S-CFE: Simple Counterfactual Explanations

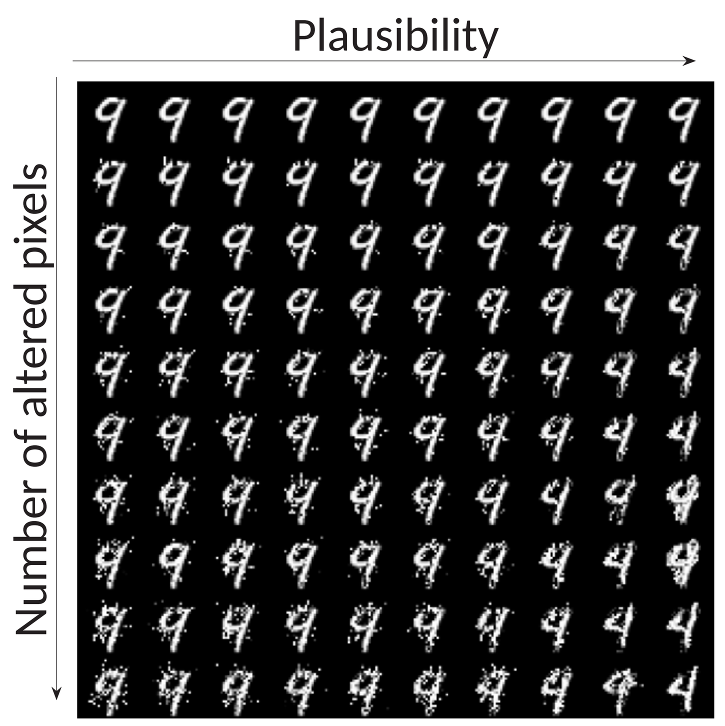
Shpresim Sadiku, Moritz Wagner, Sai Ganesh Nagarajan, Sebastian Pokutta

Cooperation: TU Berlin, ZIB
Funding: DFG Cluster of Excellence Math+, German Federal Ministry of Education and Research

Berlin Mathematics Research Center MATH+
ZIB ZUSE INSTITUTE BERLIN

Motivation

- **Opaque AI Decisions:** Machine learning models impact critical areas but lack transparency
- **Counterfactual Explanations (CFEs):** Show "what-if" changes needed to alter a model's decision
- **Basic Principles:** CFEs must be *Valid*, *Proximate*, and *Actionable*
- **Additionally,** *Plausible* and *Sparse* for realistic suggestions
- **Complex Optimization:** Finding CFEs requires solving complex mathematical problems with *non-convex* and *non-smooth* objectives



Background on CFEs

- Input space $\mathscr{X} \subseteq \mathbb{R}^d$, output space $\mathscr{Y}$
- Data $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y}\}_{i=1}^n$ generated from joint density $\psi : \mathscr{X} \times \mathscr{Y} \mapsto \mathbb{R}_+$
- Conditional input density $q(x,y) := \psi(x|y)$
- Classifier $f_l : \mathscr{X} \to \mathbb{R}^{|\mathscr{Y}|}$ and $f(x) := \arg\max_i [f_l(x)]_i$

Definition. *Given $x_f \in \mathbb{R}^d$ such that $f(x_f) = y_f$, its* ***closest sparse data-manifold CFE*** *with respect to $f(\cdot)$ and the data manifold of the target class y_{cf} is defined as $x_{cf} \in \mathscr{X}$ solving*

$$\begin{aligned} x_{cf} := \ & \arg\min_{x \in \mathscr{X}} \|x - x_f\|_2^2 \\ \text{s.t. } \ & x \in \mathscr{A} \\ & f(x) = y_{cf} \\ & q(x, y_{cf}) \geq \tau \\ & \|x - x_f\|_0 \leq m, \end{aligned} \tag{1}$$

where $\mathscr{A}$ denotes the value range for features, $m \in \mathbb{N}$ and $\tau > 0$.

