Modeling of Time Series Using Random Forests: Theoretical Developments

Richard A. Davis and Mikkel Slot Nielsen
Columbia University

"New Results on Time Series and their Statistical Applications"
Centre International de Rencontres Mathématiques
Luminy

September 14–18, 2020

Random forests

Random Forests: Idea introduced by Breiman (2001), inspired by CART Breiman, et al. (1984). (63,000 and 47,000 citations)

Leo Breiman:

- Started out as a probabilist (Breiman's Lemma for heavy tails) – influenced by Loève and Blackwell
- Textbook: *Probability*
- CART et al. (Breiman, Friedman, Olshen, and Stone.)
- Conversation with Leo Breiman–fascinating look at the man.
- *Statistical Modeling: The Two Cultures (2001)* – generative models and predictive models.

Random forests

Classical setting for random forest: $\{(X_t, Y_t)\}$ an iid sequence of observations from the model

$$Y = f(X) + \varepsilon,$$

Random forests

Classical setting for random forest: $\{(X_t, Y_t)\}$ an iid sequence of observations from the model

$$Y = f(X) + \varepsilon,$$

Goal: Estimate $f(x) = E(Y|X = x)$. Here X may be *high-dimensional* and f is *nonlinear*.

Random forests

Classical setting for random forest: $\{(X_t, Y_t)\}$ an iid sequence of observations from the model

$$Y = f(X) + \varepsilon,$$

Goal: Estimate $f(x) = E(Y|X = x)$. Here X may be *high-dimensional* and f is *nonlinear*.

Random forests Use regression trees or a recursive partition of the feature space, i.e., *independent variables* X_t.

Random forests

Classical setting for random forest: $\{(X_t, Y_t)\}$ an iid sequence of observations from the model

$$Y = f(X) + \varepsilon,$$

Goal: Estimate $f(x) = E(Y|X = x)$. Here X may be *high-dimensional* and f is *nonlinear*.

Random forests Use regression trees or a recursive partition of the feature space, i.e., *independent variables* X_t.

Advantages:

- learning algorithm that aggregates estimates over a large number of trees.
- widely used in a variety of applications including object recognition, bioinformatics, ecology, and finance.
- little tuning required
- some (Howard and Bowles (2012) claim random forests are the most successful general-purpose prediction algorithms.

Random forests

Time series setting: The time series $\{Y_t\}$ is a Nonlinear AR (NLAR) process satisfying the recursions,

$$Y_t = f(Y_{t-1}, \ldots, Y_{t-p}) + \varepsilon_t, \qquad t \geq 1,$$

with initial values $Y_0, \ldots, Y_{1-p}$. (p could be large and increasing with sample size.)

Objective: Based on observations $(X_1, Y_1), \ldots, (X_T, Y_T)$, where $X_t = (Y_{t-1}, \ldots, Y_{t-p})$, estimate $E(Y|X = x)$ using random forests.

Recursive partitions: Start with $\mathcal{P}_1 = \{\mathbb{R}^p\}$ and then construct $\mathcal{P}_{n+1}$ from $\mathcal{P}_n$ as follows:

- Select an (unsplit) node $A \in \mathcal{P}_n$
- Select a split direction i–could be random with prob $p_i, i = 1, \ldots, p$
- Determine split position $\tau \in \{x_i : x \in A\}$ chosen in accordance with some set of rules.

$$A_L := \{x \in A : x_i \leq \tau\} \quad \text{and} \quad A_R := \{x \in A : x_i > \tau\},$$

(x_i refers to the i-th entry of $x \in \mathbb{R}^p$)

- A is the *parent* node of A_L and A_R; A_L and A_R are *child* nodes of A.

Remark: A given partition Λ of $\mathbb{R}^p$ is called recursive if $\Lambda = \mathcal{P}_n$ for some $n \geq 1$, where $\mathcal{P}_1, \ldots, \mathcal{P}_n$ are obtained as above.

Choosing node, direction and position of split

Model. Observations $Y_1, \ldots, Y_T$ from the model

$$Y_t = f(X_t) + \varepsilon_t, \qquad t \geq 1,$$

where $X_t := (Y_{t-1}, \ldots, Y_{t-p})$ and $\xi := (Y_0, \ldots, Y_{1-p})$.
Group these into input-out pairs,

$$\mathcal{D}_T = \{(X_1, Y_1), \ldots, (X_T, Y_T)\}.$$

- split direction i: probability of splitting in i^{th} direction is $p_i = p_i(\mathcal{D}_T)$.
- split position τ: Breiman suggests maximizing impurity, i.e., minimize the sum of variances

$$\frac{1}{|\{i : X_i \leq \tau\}|} \sum_{X_i \leq \tau} (Y_i - \bar{Y}_{\leq \tau})^2 + \frac{1}{|\{i : X_i > \tau\}|} \sum_{X_i > \tau} (Y_i - \bar{Y}_{> \tau})^2$$

  Similar in spirit to change-point problem in a random field. Choice of τ may depend on $\mathcal{D}_T$ and other independent random mechanism Θ.

