both the larger and the smaller data sets, LinCDE boosting performs the best concerning the log-likelihood. The advantage is more significant in the smaller data set. The reason is that in the larger data set, there are more covariates, and the conditional mean explains a larger proportion of the variation in response, while in the smaller data set, the conditional distribution after the centering contains more information, such as the bimodality, which can be learnt by LinCDE boosting.

data	method	-log-like	quantile loss				
			5%	25 %	50%	75 %	95 %
Geyser	QRF	1.55 (0.12)	0.09 (0.01)	0.30 (0.02)	0.42 (0.03)	0.33 (0.02)	0.10 (0.01)
	DB	1.28 (0.10)	0.09 (0.01)	0.27 (0.02)	0.37 (0.03)	0.30 (0.02)	0.09 (0.01)
	LinCDE	**1.16** (0.07)	0.09 (0.01)	0.28 (0.02)	0.37 (0.03)	0.30 (0.02)	0.09 (0.01)
Height (age 14 - 40)	QRF	3.30 (0.03)	0.63 (0.03)	1.72 (0.06)	2.19 (0.07)	1.77 (0.07)	0.69 (0.04)
	DB	3.29 (0.04)	0.65 (0.03)	1.84 (0.05)	2.24 (0.06)	1.90 (0.07)	0.72 (0.05)
	LinCDE	**3.19** (0.03)	0.64 (0.04)	1.82 (0.05)	2.20 (0.06)	1.88 (0.07)	0.71 (0.04)
Height (age 14 - 17)	QRF	3.93 (0.17)	0.95 (0.12)	2.99 (0.24)	3.87 (0.25)	3.19 (0.23)	1.12 (0.20)
	DB	4.21 (0.16)	1.04 (0.04)	3.62 (0.19)	4.90 (0.28)	4.39 (0.38)	1.77 (0.32)
	LinCDE	**3.61** (0.06)	0.84 (0.05)	2.92 (0.16)	3.79 (0.23)	3.05 (0.19)	0.96 (0.10)
Bone density	QRF	-1.67 (0.11)	0.004 (0.001)	0.012 (0.001)	0.015 (0.001)	0.013 (0.001)	0.004 (0.001)
	DB	-1.49 (0.03)	0.007 (0.000)	0.015 (0.001)	0.014 (0.001)	0.016 (0.001)	0.007 (0.000)
	LinCDE	**-1.89** (0.06)	0.004 (0.000)	0.011 (0.001)	0.014 (0.001)	0.013 (0.001)	0.005 (0.001)

Table 3: Comparison of LinCDE boosting, QRF, and DB on real data sets. We display negative log-likelihoods and quantile losses at $\{5\%, 25\%, 50\%, 75\%, 95\%\}$ levels. (For all metrics, the smaller the better.) Standard deviations are in the parentheses.

8.3 Relative Spinal Bone Mineral Density Data

The relative spinal bone mineral density (spnbmd) data contains 485 observations on 261 North American adolescents. The response is the difference in spnbmd taken on two consecutive visits divided by the average. There are three covariates: sex, race, and age (the average age over the two visits). We estimate the conditional distributions of the spnbmd. The comparisons of log-likelihood and quantile losses are summarized in Table 3. LinCDE boosting performs the best concerning the log-likelihood.

The scatterplot of spnbmd versus age demonstrates serious heteroscedasticity: the variances reach the climax at age 12 and decrease afterward. In Figure 13, we plot LinCDE boosting's estimates at age 12, 15 and 20 of white females. The spreads of the conditional densities decrease as the age grows. At age 12, the spnmd distribution is right-skewed, while those at age 15 and 20 are approximately symmetric.

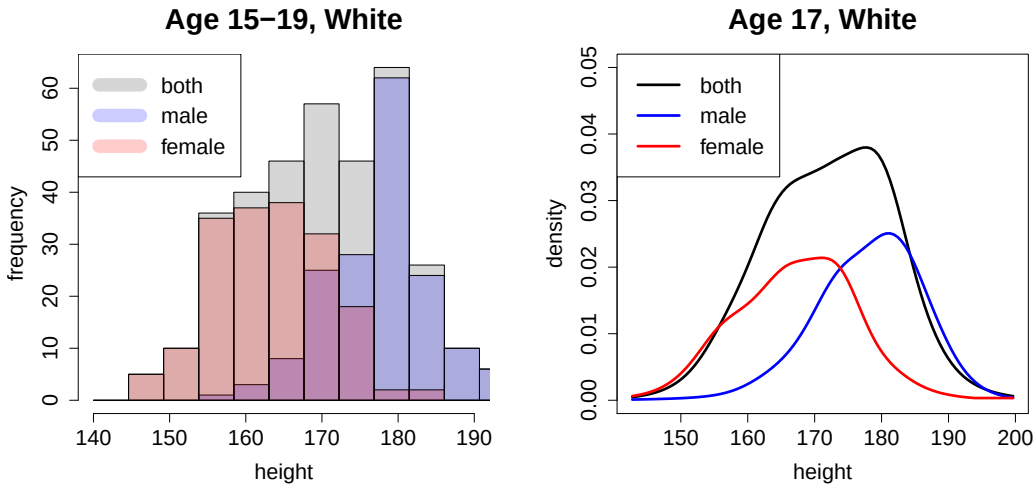


Figure 12: LinCDE boosting applied to the height data. On the left, we plot the histogram of heights of white teenagers in the age range 15-19. On the right, we plot the average estimated density by LinCDE boosting from the larger data set at age 17, race white, and other covariates fixed at the corresponding medians. The estimated conditional density without the gender is on the borderline between unimodal and bimodal. Furthermore, we contrast the histograms of male and female heights against that of the combined data in the left plot, explaining the formation of the quasi-bimodality. We also depict the estimated densities of females, males (right panel), which accord with the histograms (left panel).

9. Discussion

In this paper, we propose LinCDE boosting for conditional density estimation. LinCDE boosting poses no specific parametric assumptions of the density family. The estimates reflect a variety of distributional characteristics. In the presence of unrelated nuisance covariates, LinCDE boosting is able to focus on the influential ones.

So far, we have discussed only univariate responses. Multivariate responses emerge in multiple practical scenarios: locations on a 2D surface, joint distributions of health indices, to name a few. Lindsey's method and thus LinCDE boosting can be easily generalized to multivariate responses. Assume the responses are p-dimensional, multivariate LinCDE boosting considers the density (11) with sufficient statistics involving $y^{(1)}$ to $y^{(p)}$. As an illustrative example, if we use sufficient statistics $\{y^{(i)}\}$ and $\{y^{(i)}y^{(j)}\}$, the resulting density will be a multivariate Gaussian (see Efron and Hastie, 2016, chap. 8.3 for the galaxy data example). The response discretization now divides the hyper-rectangle response support into equal-sized p-dimensional bins and the rest of LinCDE boosting procedures carry over. The cost of multivariate responses is the exponentially growing number of bins and sufficient statistics, which requires more samples as well as computational power. In contrast, conditional quantiles for multi-dimensional responses are relatively less straightforward but several promising proposals have been proposed (Barnett, 1976; Chaudhuri, 1996; Kong

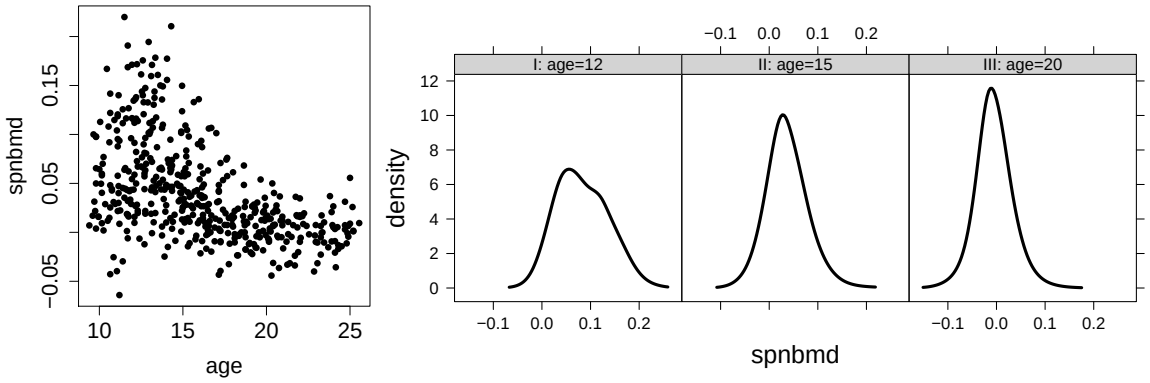


Figure 13: LinCDE boosting applied to the relative spinal bone mineral density data. On the left, we display the scatterplot of spnbmd versus age. There is serious heteroscedasticity. On the right, we plot LinCDE boosting's estimates at age 12, 15 and 20 of white females. The spreads of the conditional densities decrease as the age grows. At age 12, the spnbmd distribution is right-skewed, while those at age 15 and 20 are approximately symmetric.

and Mizera, 2012; Carlier et al., 2016). We refer readers to Koenker (2017) and references therein.

There are several exciting applications of LinCDE boosting.

- *Online learning.* Online learning processes the data that become available in a sequential order, such as stock prices and online auctions. As opposed to batch learning techniques which generate the best predictor by learning on the entire training data set once, online learning updates the best predictor for future data at each step. Online updating of LinCDE boosting is simple: we input the previous conditional density estimates as offsets, and modify them to fit new data.

- *Conditional density ratio estimation.* A stream of work studies the density ratio model (DRM), particularly the semi-parametric DRM. The density ratio can be used for importance sampling, two-sample testing, outlier detection (see Sugiyama et al., 2012, for an extensive review). Boosting tilts the baseline estimate parametrically based on the smaller group.

One future research direction is adding adaptive ridge penalty in LinCDE boosting. Recall the fits in the *LGGMD* setting (Figure 8) where the conditional densities change from a relatively smooth Gaussian density to a curvy bimodal Gaussian mixture, the estimates get stuck in between: in the smooth Gaussian subregions, the estimates produce unnecessary curvatures; in the bumpy Gaussian mixture subregions, the estimates are not sufficiently wavy. The lack-of-fit can be attributed to the universal constraint on the degrees of freedom: the constraint may be too stringent in some subregions while lenient in others.

10. Software

Software for LinCDE is made available as an R package at `https://github.com/ZijunGao/LinCDE`. The package can be installed from GitHub with

```
install.packages("devtools")
devtools::install_github("ZijunGao/LinCDE", build_vignettes = TRUE)
```

The package comes with a detailed vignette discussing hyper-parameter tuning and demonstrating a number of simulated and real data examples.

Acknowledgments

This research was partially supported by grants DMS-2013736 and IIS 1837931 from the National Science Foundation, and grant 5R01 EB 001988-21 from the National Institutes of Health.