The Need for Plausibility

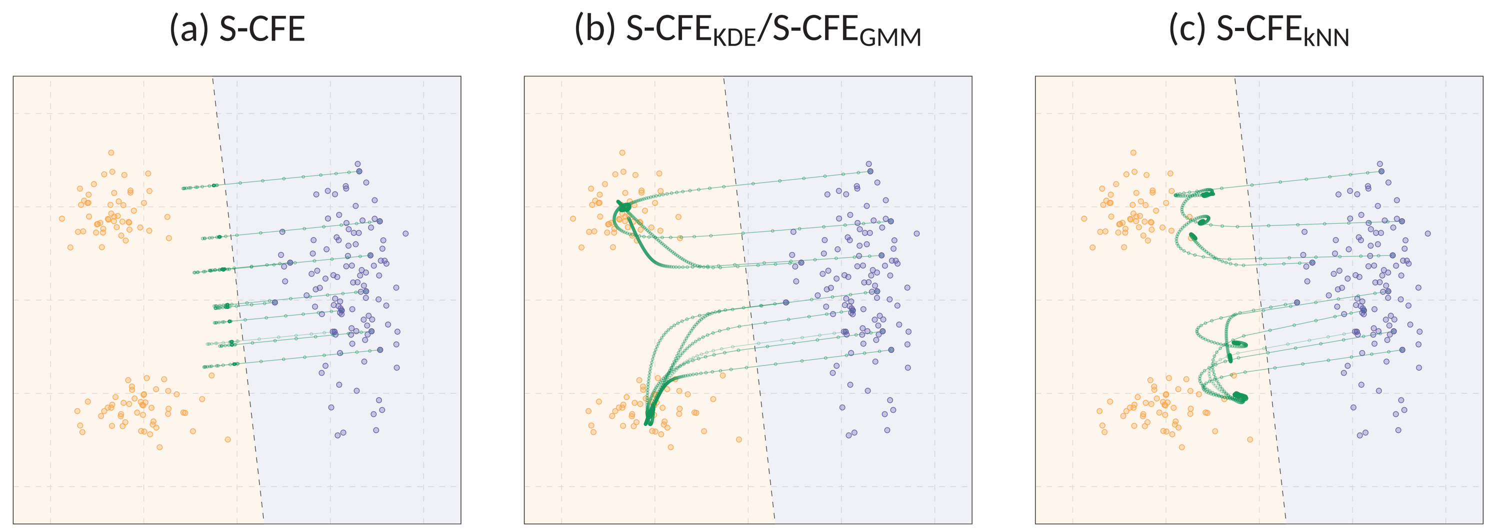


Figure 1: Synthetic 2D Gaussian dataset demonstrating (a) methods without a plausibility term vs. (b)-(c) methods combined with a plausibility term.

Results for DNN classifiers

Dataset	Method	Validity (std)	ℓ_2 (std)	ℓ_0 (std)	LOF (std)	Time
Housing 12 features	S-CFE$_{KDE}$	**100** (0.00)	**2.59** (1.21)	**2.00** (0.00)	1.23 (0.29)	12.7
	S-CFE$_{GMM}$	**100** (0.00)	2.91 (1.38)	**2.00** (0.00)	**1.12** (0.26)	13.3
	S-CFE$_{k\text{NN}}$	**100** (0.00)	3.64 (1.73)	**2.00** (0.00)	1.17 (0.31)	5.85
	DCFE	**100** (0.00)	3.50 (1.68)	6.86 (1.42)	1.27 (0.38)	**5.33**
	CEM	94.0 (0.23)	2.93 (2.23)	2.99 (1.17)	1.36 (0.60)	7.51
Wine 13 features	S-CFE$_{KDE}$	**100** (0.00)	3.31 (1.16)	**2.00** (0.00)	0.99 (0.01)	12.4
	S-CFE$_{GMM}$	**100** (0.00)	3.44 (1.09)	**2.00** (0.00)	**0.98** (0.02)	13.1
	S-CFE$_{k-\text{NN}}$	**100** (0.00)	4.04 (1.59)	**2.00** (0.00)	1.01 (0.07)	5.80
	DCFE	**100** (0.00)	**3.21** (2.70)	7.13 (1.31)	1.03 (0.18)	**4.95**
	CEM	92.0 (0.29)	5.40 (3.25)	5.14 (2.68)	1.07 (0.14)	5.71
MNIST 784 features	S-CFE$_{GMM}$	99.1 (0.09)	**6.74** (2.92)	**25.0** (0.00)	**1.21** (0.18)	55.3
	S-CFE$_{k-\text{NN}}$	**99.8** (0.04)	7.04 (2.99)	**25.0** (0.00)	1.30 (0.22)	13.1
	DCFE	99.3 (0.08)	8.06 (3.48)	118 (6.30)	1.32 (2.24)	**11.8**