Regression tree estimate

Regression tree estimate: For a recursive tree partition Λ, the estimate is

$$T_\Lambda(x) = \frac{1}{|\{t : X_t \in A_\Lambda(x)\}|} \sum_{t=1}^{T} Y_t 1_{A_\Lambda(x)}(X_t), \qquad x \in \mathbb{R}^p\,, \tag{1}$$

where A_Λ is the unique leaf in the partition containing x.

Partition-optimal tree:

$$T^*_\Lambda(x) := \mathbb{E}_\Lambda[Y \mid X \in A_\Lambda(x)] \tag{2}$$

- (X, Y) is a copy of (X_1, Y_1) and indep of $(\mathcal{D}_T, \Theta)$
- $\mathbb{E}_\Lambda$ denotes expectation with respect to the conditional probability measure $\mathbb{P}_\Lambda := \mathbb{P}(\,\cdot \mid \mathcal{D}_T, \Theta)$.
- The set $A_\Lambda(x)$ is treated as non-random in (2).
- We'll show $T_\Lambda(x) - T^*_\Lambda(x)$ goes to 0 uniformly (in some sense)

Assumptions on the model

A.1 ε_1 has pdf $h_\varepsilon(x) > 0$ on $\mathbb{R}$ and, for some $c \in (0, \infty)$,

$$\mathbb{E}[|\varepsilon_1|^m] \leq m! c^{m-2}, \qquad m \geq 3 \quad \text{(Bernstein Condition)} \tag{3}$$

The cdf $F_\varepsilon(x) = \int_{-\infty}^{x} h_\varepsilon(y)\, dy$ of ε_1 satisfies

$$\sup_{x \in \mathbb{R}} \frac{F_\varepsilon(x + \tau)}{F_\varepsilon(x)} < \infty \tag{4}$$

for any $\tau \in (0, \infty)$.

Assumptions on the model

A.1 ε_1 has pdf $h_\varepsilon(x) > 0$ on $\mathbb{R}$ and, for some $c \in (0, \infty)$,

$$\mathbb{E}[|\varepsilon_1|^m] \leq m! c^{m-2}, \qquad m \geq 3 \quad \text{(Bernstein Condition)} \tag{3}$$

The cdf $F_\varepsilon(x) = \int_{-\infty}^{x} h_\varepsilon(y)\, dy$ of ε_1 satisfies

$$\sup_{x \in \mathbb{R}} \frac{F_\varepsilon(x + \tau)}{F_\varepsilon(x)} < \infty \tag{4}$$

for any $\tau \in (0, \infty)$.

A.2 f is bounded, i.e., $M := \sup_{x \in \mathbb{R}^p} |f(x)| < \infty$..

Assumptions on the model

A.1 ε_1 has pdf $h_\varepsilon(x) > 0$ on $\mathbb{R}$ and, for some $c \in (0, \infty)$,

$$\mathbb{E}[|\varepsilon_1|^m] \leq m! c^{m-2}, \qquad m \geq 3 \quad \text{(Bernstein Condition)} \tag{3}$$

The cdf $F_\varepsilon(x) = \int_{-\infty}^{x} h_\varepsilon(y)\, dy$ of ε_1 satisfies

$$\sup_{x \in \mathbb{R}} \frac{F_\varepsilon(x + \tau)}{F_\varepsilon(x)} < \infty \tag{4}$$

for any $\tau \in (0, \infty)$.

A.2 f is bounded, i.e., $M := \sup_{x \in \mathbb{R}^p} |f(x)| < \infty$..

A.3 Minimum number of points k in a leaf satisfies $k/(\log T)^4 \to \infty$ as $T \to \infty$.

Comments on the assumptions

A1: ε_1 has pdf, positive, Bernstein, etc.

- gives a geometrically ergodic stationary solution to Markov chain (could get by with less)
- Bernstein condition implies ε_1 is subexpoential, $P(|\varepsilon_1| > x) \leq \gamma_1 e^{-\gamma_2 x}$ for $\gamma_1, \gamma_2 > 0$.
- Left tail condition (4) is implied by $\lim_{x\to\infty} \frac{h_\varepsilon(x)}{h_\varepsilon(x+\tau)}$ exists and is nonzero for $\tau > 0$.