Appendix A. Regularization Parameter and Degrees of Freedom

We discuss how the hyper-parameter λ relates to the degrees of freedom.[12] According to Eq. (12.74) in Efron and Hastie (2016), the effective number of parameters in Poisson models takes the form

$$\mathrm{df} = \sum_{b=1}^{B} \mathrm{cov}(\hat{\eta}_b, n_b), \tag{27}$$

where n_b are responses and $\hat{\eta}_b$ are estimates of the natural parameter. In the following Proposition 7, we obtain an approximation of Eq. (27) in the ridge Poisson regression via a quadratic expansion of the objective function.

Proposition 7 *Assume the responses n_b are generated independently from the Poisson model with conditional means $\mu_b^* = \kappa_b e^{z_b^\top \beta^*}$ (z includes the intercept) and the corresponding natural parameter $\eta_b^* = \log(\mu_b^*)$. Let $\lambda' > 0$ be the hyper-parameter of the ridge penalty.*[13]

- *For arbitrary β, the second order Taylor expansion at β^* of the Poisson log-likelihood with ridge penalty is*

$$\begin{aligned}
&\sum_{b=1}^{B} n_b \left(\log(\kappa_b) + z_b^\top \beta\right) - \kappa_b e^{z_b^\top \beta} - log(n_b!) - \lambda' \sum_{j=1}^{k} \omega_j \beta_j^2 \\
&\qquad = -\frac{1}{2}(Z\beta + K - \zeta)^\top W (Z\beta + K - \zeta) - \lambda' \beta^\top \Omega \beta + C, \\
&Z_{b\cdot} = z_b^\top, \quad K_b = \log(\kappa_b), \quad \zeta_b = \eta_b^* + \frac{n_b - \mu_b^*}{\mu_b^*}, \\
&W = \mathrm{diag}(\mu_1^*, \ldots, \mu_B^*), \quad \Omega = \mathrm{diag}(\omega_1, \ldots, \omega_k),
\end{aligned} \tag{28}$$

  for some constant $C > 0$ independent of β.

- *Let $\hat{\beta}$ be the minimizer of the quadratic approximation* (28)*, then*

$$df = \sum_{b=1}^{B} cov\left(\log(\kappa_b) + z_b^\top \hat{\beta}, y_b\right) = tr\left(\left(H_{\beta^*} + 2\lambda'\Omega\right)^{-1} H_{\beta^*}\right), \tag{29}$$

  where $H_{\beta^*} = Z^\top W Z$ is the Hessian matrix of the negative Poisson log-likelihood evaluated at β^*.

We prove Proposition 7 in Appendix F. As a corollary, if Ω is a scalar multiple of the identity matrix, Eq. (29) agrees with the degrees of freedom formula Eq. (7.34) in Hastie et al. (2009). In practice, β^* is unknown and we plug $\hat{\beta}$ in Eq. (29) to compute the number of effective parameters. We remark that Eq. (29) includes one degree of freedom for the intercept and in the LinCDE package we will use Eq. (29) minus one as the degrees of freedom.

12. The degrees of freedom are derived under the Poisson approximation (7) of the original likelihood.
13. $\lambda' = n\lambda$ for the λ in (8).

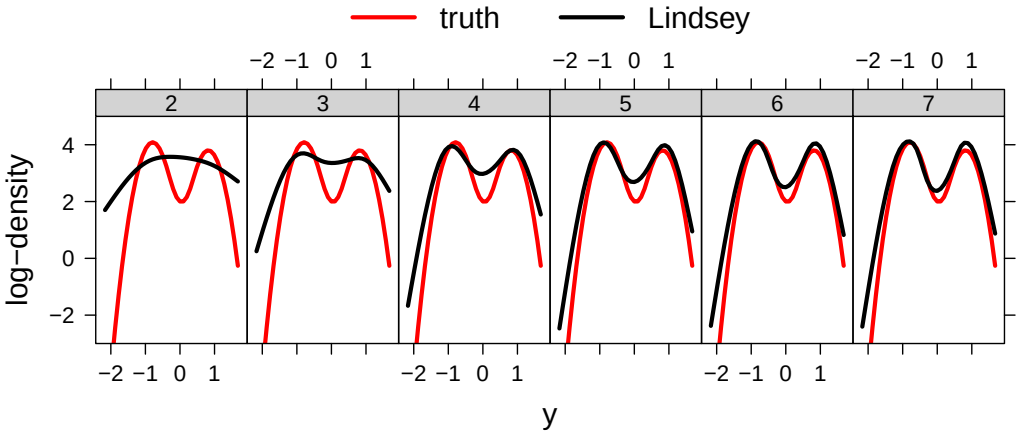


Figure 14: Estimated log-densities by Lindsey's method under different degrees of freedom. We use 10 transformed natural cubic splines as basis functions and increase the degrees of freedom in Lindsey's method from 2 to 7 (left to right).

Figure 14 plots the estimated log-densities of a Gaussian mixture distribution from Lindsey's method under different degrees of freedom. As the degrees of freedom (excluding the intercept) increase from 2 to 7, the log-density curves evolve from approximately quadratic to significantly bimodal. At 7 degrees of freedom, the estimated log-density is reasonably close to the underlying truth.

Appendix B. Linear Transformation for Basis Construction

We expand on the linear transformation used in the basis construction in Section 3. Let the eigen-decomposition of Ω be $UDU^\top$, where $U \in \mathbb{R}^{k\times k}$ is orthonormal, and $D \in \mathbb{R}^{k\times k}$ is a diagonal matrix with non-negative diagonal values ordered decreasingly. Define the linear-transformed cubic spline bases $\tilde{z}(y) = U^\top z(y)$ and the corresponding coefficients $\tilde{\beta} = U^\top \beta$. Then $\tilde{z}(y)^\top \tilde{\beta} = z(y)^\top \beta$ and $\tilde{\Omega} = U^\top \Omega U = U^\top U D U^\top U = D$. Therefore, $z(y) \mapsto U^\top z(y)$ is the desired linear transformation.

Appendix C. Approximation Performance of Proposition 2

In Figures 15 and 16, we empirically demonstrate the efficacy of the quadratic approximation suggested by Proposition 2 . We let the conditional densities depend solely on $x^{(1)}$ and jump (Figure 15) or vary continuously (Figure 16) in conditional variance, modality, or skewness. We observe that at candidate splits where the left child and the right child are similar, e.g., all the candidate splits based on $x^{(2)}$, the quadratic form and the exact log-likelihood difference are almost the same. At candidate splits where the left child and the right child are different, e.g., $x^{(1)} = 0$ in Figure 15, the exact log-likelihood difference is large, and the quadratic form is sufficiently close to the difference to determine the optimal split. We remark that to gain robustness, we set the quadratic approximation to zero if one of

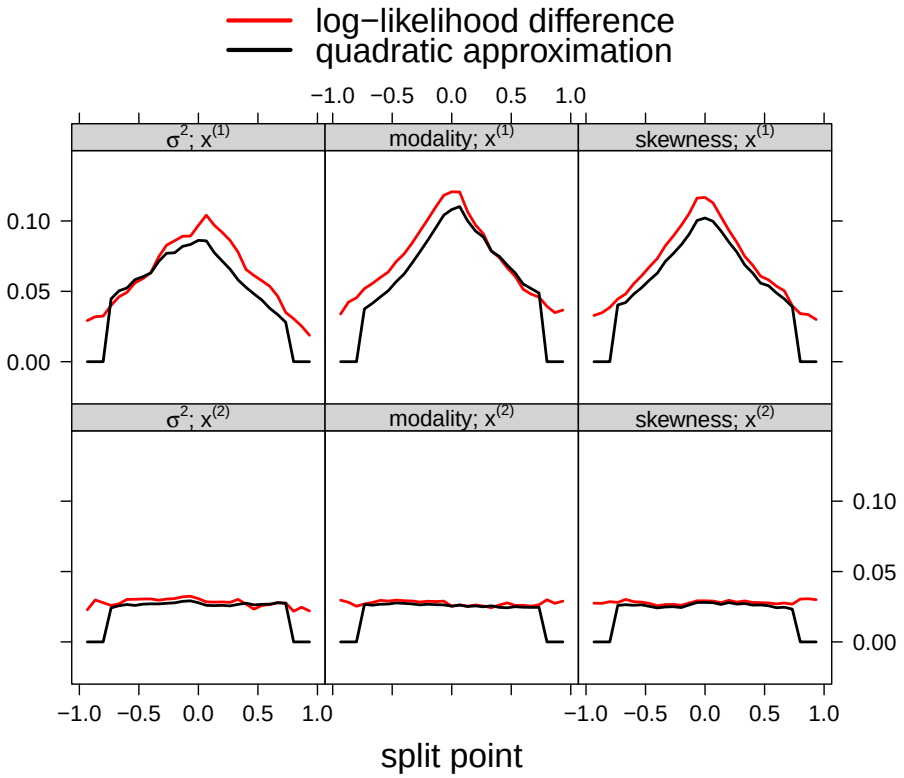
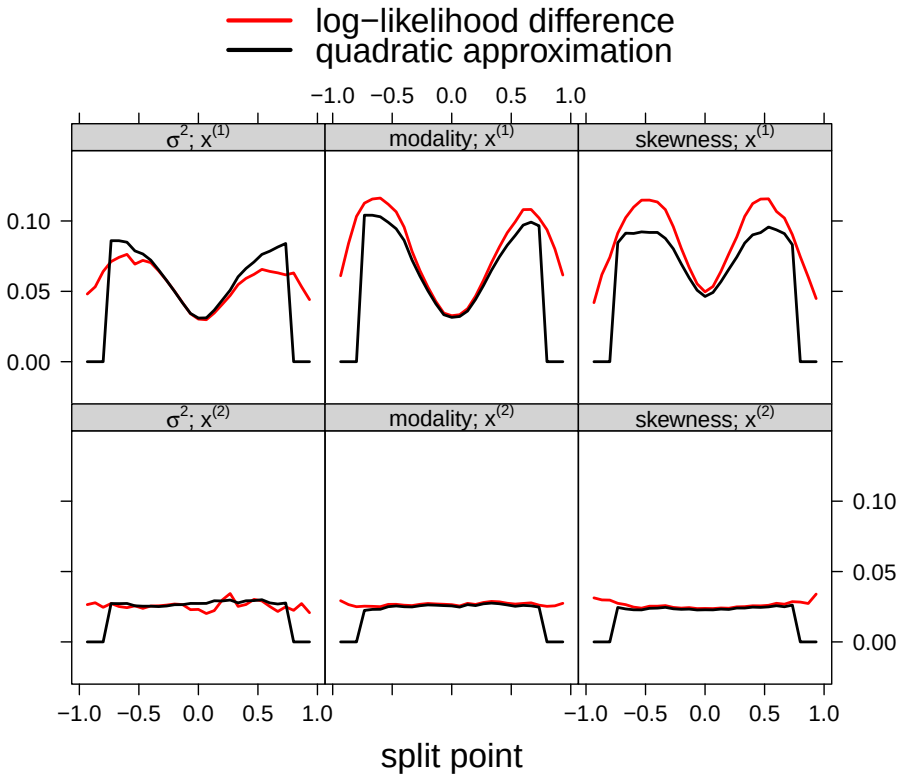


Figure 15: Comparison of the log-likelihood difference (13) and the quadratic approximation in Proposition 2. There are two subregions: $x^{(1)} \leq 0$ and $x^{(1)} > 0$. In different subregions, the conditional densities are different in conditional variance (σ^2), modality, or skewness (each column corresponds to a type of difference). In the same subregion, the conditional density does not change. In each trial, we sample 100 observations and 5 covariates. We use 5 transformed natural cubic splines as basis functions, and 30 candidate splits for each covariate equally spread across their ranges. We plot the log-likelihood difference (red) and the quadratic approximation (black) for candidate splits of $x^{(1)}$ (influential) and $x^{(2)}$ (nuisance), respectively. The results are aggregated over 100 times.

the candidate split's children contain less than 10 samples, which leads to the imperfect approximation at the boundary candidate splits.