A2: f bounded. Implicit in virtually all theoretical work since the input vector is assumed to live on $[0,1]^p$ and f is continuous.

A3: $k/(\log T)^4 \to \infty$. Number of points in the leaves has to grow, but not too fast. $(\log T)^4$ is needed in order to apply Bernstein-like inequality to strong mixing sequences and bound has to apply uniformly across all trees.

First result: a concentration inequality

Theorem 1

Suppose that A.1–A.3 are satisfied. Then there exists a constant $\beta \in (0, \infty)$ *such that*

$$\sup_{(x,\Lambda)\in\mathbb{R}^p\times\mathcal{V}_k} |T_\Lambda(x) - T^*_\Lambda(x)| \leq \beta \frac{(\log T)^2}{\sqrt{k}} \tag{5}$$

with probability at least $1 - 4T^{-1}$ *for all sufficiently large* T*.*

($\mathcal{V}_k$ = all partition whose leafs contain at least k points.)

Remark $|T_\Lambda(x) - T^*_\Lambda(x)|$ is the deviation of *mean-corrected* sample average of at least k observations. For some choices of pairs (x, Λ) the number is exactly k and hence error is $1/\sqrt{k}$. Upper bound includes $(\log T)^2$, small price to get uniform rates.

A Corollary for forests

Forests Let $\mathcal{W}_k := \{\Lambda \subseteq \mathcal{V}_k : |\Lambda| < \infty\}$ = collection of k-valid partitions (trees). Let $\mathbf{\Lambda} = \{\Lambda_1, \ldots, \Lambda_B\}$ be B trees in $\mathcal{W}_k$.

Random forest *estimate* and its partition optimal-counterpart are

$$H_{\mathbf{\Lambda}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_{\Lambda_b}(x) \quad \text{and} \quad H^*_{\mathbf{\Lambda}}(x) = \frac{1}{B}\sum_{b=1}^{B} T^*_{\Lambda_b}(x), \qquad x \in \mathbb{R}^p.$$

Corollary 2

Suppose that A.1–A.3 are satisfied. Then there exists a constant $\beta \in (0, \infty)$ *such that*

$$\sup_{(x,\mathbf{\Lambda}) \in \mathbb{R}^p \times \mathcal{W}_k} |H_{\mathbf{\Lambda}}(x) - H^*_{\mathbf{\Lambda}}(x)| \leq \beta \frac{(\log T)^2}{\sqrt{k}}$$

with probability at least $1 - 4T^{-1}$ *for all sufficiently large* T.

Notes.

- This development follows that of Wager and Walther (2015) in the iid setting.
- All the trees are constructed from the same data set $\mathcal{D}_T$.
- Breiman (2001) uses an initial bootstrap step before growing the trees.
- If f is smooth and if the diameter of each leaf shrinks to 0, then the forest estimates should be consistent.

Getting to consistency.

Construction of the trees. For $\alpha \in (0, .5)$, $m \geq 2k$, Λ is an (α, k, m)-valid partition ($\Lambda \in \mathcal{V}_{\alpha,k,m}$) if

(i) Any currently unsplit node with at least m data points will eventually be split.

(ii) The probability $\rho_i = \rho_i(\mathcal{D}_T)$ that a given (feasible) node is split along the i-th direction is bounded from below for all $i = 1, \ldots, p$ by a strictly positive constant.

(iii) The split position is chosen such that each child node contains at least a fraction $\alpha \in (0, 1/2)$ of the data points in its parent node.

(iv) All leaves of the tree contain at least k data points.

A4: The function f in (1) is C-Lipschitz, that is,

$$|f(x') - f(x)| \leq C\|x' - x\| \qquad \text{for all } x, x' \in \mathbb{R}^p$$

A5: It holds that $\log(T/m)/\log(\alpha^{-1}) \to \infty$ as $T \to \infty$.

Theorem 3

Let $\hat{f}_T$ be an (α, k, m)-forest and suppose that A.1–A.5 are satisfied. Then the following hold:

(a) *$\hat{f}_T$ is a pointwise consistent estimator of f in the sense that*

$$\hat{f}_T(x) \longrightarrow f(x) \qquad \text{in probability as } T \to \infty.$$

for any $x \in \mathbb{R}^p$.

(b) *$\hat{f}_T(X)$ is a consistent estimator of the conditional mean $\mathbb{E}[Y \mid X]$ in the sense that*

$$\hat{f}_T(X) \longrightarrow \mathbb{E}[Y \mid X] \qquad \text{in probability as } T \to \infty.$$

Main ideas in the arguments

Transform X_t to $[0,1]^p$. Choose a mapping

$$\iota_h \colon (x_1, \ldots, x_p) \longmapsto (F_h(x_1), \ldots, F_h(x_p))$$

such that $Z_t = \iota_h(X_t)$ has pdf h_Z on $[0,1]^p$ with

$$\zeta^{-1} \leq h_Z(z) \leq \zeta \quad \text{for all } z \in [0,1]^p.$$

This is purely conceptual for proofs–not needed in practice.

Main ideas in the arguments

Transform X_t to $[0,1]^p$. Choose a mapping

$$\iota_h \colon (x_1, \ldots, x_p) \longmapsto (F_h(x_1), \ldots, F_h(x_p))$$

such that $Z_t = \iota_h(X_t)$ has pdf h_Z on $[0,1]^p$ with

$$\zeta^{-1} \le h_Z(z) \le \zeta \quad \text{for all } z \in [0,1]^p.$$

This is purely conceptual for proofs–not needed in practice.

Concentration inequality. Use Bernstein-type inequality to show

$$\log \mathbb{P}\Big(\Big| \frac{\#R}{T} - \mu(R) \Big| > x \Big) \lesssim - \frac{x^2 T}{\nu_R^2 + T^{-1} + x(\log T)^2}, \qquad x > 0,$$

($\#R = |\{t : Z_t \in R\}$ and $\nu_R^2 := \mathrm{Var}(1_R(Z_1)) + 2\sum_{t=1}^{\infty} |\mathrm{Cov}(1_R(Z_{t+1}), 1_R(Z_1))|$.)

Main ideas in the arguments

Transform X_t to $[0,1]^p$. Choose a mapping

$$\iota_h \colon (x_1, \ldots, x_p) \longmapsto (F_h(x_1), \ldots, F_h(x_p))$$

such that $Z_t = \iota_h(X_t)$ has pdf h_Z on $[0,1]^p$ with

$$\zeta^{-1} \leq h_Z(z) \leq \zeta \quad \text{for all } z \in [0,1]^p.$$

This is purely conceptual for proofs–not needed in practice.

Concentration inequality. Use Bernstein-type inequality to show

$$\log \mathbb{P}\left(\left| \frac{\#R}{T} - \mu(R) \right| > x \right) \lesssim - \frac{x^2 T}{\nu_R^2 + T^{-1} + x(\log T)^2}, \qquad x > 0,$$

($\#R = |\{t : Z_t \in R\}$ and $\nu_R^2 := \mathrm{Var}(1_R(Z_1)) + 2\sum_{t=1}^{\infty} |\mathrm{Cov}(1_R(Z_{t+1}), 1_R(Z_1))|$.)

Leaves get smaller. Conditions A.1, A.2, A.5 imply $\mathrm{diam}(A_\Lambda(x)) \to 0$ as $T \to \infty$. (Proof is similar but more complex than in Meihshausen (2006).)

Simulation study

Model $Y_t = f(Y_{t-1}) + \varepsilon_t$, $\{\varepsilon_t\} \sim$ IID Laplace, $h_\varepsilon(x) = \frac{1}{2}e^{-|x|}$.

Four models:

1. $f(x) = 0.5\,\text{sign}(x)\min\{|x|, 10\}$, truncated AR(1)
2. $f(x) = -2xe^{-0.7x^2} + 3x^2e^{-0.95x^2}$, exponential AR(2)
3. $f(x) = \cos(5x)e^{-x^2}$, damped sinusoid
4. $f(x) = \min\{|x|, 0.75\}\min\{|x|, 10\}$, spline-like.

Used **ranger** package in R.

- $B = 400$ trees
- T=400, 1600, 6400
- $k = \lfloor 0.04(\log T)^4 \log\log T \rfloor$.