Appendix D. Stopping Criteria for LinCDE Trees

We discuss stopping criteria for LinCDE trees. There is no universally optimal choice. If we build a single tree learner, the preferred strategy, at least for regression and classification trees according to Breiman et al. (1984), is to grow a large tree, then prune the tree using



Continuously varying conditional densities

Figure 16: Comparison of the log-likelihood difference (13) and the quadratic approximation in Proposition 2. The conditional densities vary continuously with $|x^{(1)}|$ in variance (σ^2), modality, or skewness (each column corresponds to a type of difference). In each trial, we sample 100 observations and 5 covariates. We use 5 transformed natural cubic splines as basis functions, and 30 candidate splits for each covariate equally spread across their ranges. We plot the log-likelihood difference (red) and the quadratic approximation (black) for candidate splits of $x^{(1)}$ (influential) and $x^{(2)}$ (nuisance), respectively. The results are aggregated over 100 times.

the cost-complexity pruning. If we train a random forest learner, then Breiman (2001) recommends stopping the splitting process only when some minimum node size, default to be 5 in the package *randomForest* (Liaw and Wiener, 2002), is reached. As for tree boosting, Friedman (2001) uses trees with the same number of terminal nodes. The number of terminal nodes is treated as a hyper-parameter of the boosting algorithm and tuned to maximize the performance on the data set at hand. The aforementioned stopping criteria generalize to LinCDE trees straightforwardly. Two options are currently available in codes: (1) stop when the tree depth reaches some prefixed level; (2) stop when the decrease in the objective fails to surpass a certain number—a greedy top-down approach. We don't

recommend the criterion of stopping until a terminal node is pure, because there will be insufficient samples at terminal nodes for density estimation.

Appendix E. Approximation Error for LinCDE Boosting

We characterize the expressiveness of LinCDE boosting's function class class (18) with splines $z(y)$.

Without loss of generality, we will assume $f_{Y|X}(y|x)$ is supported on $[0,1] \times [0,1]^d$. Let S_2^s be the Sobolev space of functions $h(y)$ on $[0,1]$ for which $h^{(s-1)}$ is absolutely continuous and $\int (h^{(s)}(y))^2 dy$ is finite. Denote the space of tree boosting models with arbitrary depth and number of trees by $\mathcal{N}_1 := \{\sum_{i=1}^N a_i \mathbb{1}_{x^{(j)} \in I_j}, N \geq 1, I_j \subseteq [0,1]\}$. Denote the space of splines of order ζ on $[0,1]$ with equal-spaced knots $\{1, \ldots, k-\zeta+1\} \times \delta$, $\delta = 1/(k-\zeta+2)$, $\zeta \leq k/2$, by $\Omega_{k,\zeta}$. Let $\mathcal{F}_k$ be the exponential family where $z_j(y) \in \Omega_{k,\zeta}$, $\zeta \geq s \geq 1$, and the parameter functions $\beta_j(x) \in \mathcal{N}_1$, $1 \leq j \leq k$.

Proposition 8 *Assume the log-conditional-density function $h(y|x) := \log(f_{Y|X}(y|x)) \in S_2^s$, $s \leq k/2$, and $\|h(\cdot|x)\|_\infty, \|h^{(s)}(\cdot|x)\|_2 \leq B$, then*

$$\min_{\tilde{f}_{Y|X} \in \mathcal{F}_k} \mathbb{E}\left[D\left(f_{Y|X}(y|X) \;||\; \tilde{f}_{Y|X}(y|X)\right)\right] \leq \frac{C}{k^{2s}}, \tag{30}$$

where $C > 0$ is a constant depending on B.

The proof of Proposition 8 can be found in Section F.

Appendix F. Proofs

In this section, we provide proofs of aforementioned propositions and claims.

F.1 Proof of Claim 1

Proof. For $u \in \mathbb{R}^k$, u lies in the null space of Ω_z if and only if $u^\top \Omega u = 0$. By the definition of Ω,

$$0 = u^\top \Omega u = \int (z'''(y)^\top u)^2 dy,$$

which implies that $(z(y)^\top u)''' = z'''(y)^\top u = 0$ almost everywhere. On one hand, if $z(y)^\top u$ is linear or quadratic, then $(z(y)^\top u)'''$ is automatically zero everywhere. On the other hand, since $z(y)$ are cubic spline bases, $z(y)^\top u$ is piece-wise cubic, and $(z(y)^\top u)''' = 0$ implies that $z(y)^\top u$ is piece-wise linear or quadratic. Because $z'(y)$ and $z''(y)$ are both continuous, $z(y)^\top u$ is also second-order continuous, $z(y)^\top u$ must be linear or quadratic in y. ■

F.2 Proof of Proposition 7

Proof. For arbitrary μ_b, η_b such that $\eta_b = \log(\mu_b)$, the second order Taylor expansion at μ_b^* of the Poisson log-likelihood of the b-th sample $\ell(y_b; \eta_b)$ is

$$\ell(y_b; \eta_b) \approx \ell(y_b; \eta_b^*) + \frac{\partial}{\partial \eta_b} \ell(y_b; \eta_b^*)(\eta_b - \eta_b^*) + \frac{1}{2} \frac{\partial^2}{\partial \eta_b^2} \ell(y_b; \eta_b^*)(\eta_b - \eta_b^*)^2. \tag{31}$$

By definition,

$$\ell(y_b;\eta_b^*) = y_b\eta_b^* - \kappa_b e^{\eta_b^*} + C, \quad \frac{\partial}{\partial\eta_b}\ell(y_b;\eta_b^*) = y_b - \kappa_b e^{\eta_b^*}, \quad \frac{\partial^2}{\partial\eta_b^2}\ell(y_b;\eta_b^*) = -\kappa_b e^{\eta_b^*}, \quad (32)$$

for some constant C independent of η_b^*. Plug Eq. (32) into Eq. (31),

$$\begin{aligned}\ell(y_b;\eta_b) &\approx y_b\eta_b^* - \kappa_b e^{\eta_b^*} + (y_b - \kappa_b e^{\eta_b^*})(\eta_b - \eta_b^*) - \frac{1}{2}\kappa_b e^{\eta_b^*}(\eta_b - \eta_b^*)^2 + C \\ &= -\frac{1}{2}\kappa_b e^{\eta_b^*}\left(\eta_b - \eta_b^* - \frac{y_b - \kappa_b e^{\eta_b^*}}{\kappa_b e^{\eta_b^*}}\right)^2 + r(y_b,\eta_b^*),\end{aligned} \quad (33)$$

where the remainder $r(y_b,\eta_b^*)$ is independent of η_b. Plug $\eta_b = \log(\kappa_b) + z_b^\top\beta$, $\mu_b^* = e^{\eta_b^*}$ in Eq. (33), we sum over all samples and finish the proof of Eq. (28).

We different the quadratic approximation (28) with respect to β and obtain the score function

$$-Z^\top W(Z\beta + K - \zeta) - 2\lambda'\Omega\beta = 0.$$

We solve the score function to obtain

$$\hat\beta = \left(Z^\top WZ + 2\lambda'\Omega\right)^{-1} Z^\top W(\zeta - K).$$

Then plug $\hat\eta = K + Z\hat\beta$ into Eq. (27) and we get

$$\begin{aligned}\mathrm{df} = \sum_{b=1}^B \mathrm{cov}(\hat\eta_b, y_b) &= \mathrm{tr}\left(\mathrm{cov}\left(Z\left(Z^\top WZ + 2\lambda'\Omega\right)^{-1} Z^\top W(\zeta - K), Y\right)\right) \\ &= \mathrm{tr}\left(Z\left(Z^\top WZ + 2\lambda'\Omega\right)^{-1} Z^\top W\mathrm{cov}(\zeta, Y)\right).\end{aligned}$$

Notice that $\mathrm{cov}(\zeta_b, y_{b'}) = \mathrm{cov}(y_b, y_{b'})/\mu_b^* = \mathbb{1}_{\{b=b'\}}$, then

$$\mathrm{df} = \mathrm{tr}\left(Z\left(Z^\top WZ + 2\lambda'\Omega\right)^{-1} Z^\top W\right) = \mathrm{tr}\left(\left(Z^\top WZ + 2\lambda'\Omega\right)^{-1} Z^\top WZ\right).$$

This gives Eq. (29). The Hessian of the negative Poisson log-likelihood at β^* is

$$\begin{aligned}H_{\beta^*} &= -\nabla_\beta^2|_{\beta=\beta^*}\sum_{b=1}^B \ell(y_b;\eta_b) = \sum_{b=1}^B \nabla_\beta^2|_{\beta=\beta^*}\kappa_b e^{z_b^\top\beta} \\ &= \sum_{b=1}^B \kappa_b e^{z_b^\top\beta^*} z_b z_b^\top = Z^\top WZ^\top.\end{aligned}$$