Simulation study

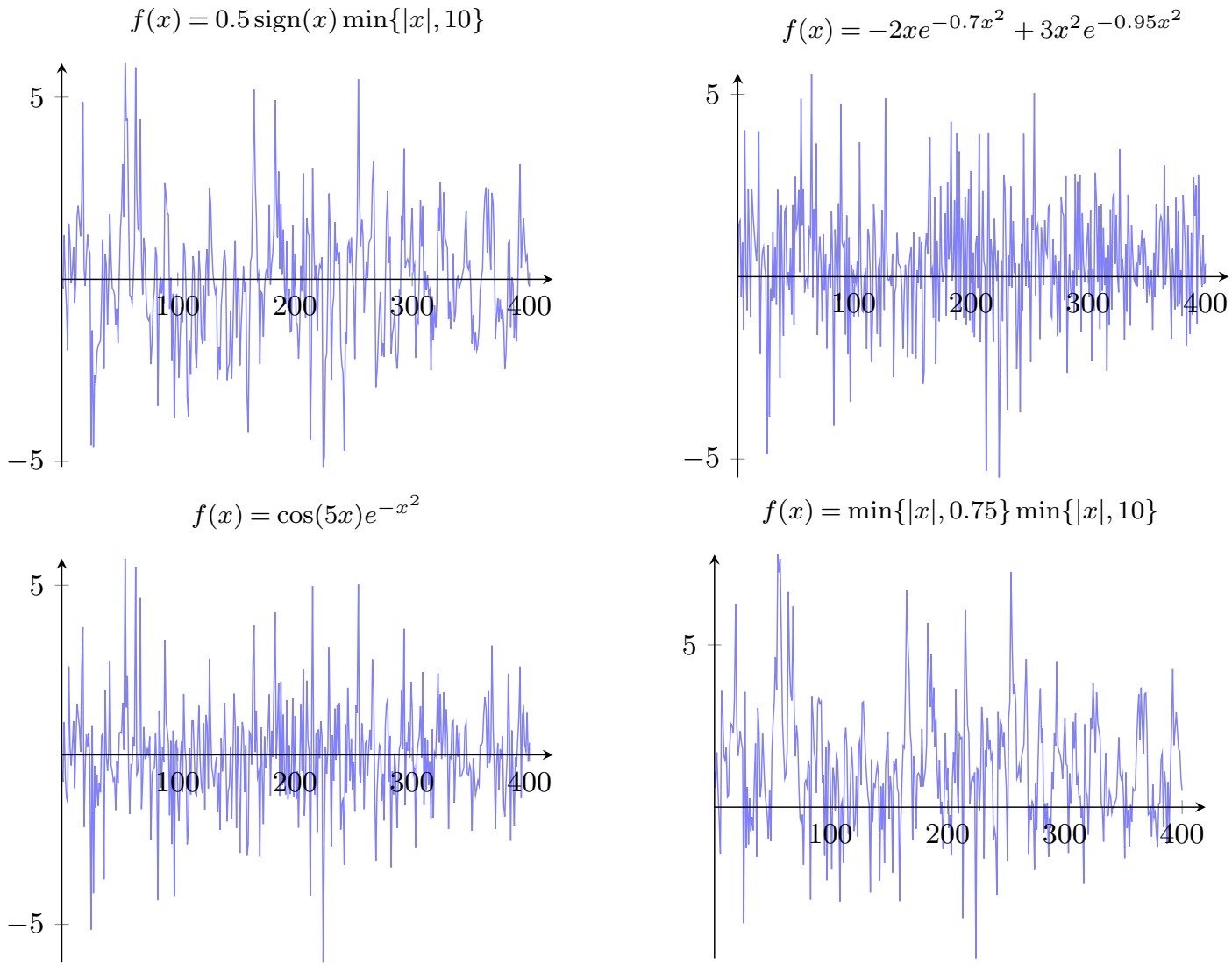
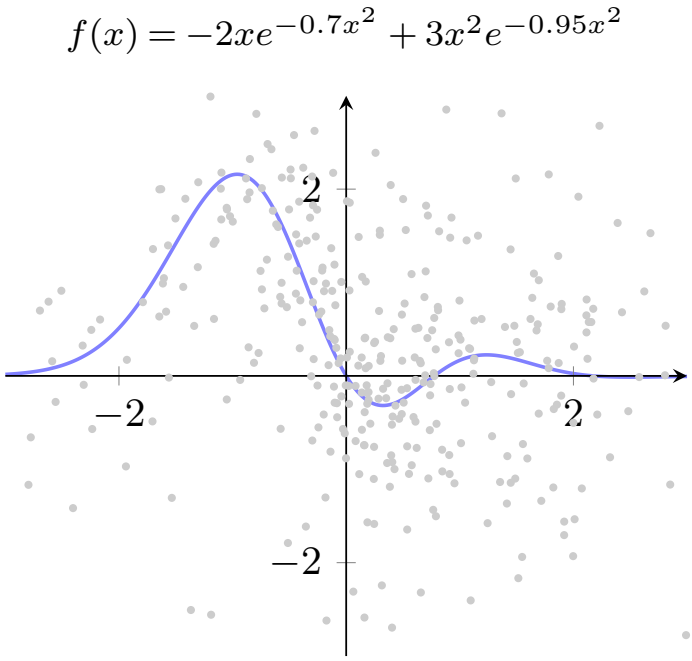


FIG 1. *Simulations of* $Y_1, \ldots, Y_{400}$ *from the model* (2.1) *for the four different specifications of* f *considered in* (4.1).

Simulation study



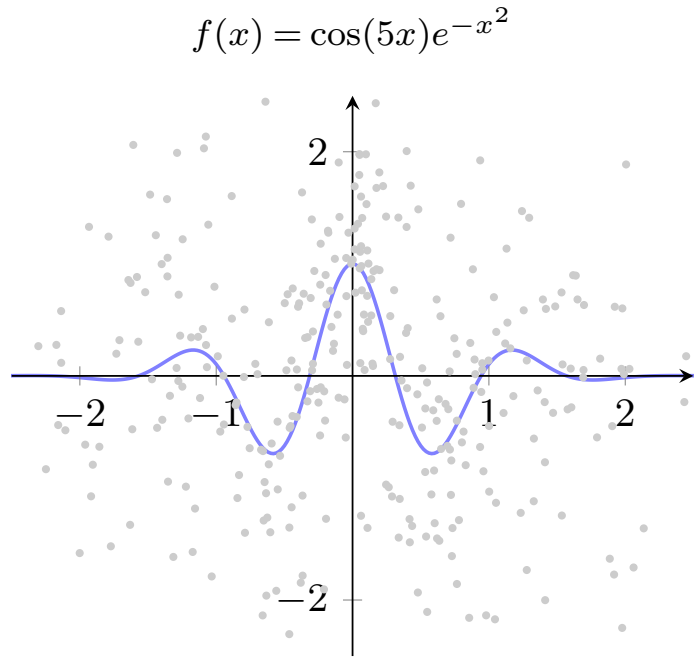


FIG 3. *Scatter plots of the data $\mathcal{D}_{400}$ under two of the specifications of f considered in* (4.1).