■

F.3 Proof of Proposition 2

Proof. We simplify the differences in the log-likelihood,

$$\begin{aligned}
\Delta\ell(\mathcal{R}, s) = &\sum_{x_i \in \mathcal{R}_L} \Big(\log(\kappa(y_i)) + z(y_i)^\top \hat{\beta}_{\mathcal{R}_{s,L}} - \psi(\hat{\beta}_{\mathcal{R}_{s,L}})\Big) \\
&+ \sum_{x_i \in \mathcal{R}_R} \Big(\log(\kappa(y_i)) + z(y_i)^\top \hat{\beta}_{\mathcal{R}_{s,R}} - \psi(\hat{\beta}_{\mathcal{R}_{s,R}})\Big) \\
&- \sum_{x_i \in \mathcal{R}} \Big(\log(\kappa(y_i)) + z(y_i)^\top \hat{\beta}_{\mathcal{R}} - \psi(\hat{\beta}_{\mathcal{R}})\Big) \\
&- \lambda n_{\mathcal{R}_{s,L}} \hat{\beta}_{\mathcal{R}_{s,L}}^\top \Omega \hat{\beta}_{\mathcal{R}_{s,L}} - \lambda n_{\mathcal{R}_{s,R}} \hat{\beta}_{\mathcal{R}_{s,R}}^\top \Omega \hat{\beta}_{\mathcal{R}_{s,R}} + \lambda n_{\mathcal{R}} \hat{\beta}_{\mathcal{R}}^\top \Omega \hat{\beta}_{\mathcal{R}} \\
= &n_{\mathcal{R}_{s,L}} \bar{z}_{\mathcal{R}_{s,L}}^\top (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) + n_{\mathcal{R}_{s,R}} \bar{z}_{\mathcal{R}_{s,R}}^\top (\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}) \\
&- n_{\mathcal{R}_{s,L}} (\psi(\hat{\beta}_{\mathcal{R}_{s,L}}) - \psi(\hat{\beta}_{\mathcal{R}})) - n_{\mathcal{R}_{s,R}} (\psi(\hat{\beta}_{\mathcal{R}_{s,R}}) - \psi(\hat{\beta}_{\mathcal{R}})) \\
&- \lambda n_{\mathcal{R}_{s,L}} \hat{\beta}_{\mathcal{R}_{s,L}}^\top \Omega \hat{\beta}_{\mathcal{R}_{s,L}} - \lambda n_{\mathcal{R}_{s,R}} \hat{\beta}_{\mathcal{R}_{s,R}}^\top \Omega \hat{\beta}_{\mathcal{R}_{s,R}} + \lambda n_{\mathcal{R}} \hat{\beta}_{\mathcal{R}}^\top \Omega \hat{\beta}_{\mathcal{R}},
\end{aligned} \tag{34}$$

where we use $n_{\mathcal{R}_{s,L}} \bar{z}_{\mathcal{R}_{s,L}} + n_{\mathcal{R}_{s,R}} \bar{z}_{\mathcal{R}_{s,R}} = n \bar{z}_{\mathcal{R}}$. The score equation of $\hat{\beta}_{\mathcal{R}}$ implies

$$\bar{z}_{\mathcal{R}} = \nabla\psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\hat{\beta}_{\mathcal{R}}. \tag{35}$$

and similarly for $\hat{\beta}_{\mathcal{R}_{s,L}}$, $\hat{\beta}_{\mathcal{R}_{s,R}}$. Plug Eq. (35) into Eq. (34), we obtain

$$\begin{aligned}
\Delta\ell(\mathcal{R}, s) = &n_{\mathcal{R}_{s,L}} \nabla\psi(\hat{\beta}_{\mathcal{R}_{s,L}})^\top (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) + n_{\mathcal{R}_{s,R}} \nabla\psi(\hat{\beta}_{\mathcal{R}_{s,R}})^\top (\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}) \\
&- n_{\mathcal{R}_{s,L}} (\psi(\hat{\beta}_{\mathcal{R}_{s,L}}) - \psi(\hat{\beta}_{\mathcal{R}})) - n_{\mathcal{R}_{s,R}} (\psi(\hat{\beta}_{\mathcal{R}_{s,R}}) - \psi(\hat{\beta}_{\mathcal{R}})) \\
&+ \lambda n_{\mathcal{R}_{s,L}} \left(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}\right)^\top \Omega \left(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}\right) \\
&+ \lambda n_{\mathcal{R}_{s,R}} \left(\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}\right)^\top \Omega \left(\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}\right).
\end{aligned} \tag{36}$$

By the Taylor expansion of $\psi(\beta)$ and $\nabla\psi(\beta)$ at $\hat{\beta}_{\mathcal{R}}$,

$$\begin{aligned}
&\nabla\psi(\hat{\beta}_{\mathcal{R}_{s,L}})^\top (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) - (\psi(\hat{\beta}_{\mathcal{R}_{s,L}}) - \psi(\hat{\beta}_{\mathcal{R}})) \\
= &\nabla\psi(\hat{\beta}_{\mathcal{R}_{s,L}})^\top (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) - \nabla\psi(\hat{\beta}_{\mathcal{R}})^\top (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) \\
&- \frac{1}{2} (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})^\top \nabla^2\psi(\hat{\beta}_{\mathcal{R}}) (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) + O(\|(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})\|_2^3) \\
= &\frac{1}{2} (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})^\top \nabla^2\psi(\hat{\beta}_{\mathcal{R}}) (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) + O(\|(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})\|_2^3).
\end{aligned} \tag{37}$$

Plug Eq. (37) into Eq. (36), we get

$$\begin{aligned}
\Delta\ell(\mathcal{R}, s) =& \frac{1}{2} n_{\mathcal{R}_{s,L}} (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})^\top \nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) \\
&+ \frac{1}{2} n_{\mathcal{R}_{s,R}} (\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}})^\top \nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) (\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}) \\
&+ \lambda n_{\mathcal{R}_{s,L}} \left(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}\right)^\top \Omega \left(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}\right) \\
&+ \lambda n_{\mathcal{R}_{s,R}} \left(\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}\right)^\top \Omega \left(\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}\right) \\
&+ O\left(\|(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})\|_2^3 + \|(\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}})\|_2^3\right) \\
=& \frac{1}{2} n_{\mathcal{R}_{s,L}} (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})^\top \left(\nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\right) (\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}}) \\
&+ \frac{1}{2} n_{\mathcal{R}_{s,R}} (\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}})^\top \left(\nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\right) (\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}}) \\
&+ O\left(\|(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})\|_2^3 + \|(\hat{\beta}_{\mathcal{R}_{s,R}} - \hat{\beta}_{\mathcal{R}})\|_2^3\right).
\end{aligned} \tag{38}$$

Finally by Eq. (35) and the assumption that $\nabla^2\psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega$ is invertible,

$$\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}} = \left(\nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} (\bar{z}_{\mathcal{R}_{z,L}} - \bar{z}_{\mathcal{R}}) + O\left(\|(\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}})\|_2^2\right). \tag{39}$$

Then $\hat{\beta}_{\mathcal{R}_{s,L}} - \hat{\beta}_{\mathcal{R}} \asymp \bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}}$ and similarly for the right child. Plug Eq. (39) into Eq. (38),

$$\begin{aligned}
\Delta\ell(\mathcal{R}, s) =& \frac{1}{2} n_{\mathcal{R}_{s,L}} (\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}})^\top \left(\nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} (\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}}) \\
&+ \frac{1}{2} n_{\mathcal{R}_{s,R}} (\bar{z}_{\mathcal{R}_{s,R}} - \bar{z}_{\mathcal{R}})^\top \left(\nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} (\bar{z}_{\mathcal{R}_{s,R}} - \bar{z}_{\mathcal{R}}) \\
&+ O\left(\|(\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}})\|_2^3 + \|(\bar{z}_{\mathcal{R}_{s,R}} - \bar{z}_{\mathcal{R}})\|_2^3\right) \\
=& \frac{n_{\mathcal{R}_{s,L}} n_{\mathcal{R}_{s,L}}}{2n_{\mathcal{R}}} (\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}_{s,R}})^\top \left(\nabla^2 \psi(\hat{\beta}_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} (\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}_{s,R}}) \\
&+ O\left(\|(\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}})\|_2^3 + \|(\bar{z}_{\mathcal{R}_{s,R}} - \bar{z}_{\mathcal{R}})\|_2^3\right),
\end{aligned}$$

and we finish the proof. ■

F.4 Proof of Claim 9

Claim 9 *Assume that in $\mathcal{R}$, y is supported on the midpoints $\{y_b\}$, then the covariance matrix approximation by Lindsey's method equals $\nabla^2\psi(\hat{\beta}_{\mathcal{R}})$.*

Proof. If y is supported on the midpoints $\{y_b\}$, then the discretization in Lindsey's method is accurate. As a result, the estimator of Lindsey's method is the exact log-likelihood maximizer $\hat{\beta}_{\mathcal{R}}$. Furthermore, the multinomial cell probabilities based on the Linsey's method's estimator is indeed the response distribution indexed by $\hat{\beta}_{\mathcal{R}}$. Next, by Lehmann and Romano (2006), $\nabla^2\psi(\hat{\beta}_{\mathcal{R}})$ equals the population covariance matrix of the sufficient statistics generated from the distribution indexed by $\hat{\beta}_{\mathcal{R}}$. Therefore, the covariance approximation by Lindsey's method equals $\nabla^2\psi(\hat{\beta}_{\mathcal{R}})$. ■

F.5 Proof of Proposition 3

Proof. We order the observations and count responses in each bin ($\tilde{O}(n_{\mathcal{R}})$). Next, we run Newton-Rapson algorithm. In each iteration, the computation of the gradient vector

$$\sum_{b=1}^{B} n_b \tilde{z}(y_b)(1 - e^{z(y_b)^\top \beta + \beta_0}).$$

is $O(kB)$, where $\tilde{z}^\top = [z^\top, 1]$. The computation of the Hessian matrix

$$-\sum_{b=1}^{B} n_b e^{z(y_b)^\top \beta + \beta_0} \tilde{z}(y_b)\tilde{z}(y_b)^\top$$

takes $O(k^2 B)$ operations. Finally, one Newton-Raphson update takes $O(k^3)$ operations. Newton-Raphson algorithm is superlinear (Boyd and Vandenberghe, 2004), thus we can regard the number of Newton-Raphson updates of constant order. The cell probabilities can be computed in $O(kB)$ time, and the covariance matrix takes $O(k^2 B)$ operations.

To compute the quadratic approximation, $\{n_{\mathcal{R}_{s,L}}, n_{\mathcal{R}_{s,R}}, \bar{z}_{\mathcal{R}_{s,L}}, \bar{z}_{\mathcal{R}_{s,R}}\}$ can be computed in $O(dn_{\mathcal{R}}k)$ by scanning through the samples once per coordinate. (If observations are not ordered beforehand, we will have $\tilde{O}(dn_{\mathcal{R}}k)$, where $\tilde{O}$ denotes the order up to log terms.) Adding the diagonal matrix Ω is cheap, and the matrix inversion takes $O(k^3)$ operations. For a candidate split, the quadratic term in Proposition 2 further takes $O(k^2)$ time. Choosing the optimal split takes $O(S)$. In summary, the complexity is $O(dn_{\mathcal{R}}k + k^2 B + k^3 + Sk^2)$. ■

F.6 Proof of Proposition 8

The results are built on the expressiveness of logspline density models and tree boosting models. We first develop an intermediate property for unconditional densities, and then combine it with the representability of tree boosting models with arbitrary depth and number of trees. We will use C to denote constants depending on B and the concrete values may vary from line to line.

For unconditional densities, let the true log-density be $h(y) \in S_2^s$ and $\|h^{(s)}(\cdot)\|_2 \le B$. By Barron and Sheu (1991), there exists a function $h_k \in \Omega_{k,\zeta}$ such that

$$\begin{aligned} \|h(\cdot) - h_k(\cdot)\|_2 &\le C\delta^s \|h^{(s)}(\cdot)\|_2 \le \frac{C}{k^s}, \\ \|h(\cdot) - h_k(\cdot)\|_\infty &\le C\delta^{s-1/2} \|h^{(s)}(\cdot)\|_2 \le \frac{C}{k^{s-1/2}}, \end{aligned} \tag{40}$$

where we use $\delta = 1/(k - \zeta + 2) \le 2/k$. Next, we turn to conditional densities. For $|h(y \mid x)| \in S_2^s$ and $\|h(\cdot \mid x)\|_\infty \le B$, by Barron and Sheu (1991, Lemma 3), there exists a unique vector solution $\beta(x)$ to the minimization problem

$$\beta(x) := \underset{\tilde{\beta}(x) \in \mathbb{R}^k}{\arg\min}\, D\left(f_{Y|X}(y \mid X = x) \parallel \kappa(y) e^{z(y)^\top \tilde{\beta}(x) - \psi(\tilde{\beta}(x))} \right).$$

We next show $\beta(x)$ is measurable on $[0,1]^d$. In fact, define the mapping

$$\Phi(\beta) := \mathbb{E}\left[z(Y)\kappa(Y) e^{z(Y)^\top \beta - \psi(\beta)} \right].$$

Again by Barron and Sheu (1991, Lemma 3), Φ is invertible and thus the inverse Φ^{-1} is differentiable. By the optimality of $\beta(x)$, we have

$$\Phi(\beta(x)) := \mathbb{E}\left[z(Y)\kappa(Y)e^{z(Y)^\top\beta(x)-\psi(\beta(x))}\right] = \mathbb{E}\left[z(Y)f_{Y|X}(Y \mid X = x)\right].$$

Notice that $\log(f_{Y|X}(Y \mid X = x)) \in S_2^s$, then $f_{Y|X}(Y \mid X = x)$ is at least $(s-1)$-th order smooth and as a result $\Phi(\beta(x))$ is differentiable with regard to x. Finally, since $\beta(x) = \Phi^{-1} \circ \Phi(\beta(x))$, we conclude $\beta(x)$ is differentiable thus measurable on $[0,1]^d$. Note that $\mathbb{E}\left[z(Y)f_{Y|X}(Y \mid X = x)\right]$ is bounded since $|h(y \mid x)| \le B$, then $\|\beta(x)\|_\infty \le C_\beta$ for some C_β depending on B.

Now we are ready to approximate $\beta(x)$ by tree boosting models. By the simple function approximation theorem, for any bounded measurable function on $[0,1]$, there exists a sequence of simple functions, i.e., linear combinations of indicator functions, such that converge to the target function point-wisely and and in L_1. Recall the tree boosting models $\{\sum_{i=1}^N a_i \mathbb{1}_{x^{(j)}\in I_j}, N \ge 1, I_j \subseteq [0,1]\}$, then given a measurable function $\beta_j(x)$ supported on the hyper-cube $[0,1]^d$, for any $\varepsilon > 0$, there exists a tree boosting model $\beta_{j,\varepsilon}(x) \in \mathcal{N}_1$ such that $|\beta_{j,\varepsilon}(x)| \le |\beta(x)| + \varepsilon$ on $[0,1]^d$ and

$$|\beta_j(x) - \beta_{j,\varepsilon}(x)| \le \varepsilon, \quad \forall x \in D_\varepsilon \subseteq [0,1]^d, \quad m(D_\varepsilon) \ge 1 - \varepsilon,$$

where D_ε is some measurable set and $m(D_\varepsilon)$ denotes the Lebesgue measure of the subset D_ε. By the triangle inequality and Cauchy-Schwarz inequality,

$$\begin{aligned}\left|h(y \mid x) - z(y)^\top\beta_\varepsilon(x)\right| &\le \left|h(y \mid x) - z(y)^\top\beta(x)\right| + \left|z(y)^\top\beta(x) - z(y)^\top\beta_\varepsilon(x)\right| \\ &\le \left|h(y \mid x) - z(y)^\top\beta(x)\right| + \|z(y)\|_2\|\beta(x) - \beta_\varepsilon(x)\|_2.\end{aligned}$$

On D_ε, $\|\beta(x) - \beta_\varepsilon(x)\|_2 \le \sqrt{k}\varepsilon$, and

$$\|h(\cdot \mid x) - z(\cdot)^\top\beta_\varepsilon(x)\|_2 \le C\left(\frac{1}{k^s} + \sqrt{k}\varepsilon\|z(\cdot)\|_2\right),$$

where we use (40). On D_ε^c, since $|h(y \mid x)| \le B$ and $|\beta_{j,\varepsilon}(x)| \le C_\beta + \varepsilon$, then

$$\|h(\cdot \mid x) - z(\cdot)^\top\beta_\varepsilon(x)\|_\infty \le \|h(\cdot \mid x)\|_\infty + \|z(\cdot)^\top\beta_\varepsilon(x)\|_\infty \le C.$$

Let $f_\varepsilon(y \mid x) := \kappa(y)e^{z(y)^\top\beta_\varepsilon(x)-\psi(\beta_\varepsilon(x))}$, then by Lemma 1 in Barron and Sheu (1991),

$$\begin{aligned}&\mathbb{E}\left[D\left(f(y \mid X) \parallel f_\varepsilon(y \mid X)\right)\right] \\ =\ &\mathbb{E}\left[D\left(f(y \mid X) \parallel f_\varepsilon(y \mid X)\right),\ X \in D_\varepsilon\right] \\ &+ \mathbb{E}\left[D\left(f(y \mid X) \parallel f_\varepsilon(y \mid X)\right),\ X \in D_\varepsilon^c\right] \\ \le\ &C\left(\frac{1}{k^{2s}} + k\varepsilon^2\|z(\cdot)\|_2^2 + \varepsilon\right).\end{aligned} \tag{41}$$

Since (41) is valid for arbitrary $\varepsilon > 0$, let $\varepsilon \to 0$ and we finish the proof. ■

F.7 Proof of Proposition 4

Proof. Without loss of generality, assume $\mathcal{R}$ is the covariate space. For simplicity, we omit the superscript of $\beta^t(x)$ and denote the natural parameter functions by $\beta(x)$.

We first show $\gamma^* = \arg\max_\gamma \ell(\mathcal{R}; \beta(x) + \gamma)$. For $\lambda = 0$, if $\Delta\gamma = 0$, the KKT condition of the last Poisson regression is

$$\frac{1}{n}\sum_{b=1}^{B} n_b z(y_b) = \sum_{b=1}^{B} \bar{p}_b(\mathcal{R}; \beta(x) + \gamma) z(y_b), \tag{42}$$

If in $\mathcal{R}$, Y is supported on the midpoints $\{y_b\}$,

$$\nabla\psi(\beta(x_i) + \gamma) = \frac{1}{B}\sum_{b=1}^{B} p_b(\beta(x_i) + \gamma) z(y_b). \tag{43}$$

Therefore, by Eq. (42) and Eq. (43),

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\nabla\psi(\beta(x_i)+\gamma) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{b=1}^{B} p_b(\beta(x_i)+\gamma)z(y_b) = \sum_{b=1}^{B} z(y_b)\cdot\frac{1}{n}\sum_{i=1}^{n} p_b(\beta(x_i)+\gamma)\\ &= \sum_{b=1}^{B}\bar{p}_b(\mathcal{R};\beta(x)+\gamma)z(y_b) = \frac{1}{n}\sum_{b=1}^{B} n_b z(y_b) = \frac{1}{n}\sum_{i=1}^{n} z(y_i),\end{aligned} \tag{44}$$

which implies the KKT condition of maximizing $\ell(\mathcal{R}; \beta(X) + \gamma)$ is satisfied at γ^*. Since the log-likelihood $\ell(\mathcal{R}; \beta(X) + \gamma)$ is strictly concave, then γ^* is the unique maximizer.

We next prove the algorithm converges. Plug in the MLE estimate of the intercept of the Poisson regression into the log-likelihood and we have

$$\begin{aligned}\ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma+\Delta\gamma) =& \ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma) + \sum_{b=1}^{B} n_b z(y_b)^\top\Delta\gamma\\ &- n\underbrace{\log\left(\sum_{b=1}^{B}\bar{p}_b(\mathcal{R};\beta(x)+\gamma)\exp\left\{z(y_b)^\top\Delta\gamma\right\}\right)}_{:=(I)}.\end{aligned} \tag{45}$$

By Jensen's inequality,

$$\begin{aligned}(I) &= \log\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{b=1}^{B} p_b(\beta(x_i)+\gamma)\exp\left\{z^\top(y_b)\Delta\gamma\right\}\right)\\ &\geq \frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{b=1}^{B} p_b(\beta(x_i)+\gamma)\exp\left\{z^\top(y_b)\Delta\gamma\right\}\right).\end{aligned} \tag{46}$$

Notice that

$$\begin{aligned}
&\psi(\beta(x_i)+\gamma+\Delta\gamma)-\psi(\beta(x_i)+\gamma)\\
=&-\log\left(\frac{\sum_{b=1}^{B}e^{z(y_b)^\top(\gamma+\Delta\gamma)}}{\sum_{b=1}^{B}e^{z(y_b)^\top\gamma}}\right)\\
=&-\log\left(\sum_{b=1}^{B}p_b(\beta(x_i)+\gamma)\exp\left\{z(y_b)^\top\Delta\gamma\right\}\right).
\end{aligned}\tag{47}$$

Therefore, by Eq. (45), Eq. (46) and Eq. (47),

$$\begin{aligned}
&\ell(\mathcal{R};\beta(x)+\gamma+\Delta\gamma)-\ell(\mathcal{R};\beta(x)+\gamma)\\
=&\sum_{i=1}^{n}z(y_i)^\top\Delta\gamma-\log\left(\sum_{b=1}^{B}p_b(\beta(x_i)+\gamma)\exp\left\{z(y_b)^\top\Delta\gamma\right\}\right)\\
\geq&\sum_{i=1}^{n}z(Y_i)^\top\Delta\gamma-n\cdot(I)=\sum_{b=1}^{B}n_bz(y_b)^\top\Delta\gamma-n\cdot(I)\\
=&\ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma+\Delta\gamma)-\ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma).
\end{aligned}$$