Simulation study

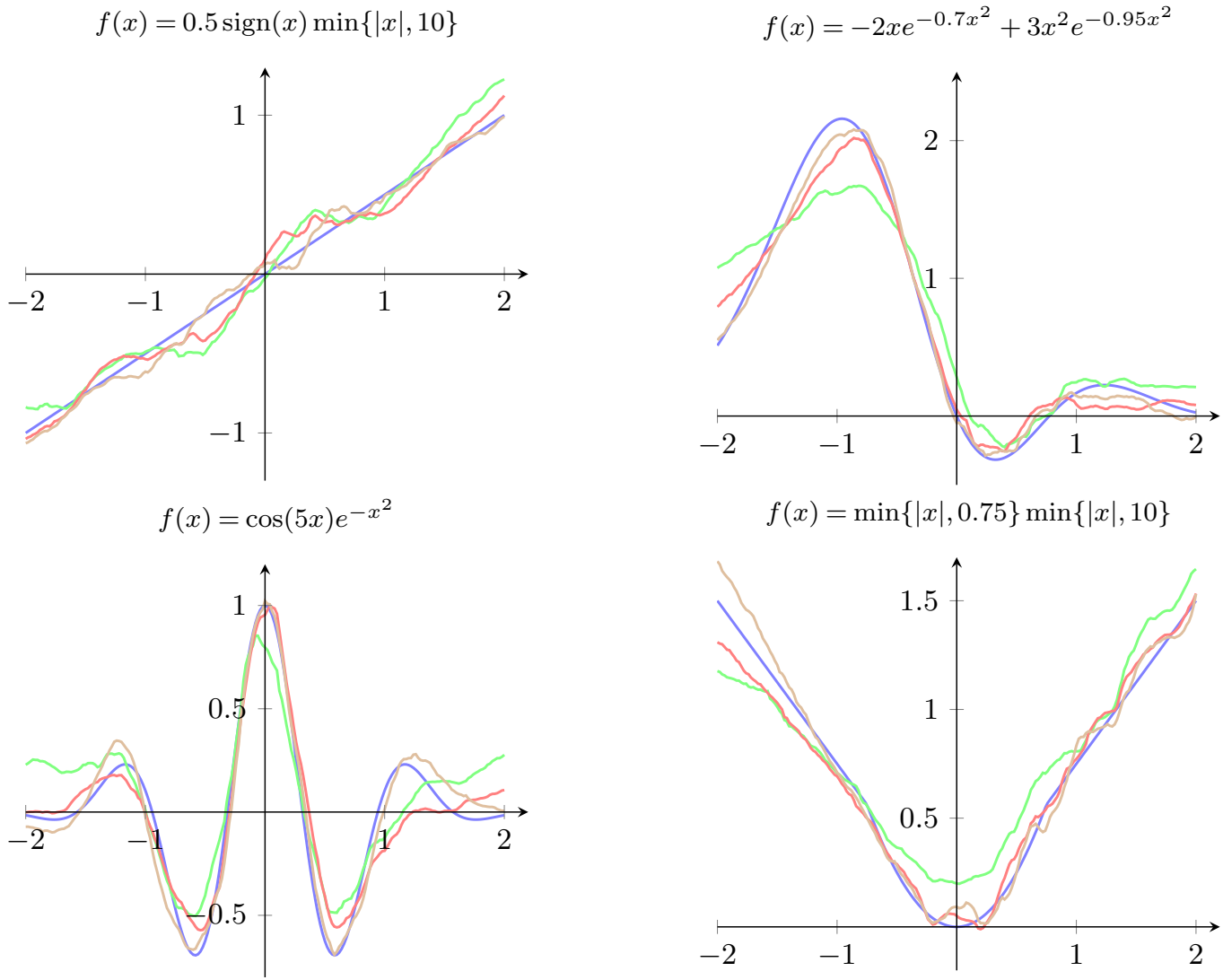


FIG 2. *The four specifications of f considered in (4.1) (blue) and the corresponding random forest estimator $\hat{f}_T$ based on sample sizes of $T = 400$ (green), $T = 1600$ (red) and $T = 6400$ (brown).*