By the strong concavity of ℓ_{poisson}, there exists a universal $\delta>0$ such that

$$\ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma+\Delta\gamma)-\ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma)\geq\delta\|\Delta\gamma\|_2^2.$$

Thus,

$$\begin{aligned}
&\ell(\mathcal{R};\beta(x)+\gamma+\Delta\gamma)-\ell(\mathcal{R};\beta(x)+\gamma)\\
\geq&\ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma+\Delta\gamma)-\ell_{\text{poisson}}(\mathcal{R};\beta(x)+\gamma)\geq\delta\|\Delta\gamma\|_2^2.
\end{aligned}$$

Since the conditional log-likelihood $\ell(\mathcal{R};\beta(x)+\gamma)$ is bounded, thus can not increase linearly forever. Therefore, $\|\Delta\gamma\|_2\to0$, i.e. the algorithm converges. ■

F.8 Proof of Proposition 5

Proof. The proof follows that of Proposition 2. We simplify the difference of log-likelihoods,

$$\begin{aligned}
\Delta\ell^t(\mathcal{R}, s) = &\sum_{x_i\in\mathcal{R}_L} \Big(z(y_i)^\top \gamma^t_{\mathcal{R}_{s,L}} - \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,L}})\Big) \\
&+ \sum_{x_i\in\mathcal{R}_R} \Big(z(y_i)^\top \gamma^t_{\mathcal{R}_{s,R}} - \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,R}})\Big) \\
&- \sum_{x_i\in\mathcal{R}} \Big(z(y_i)^\top \gamma^t_{\mathcal{R}} - \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}})\Big) \\
&- \lambda n_{\mathcal{R}_{s,L}} \gamma^{t\top}_{\mathcal{R}_{s,L}} \Omega \gamma^t_{\mathcal{R}_{s,L}} - \lambda n_{\mathcal{R}_{s,R}} \gamma^{t\top}_{\mathcal{R}_{s,R}} \Omega \gamma^t_{\mathcal{R}_{s,R}} + \lambda n_{\mathcal{R}} \gamma^{t\top}_{\mathcal{R}} \Omega \gamma^t_{\mathcal{R}} \\
=&n_{\mathcal{R}_{s,L}} \bar{z}^\top_{\mathcal{R}_{s,L}} (\gamma^t_{\mathcal{R}_{s,L}} - \gamma^t_{\mathcal{R}}) + n_{\mathcal{R}_{s,R}} \bar{z}^\top_{\mathcal{R}_{s,R}} (\gamma^t_{\mathcal{R}_{s,R}} - \gamma^t_{\mathcal{R}}) \\
&- \sum_{x_i\in\mathcal{R}_{s,L}} \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,L}}) - \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}}) \\
&- \sum_{x_i\in\mathcal{R}_{s,R}} \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,R}}) - \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}}) \\
&- \lambda n_{\mathcal{R}_{s,L}} \gamma^{t\top}_{\mathcal{R}_{s,L}} \Omega \gamma^t_{\mathcal{R}_{s,L}} - \lambda n_{\mathcal{R}_{s,R}} \gamma^{t\top}_{\mathcal{R}_{s,R}} \Omega \gamma^t_{\mathcal{R}_{s,R}} + \lambda n_{\mathcal{R}} \gamma^{t\top}_{\mathcal{R}} \Omega \gamma^t_{\mathcal{R}}.
\end{aligned} \tag{48}$$

The score equation of $\gamma^t_{\mathcal{R}}$ implies

$$\bar{z}_{\mathcal{R}} = \frac{1}{n_{\mathcal{R}}} \sum_{i=1}^{n} \nabla\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}}) + 2\lambda\Omega\gamma^t_{\mathcal{R}}, \tag{49}$$

and similarly for $\gamma^t_{\mathcal{R}_{s,L}}$, $\gamma^t_{\mathcal{R}_{s,R}}$. Plug Eq. (49) into Eq. (48), we obtain

$$\begin{aligned}
-\Delta\ell(\mathcal{R}, s) = &\sum_{x_i\in\mathcal{R}_{s,L}} \nabla\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,L}})^\top (\gamma^t_{\mathcal{R}_{s,L}} - \gamma^t_{\mathcal{R}}) \\
&+ \sum_{x_i\in\mathcal{R}_{s,R}} \nabla\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,R}})^\top (\gamma^t_{\mathcal{R}_{s,R}} - \gamma^t_{\mathcal{R}}) \\
&- \sum_{x_i\in\mathcal{R}_{s,L}} \Big(\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,L}}) - \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}})\Big) \\
&- \sum_{x_i\in\mathcal{R}_{s,R}} \Big(\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}_{s,R}}) - \psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}})\Big) \\
&+ \lambda n_{\mathcal{R}_{s,L}} \left(\gamma^t_{\mathcal{R}_{s,L}} - \gamma^t_{\mathcal{R}}\right)^\top \Omega \left(\gamma^t_{\mathcal{R}_{s,L}} - \gamma^t_{\mathcal{R}}\right) \\
&+ \lambda n_{\mathcal{R}_{s,R}} \left(\gamma^t_{\mathcal{R}_{s,R}} - \gamma^t_{\mathcal{R}}\right)^\top \Omega \left(\gamma^t_{\mathcal{R}_{s,R}} - \gamma^t_{\mathcal{R}}\right).
\end{aligned} \tag{50}$$

By the Taylor expansion of $\psi(\beta)$ and $\nabla\psi(\beta)$,

$$\begin{aligned}
&\nabla\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}_{s,L}})^\top(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})-(\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}_{s,L}})-\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}}))\\
=&\nabla\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}_{s,L}})^\top(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})-\nabla\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}})^\top(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\\
&-\frac{1}{2}(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})^\top\nabla^2\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}})(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})+O(\|(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\|_2^3)\\
=&\frac{1}{2}(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})^\top\nabla^2\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}})(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})+O(\|(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\|_2^3).
\end{aligned}\tag{51}$$

Plug Eq. (51) into Eq. (50), we get

$$\begin{aligned}
&-\Delta\ell(\mathcal{R},s)\\
=&\frac{1}{2}\sum_{x_i\in\mathcal{R}_{s,L}}(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})^\top\nabla^2\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}})(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\\
&+\frac{1}{2}\sum_{x_i\in\mathcal{R}_{s,R}}(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}})^\top\nabla^2\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}})(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}})\\
&+\lambda n_{\mathcal{R}_{s,L}}\left(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}}\right)^\top\Omega\left(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}}\right)\\
&+\lambda n_{\mathcal{R}_{s,R}}\left(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}}\right)^\top\Omega\left(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}}\right)\\
&+O\left(\|(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\|_2^3+\|(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}})\|_2^3\right)\\
=&\frac{1}{2}(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})^\top\left(\sum_{x_i\in\mathcal{R}_{s,L}}\nabla^2\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}})+2\lambda n_{\mathcal{R}_{s,L}}\Omega\right)(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\\
&+\frac{1}{2}(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}})^\top\left(\sum_{x_i\in\mathcal{R}_{s,R}}\nabla^2\psi(\beta^t(X_i)+\gamma^t_{\mathcal{R}})+2\lambda n_{\mathcal{R}_{s,R}}\Omega\right)(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}})\\
&+O\left(\|(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\|_2^3+\|(\gamma^t_{\mathcal{R}_{s,R}}-\gamma^t_{\mathcal{R}})\|_2^3\right).
\end{aligned}\tag{52}$$

By Eq. (49),

$$\begin{aligned}
&\frac{n_{\mathcal{R}_{s,R}}}{n_{\mathcal{R}}}\left(\bar{z}_{\mathcal{R}_{s,L}}-\bar{z}_{\mathcal{R}_{s,R}}\right)=\bar{z}_{\mathcal{R}_{z,L}}-\bar{z}_{\mathcal{R}}\\
=&\frac{1}{n_{\mathcal{R}_{s,L}}}\sum_{x_i\in\mathcal{R}_{s,L}}\nabla^2\psi(\beta^t(X_i)+\lambda^t_{\mathcal{R}})(\lambda^t_{\mathcal{R}_{s,L}}-\lambda^t_{\mathcal{R}})\\
&+\frac{n_{\mathcal{R}_{s,R}}}{n_{\mathcal{R}}}\left(\bar{z}^t_{\mathcal{R}_{s,L}}-\bar{z}^t_{\mathcal{R}_{s,R}}\right)+O\left(\|(\gamma^t_{\mathcal{R}_{s,L}}-\gamma^t_{\mathcal{R}})\|_2^2\right),
\end{aligned}$$

where $\bar{z}^t_{\mathcal{R}_{s,L}} := \sum_{x_i \in \mathcal{R}_{s,L}} \nabla\psi(\beta^t(X_i) + \lambda^t_{\mathcal{R}})/n_{\mathcal{R}_{s,L}}$, and similarly for $\bar{z}^t_{\mathcal{R}_{s,R}}$. Finally, by the assumption of invertibility,

$$\begin{aligned}
&\gamma^t_{\mathcal{R}_{s,L}} - \gamma^t_{\mathcal{R}} \\
&= \left(\frac{1}{n_{\mathcal{R}_{s,L}}} \sum_{x_i \in \mathcal{R}_{s,L}} \nabla^2\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} \frac{n_{\mathcal{R}_{s,R}}}{n_{\mathcal{R}}} \\
&\quad \left(\left(\bar{z}_{\mathcal{R}_{s,L}} - \bar{z}_{\mathcal{R}_{s,R}}\right) - \left(\bar{z}^t_{\mathcal{R}_{s,L}} - \bar{z}^t_{\mathcal{R}_{s,R}}\right)\right) + O\left(\|(\gamma^t_{\mathcal{R}_{s,L}} - \gamma^t_{\mathcal{R}})\|_2^2\right) \qquad (53) \\
&= \frac{n_{\mathcal{R}_{s,R}}}{n_{\mathcal{R}}} \left(\frac{1}{n_{\mathcal{R}_{s,L}}} \sum_{x_i \in \mathcal{R}_{s,L}} \nabla^2\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} \left(\bar{r}^t_{\mathcal{R}_{z,L}} - \bar{r}^t_{\mathcal{R}_{z,R}}\right) \\
&\quad + O\left(\|(\gamma^t_{\mathcal{R}_{s,L}} - \gamma^t_{\mathcal{R}})\|_2^2\right).
\end{aligned}$$

Plug Eq. (53) into Eq. (52),

$$\begin{aligned}
&- \Delta\ell(\mathcal{R}, s) \\
&= \frac{n_{\mathcal{R}_{s,L}} n^2_{\mathcal{R}_{s,R}}}{2n^2_{\mathcal{R}}} (\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}_{s,R}})^\top \left(\frac{1}{n_{\mathcal{R}_{s,L}}} \sum_{x_i \in \mathcal{R}_{s,L}} \nabla^2\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} \\
&\quad (\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}_{s,R}}) + \frac{n^2_{\mathcal{R}_{s,L}} n_{\mathcal{R}_{s,R}}}{2n^2_{\mathcal{R}}} (\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}_{s,R}})^\top \\
&\quad \left(\frac{1}{n_{\mathcal{R}_{s,R}}} \sum_{x_i \in \mathcal{R}_{s,R}} \nabla^2\psi(\beta^t(X_i) + \gamma^t_{\mathcal{R}}) + 2\lambda\Omega\right)^{-1} (\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}_{s,R}}) \\
&\quad + O\left(\|(\bar{r}^t_{\mathcal{R}_{s,L}} - \bar{r}^t_{\mathcal{R}})\|_2^3 + \|(\bar{r}^t_{\mathcal{R}_{s,R}} - \bar{r}^t_{\mathcal{R}})\|_2^3\right),
\end{aligned}$$

and we finish the proof. ■

F.9 Proof of Proposition 6

Proof. We first evaluate the complexity of the fitting step of LinCDE boosting. The computation of offset is $O(n_{\mathcal{R}}kB)$, and we store the cell probabilities $p_b(\beta^t(x_i))$. The first step of the fitting runs a penalized Lindsey's method and takes $O(n_{\mathcal{R}} + k^2B + k^3)$ as shown in the proof of 3. The second step in the fitting takes $O(n_{\mathcal{R}}B + kB)$, and we update the cell probabilities to $p_b(\beta^t(x_i) + \gamma^t_{\mathcal{R}})$.

It takes $O(n_{\mathcal{R}}k^2B)$ to compute the surrogate normalization matrix $\tilde{\Psi}^t(\gamma^t_{\mathcal{R}})$ and an extra $O(k^3)$ for matrix inversion. It takes $\tilde{O}(dn_{\mathcal{R}}kB)$ to compute all average residuals $\bar{r}^t_{\mathcal{R}}$. Finally, the quadratic approximation for all candidate splits and choosing the optimal one takes $O(Sk^2)$. In summary, the time complexity is $\tilde{O}(dn_{\mathcal{R}}kB + n_{\mathcal{R}}k^2B + k^3 + Sk^2)$. ■

Appendix G. Additional Figures

In this section, we provide additional numerical results.

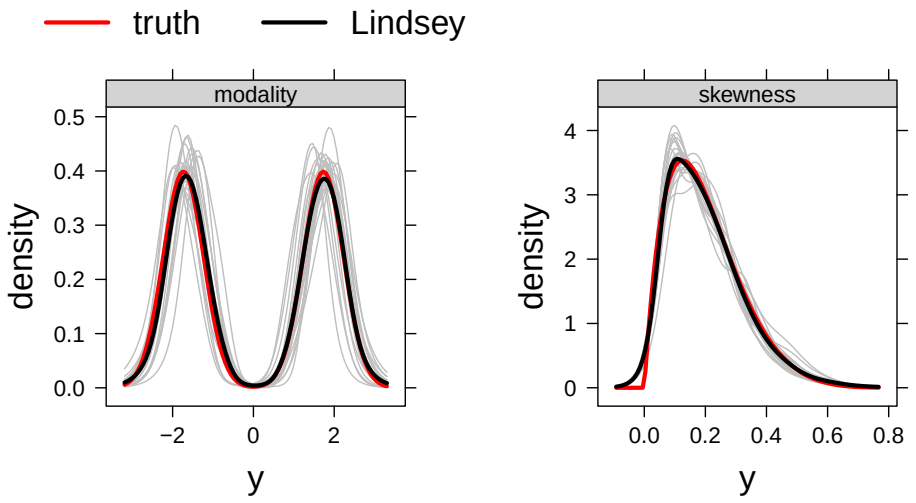


Figure 17: Estimation of bimodal and skewed densities using our regularized Lindsey's method. In the left panel, the true density is a Gaussian mixture, and in the right panel, the true density is a beta distribution. In each trial, we sample 400 observations. We generate 10 natural cubic splines with knots equally spread across the range of observations, and tune the penalty parameter λ to achieve an effective 5 degrees of freedom. We repeat each setting 100 times and plot the average fits against the true densities.

G.1 Additional Simulations

Figure 17 displays the performance of Lindsey's method in two toy examples. One target density is bimodal, and the other is skewed. In Lindsey's method, we use natural cubic splines, which are arguably the most commonly used splines because they provide good and seamless fits and are easy to implement, and transform them as discussed. In both examples, the estimated densities of Lindsey's method match the true densities quite closely, except for tiny gaps at boundaries and peaks.

Figure 18 presents the goodness-of-fit measure (24) of the three methods under the *LGD* and *LGGMD* settings with 50 bins. LinCDE boosting benefits from finer grids. In contrast, distribution boosting and quantile regression forest are hurt by larger numbers of bins due to higher variances, especially in the *LGGMD* setting where the densities themselves are bumpy.

Similar to AAE, another commonly used CDF-based metric is the cram/'er-von Mises distance (Friedman, 2019):

$$\text{CVM} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\hat{F}(q(u_j \mid x_i) \mid x_i) - F(q(u_j \mid x_i) \mid x_i)\right)^2, \tag{54}$$

where $\{u_j\}$ is an evenly spaced grid on $[0, 1]$, and $q(u \mid x)$ denotes the u quantile at the covariate value x. Figure 19 depicts the cram/'er-von Mises distance. In both settings,

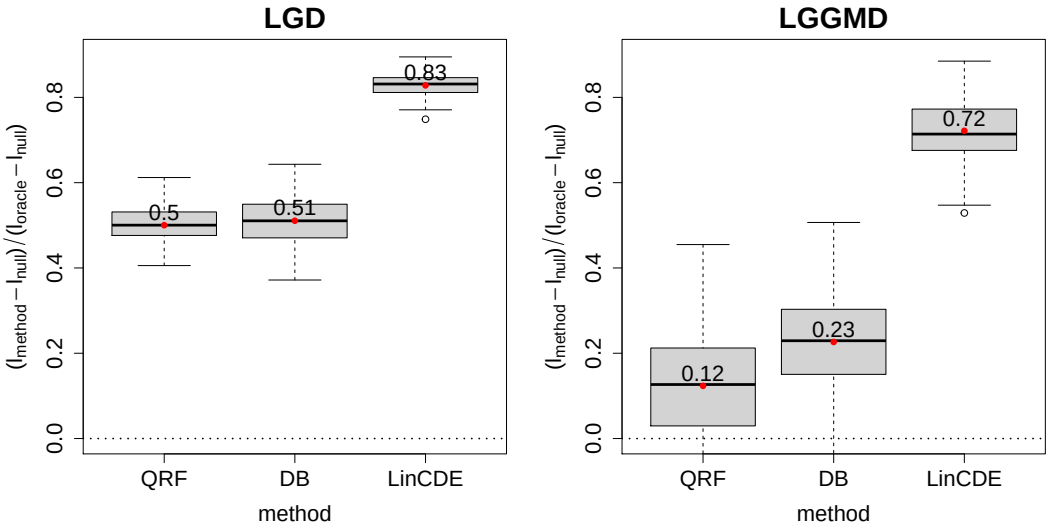


Figure 18: Box plots of goodness-of-fit measure (24) in the *LGD* (left panel) and *LGGMD* (right panel) settings. The densities are computed as (25) with 50 bins. QRF stands for quantile regression forest, DB stands for distribution boosting, LinCDE stands for LinCDE boosting respectively.

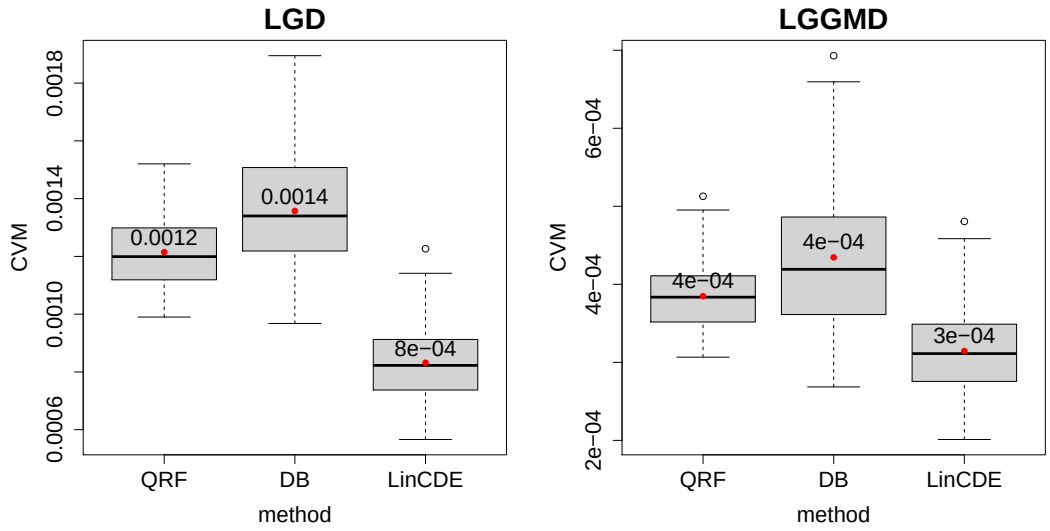


Figure 19: Box plots of CVM distance (54) in the *LGD* (left panel) and *LGGMD* (right panel) settings. CVM is a metric naturally defined for conditional CDF estimates (Friedman, 2019), and a smaller value indicates a better estimate.

LinCDE boosting outperforms the other two. The results are consistent with those of AAE.