Simulation study

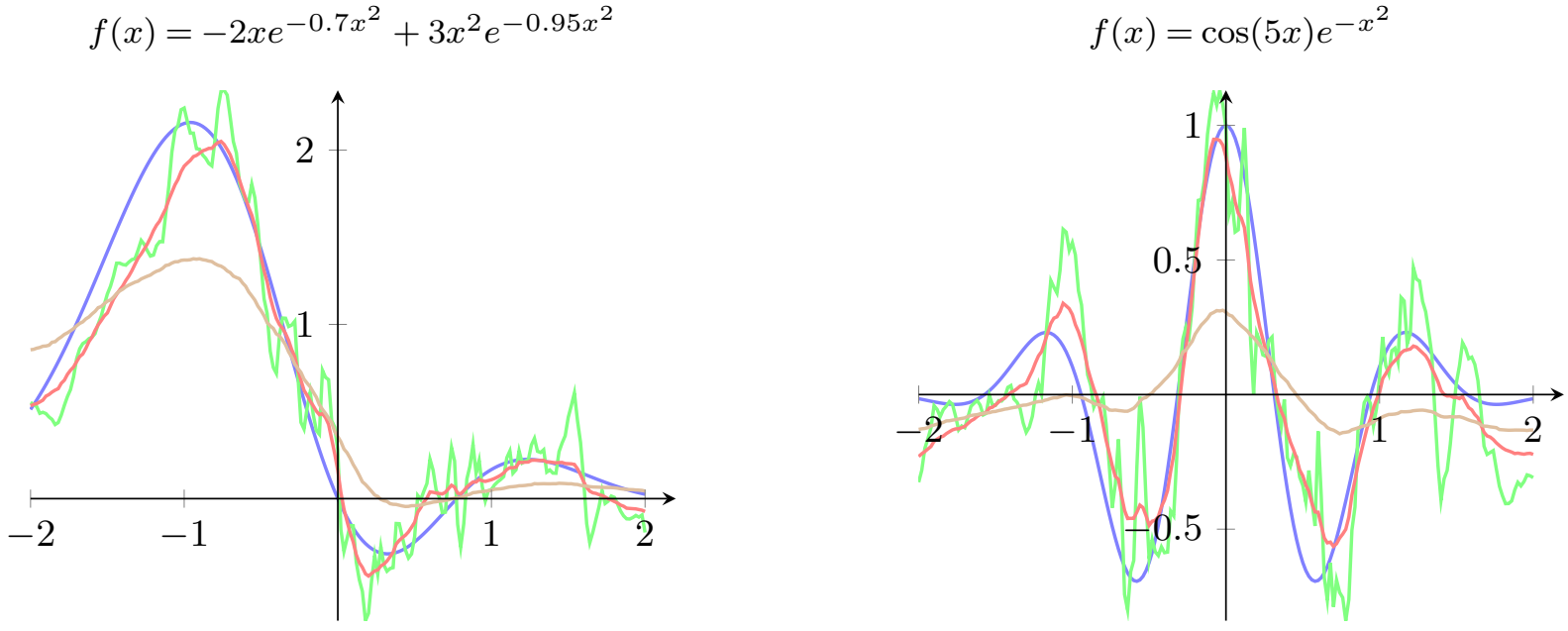


FIG 4. *Two of the specifications of f considered in (4.1) (blue) and the corresponding random forest estimator $\hat{f}_{1600}$ with $k = 40$ (green), $k = 160$ (red) and $k = 640$ (brown).*

Simulation study $p = 2$

Now we take $p = 2$ and

$$f(x_1, x_2) = x_1 e^{-0.6x_1^2} - 2(x_1^2 e^{-0.3x_1^2} + x_2 e^{-0.7x_2^2}) + 3x_2^2 e^{-0.95x_2^2}. \qquad (6)$$

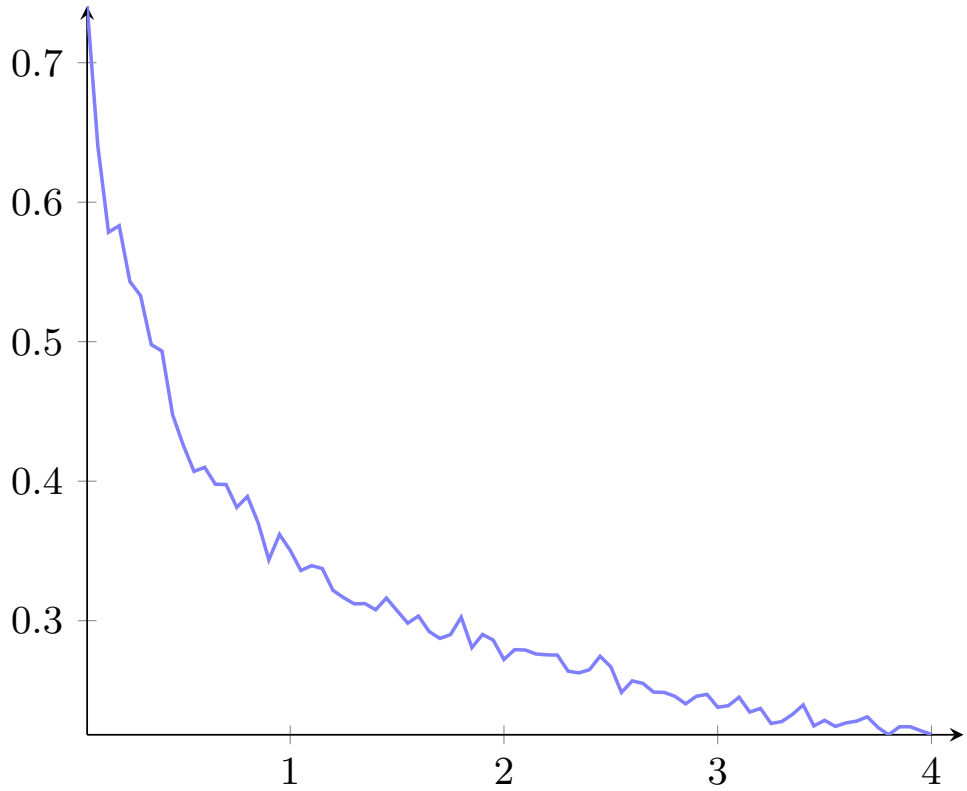


FIG 5. *The mean squared error* (4.3) *of the random forest estimator* $\hat{f}_T$ *as a function of* $10^{-4}T$ *when* f *is given by* (4.2).

Some references

- BIAU, G. (2012). Analysis of a random forests model. J
- BIAU, G. and DEVROYE, L. (2010). On the layered nearest neighbor estimate, the bagged nearest neighbor estimate and the random forest method in regression and classification.
- BIAU, G. and SCORNET, E. (2016). A random forest guided tour.
- BREIMAN, L. (1996). Bagging predictors.
- BREIMAN, L. (2001). Random forests
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). Classification and regression trees.
- MEINSHAUSEN, N. (2006). Quantile regression forests.
- MERLEVDE, F., PELIGRAD, M. and RIO, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions.
- WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using randomforests
- WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests.

Takeaway Message

Enjoy Luminy