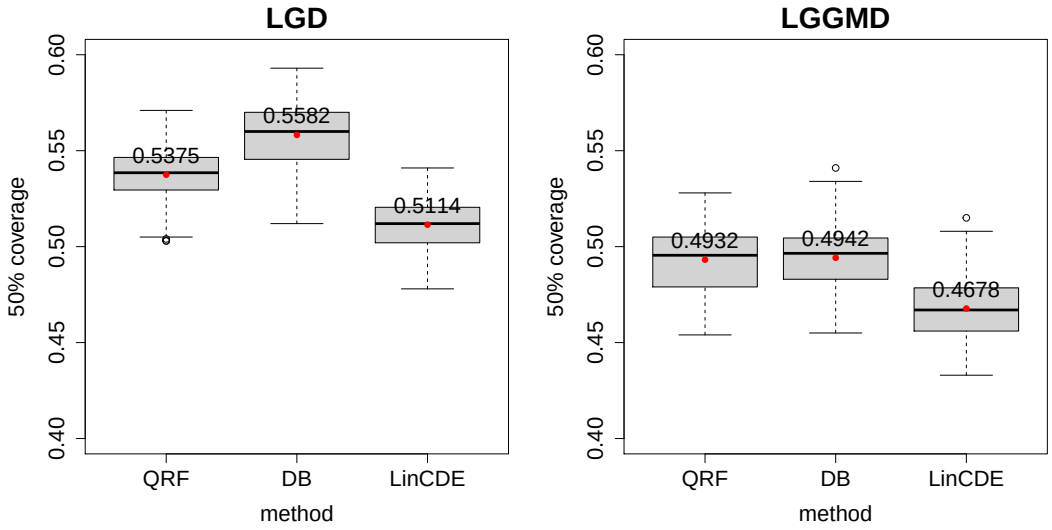


Figure 20: Coverage of 50% prediction intervals (ideally 50%) in the *LGD* (left panel) and *LGGMD* (right panel) settings.

In Figure 20, we plot the coverages of the 50% prediction intervals. The results are consistent with those of 90% prediction intervals.

G.2 Air Pollution Data

The air pollution data (Wu and Dominici, 2020) focuses on PM2.5 exposures in the United States.[14] The responses are 3108 county-level PM2.5 exposures averaged from 2000 to 2018. We incorporate 16 weather, socio-economic, and demographic covariates, such as winter relative humidity, house value, and proportion of white people. The target is estimating the conditional density of the average PM2.5 exposure. We split the data into training, validation, and test (proportion 2:1:1), and tune on the hold-out validation data. We also identify influential covariates which may help find the culprits of air pollution and manage regional air quality.

The PM2.5 exposure varies vastly across counties. For example, the average PM2.5 of west coast counties may soar up to $12\mu g/m^3$ due to frequent wildfires, while those of rural areas in Central America are typically below $8\mu g/m^3$. The difference in PM2.5 levels induces the disjoint support issue for LinCDE in Section 6, and thus we employ the centering. The comparisons of log-likelihood and quantile losses are summarized in Table 4. Centered LinCDE performs the best in log-likelihood and is comparable in quantile losses.

In Figure 21, we display how the estimated conditional densities change with respect to winter relative humidity and house value—top influential covariates identified by LinCDE.

- Winter relative humidity affects the locations of the conditional densities: as the humidity goes up, the PM2.5 concentration first increases, then decreases. One hy-

14. PM2.5: particulate particles 2.5 microns or less in diameter. The data represents 98% of the population of the United States.

pothesis of the inverted U-shaped relationship is that in dry counties, wind speed that is inversely correlated with the humidity is a powerful factor for PM2.5; in wet counties, moisture particles that accelerate the deposition process of PM2.5 are the driving force.

- House value is impactful to the scales of the conditional densities: more expensive houses are associated with more variable PM2.5 exposures. We conjecture that rural areas are generally low in PM2.5 while urban areas vary. Higher house values suggest a higher proportion of urban areas and thus less homogeneous air quality.

data	method	-log-like	quantile loss				
			5%	25 %	50%	75 %	95 %
Air pollution	QRF	0.95 (0.02)	0.077 (0.002)	0.189 (0.005)	0.229 (0.007)	0.206 (0.006)	0.087 (0.002)
	DB	1.27 (0.04)	0.099 (0.007)	0.247 (0.009)	0.300 (0.010)	0.244 (0.008)	0.093 (0.006)
	LinCDE	**0.89** (0.03)	0.063 (0.003)	0.185 (0.006)	0.233 (0.007)	0.194 (0.007)	0.072 (0.004)

Table 4: Comparison of LinCDE boosting, QRF, and DB on the air pollution data. We display the negative log-likelihoods and quantile losses at $\{5\%, 25\%, 50\%, 75\%, 95\%\}$ levels. Standard deviations are in the parentheses.

References

Barry C Arnold, Enrique Castillo, José-Mariá Sarabia, and Jose M Sarabia. *Conditional Specification of Statistical Models*. Springer, 1999.

Adelchi Azzalini and Adrian W Bowman. A look at some data on the old faithful geyser. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(3):357–365, 1990.

Nicholas M Ball, Robert J Brunner, Adam D Myers, Natalie E Strand, Stacey L Alberts, and David Tcheng. Robust machine learning applied to astronomical data sets. iii. probabilistic photometric redshifts for galaxies and quasars in the sdss and galex. *The Astrophysical Journal*, 683(1):12, 2008.

Vic Barnett. The ordering of multivariate data (with comments). *Journal of the Royal Statistical Society: Series A (General)*, 139(3):318–354, 1976.

Andrew R Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29, 2019.

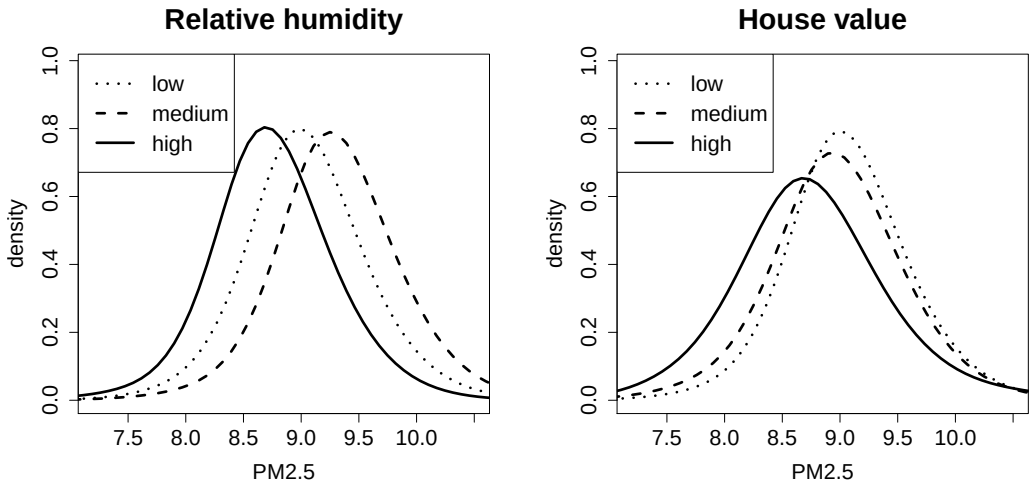


Figure 21: LinCDE boosting applied to the air pollution data. We plot the estimated densities of LinCDE boosting across different levels of top influential covariates: winter relative humidity (left panel) and house value (right panel). Other covariates are fixed at the corresponding medians. Winter relative humidity is influential to the locations of the conditional densities: as the humidity goes up, the PM2.5 concentration first increases, then decreases. House value is impactful to the scales of the conditional densities: more expensive houses are associated with more variable PM2.5 exposures.

Christopher M Bishop. Mixture density networks. Technical report, 1994.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees*. CRC press, 1984.

Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

Probal Chaudhuri. Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of multivariate analysis*, 39(2):246–269, 1991a.

Probal Chaudhuri. Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of statistics*, 19(2):760–777, 1991b.

Probal Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.

Probal Chaudhuri and Wei-Yin Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8(5):561–576, 2002.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Carol DeSantis, Jiemin Ma, Leah Bryan, and Ahmedin Jemal. Breast cancer statistics, 2013. *CA: a cancer journal for clinicians*, 64(1):52–62, 2014.

Jay L Devore and Roxy Peck. *Statistics: the Exploration and Analysis of Data*. Brooks/Cole, Cengage Learning, 5 edition, 2005.

Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference*, volume 5. Cambridge University Press, 2016.

Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.

Silverio Foresi and Franco Peracchi. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90(430):451–466, 1995.

Jerome Friedman and Balasubramanian Narasimhan. *conTree: Contrast Trees and Boosting*, 2020. R package version 0.2-4.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Jerome H Friedman. Contrast trees and distribution boosting. *arXiv preprint arXiv:1912.03785*, 2019.

Peter Hall, Rodney CL Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, 94(445):154–163, 1999.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. Springer, New York, second edition, 2009.

Xuming He, Pin Ng, and Stephen Portnoy. Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):537–550, 1998.

Rafael Izbicki and Ann B Lee. Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4): 1297–1316, 2016.

Roger Koenker. Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176, 2017.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, 46(1):33–50, 1978.

Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81 (4):673–680, 1994.

Linglong Kong and Ivan Mizera. Quantile tomography: using quantiles with multivariate data. *Statistica Sinica*, 22:1589–1610, 2012.

Charles Kooperberg and Charles J Stone. A study of logspline density estimation. *Computational Statistics & Data Analysis*, 12(3):327–347, 1991.

Charles Kooperberg and Charles J Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328, 1992.

Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer, 2006.

Youjuan Li, Yufeng Liu, and Ji Zhu. Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

J. K. Lindsey. Comparison of probability distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):38–47, 1974.

Benoît R Mâacsse and Young K Truong. Conditional logspline density estimation. *Canadian Journal of Statistics*, 27(4):819–832, 1999.

H. Markowitz. Portfolio selection. *Investment under Uncertainty*, 1959. URL `https://ci.nii.ac.jp/naid/10027840150/en/"`.

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

Nicolai Meinshausen. *quantregForest: Quantile Regression Forests*, 2017. URL `https://CRAN.R-project.org/package=quantregForest`. R package version 1.3-7.

Laura Moody, Suparna Mantha, Hong Chen, and Yuan-Xiang Pan. Computational methods to identify bimodal gene expression and facilitate personalized treatment in cancer patients. *Journal of Biomedical Informatics: X*, 1:100001, 2019.

Panagis Moschopoulos and Joan G Staniswalis. Estimation given conditionals from an exponential family. *The American Statistician*, 48(4):271–275, 1994.

Mark F Schilling, Ann E Watkins, and William Watkins. Is human height bimodal? *The American Statistician*, 56(3):223–229, 2002.

Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982.

Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, 1986.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.

Charles J Stone. Asymptotics for doubly flexible logspline response models. *The annals of statistics*, 19(4):1832–1854, 1991a.

Charles J Stone. Multivariate log-spline conditional models. *Technical report, University of California, Berkeley*, 1991b.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 781–788, 2010.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Grace Wahba. Smoothing noisy data with spline functions. *Numerische mathematik*, 24(5): 383–393, 1975.

Grace Wahba. Optimal smoothing of density estimates, in classification and clustering. *Academic Press, New York*, pages 423–458, 1977.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Nethery R. C. Sabath M. B. Braun D. Wu, X. and F. Dominici. Air pollution and covid-19 mortality in the united states: Strengths and limitations of an ecological regression analysis. *Science advances*, 6(45):p.eabd4049, 2020.

Keming Yu and M. C. Jones. Local linear quantile regression. *Journal of the American statistical Association*, 93(441):228–237, 1998.