Robustness of Plausible CFEs to Input Shifts

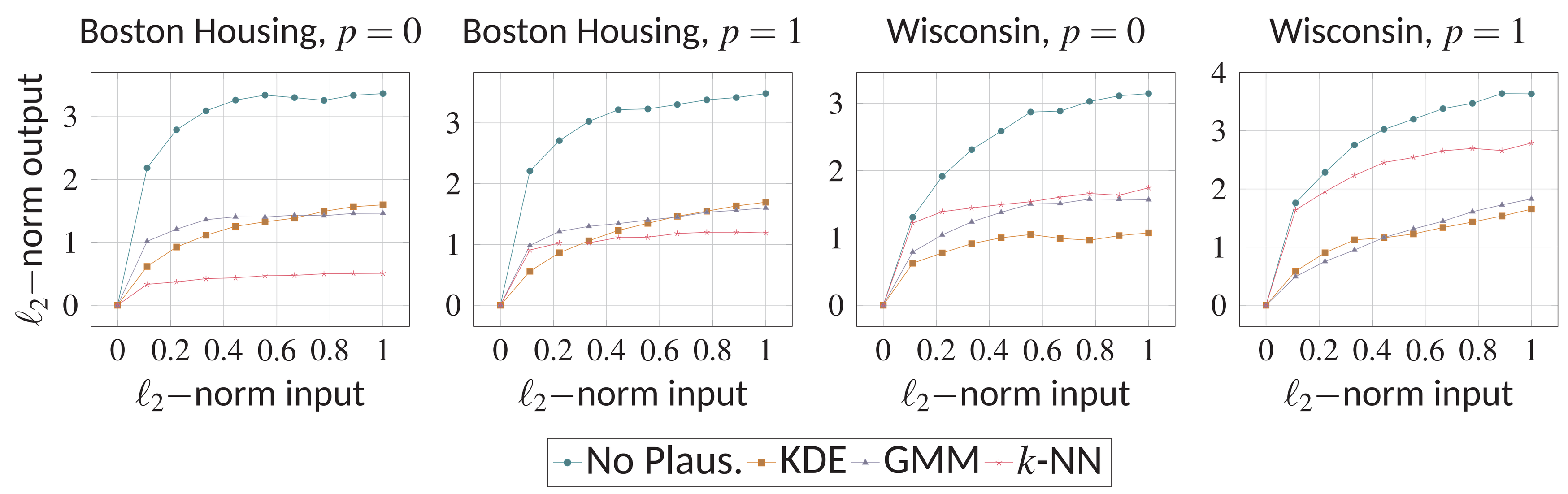


Figure 2: Robustness of the different methods. The distance of the input data points to the original data points on the x-axis and the distance of the generated CFEs to the CFE generated from the original data points on the y-axis. Tested on 100 data points from each data set.

A Simple Algorithm for Generating CFEs

Two main issues with solving Eq. (1)

1. Conditional distribution $q(\cdot, y)$ is unknown
   $\hookrightarrow$ Utilize plausibility constraints based on density estimates
2. $0-$norm for sparsity leads to NP-hard problems
   $\hookrightarrow$ Utilize accelerated proximal gradient (APG) method

- Replace validity, plausibility, and sparsity constraints with penalty terms; enforce actionability via indicator function

$$\begin{aligned} x_{cf} := \ & \arg\min_{x \in \mathbb{R}^d} \|x - x_f\|_2^2 + I_{\mathscr{A}}(x) + \gamma \mathscr{L}_f(x, y_{cf}) \\ & - \tau \hat{q}(x, y_{cf}) + \beta \|x - x_f\|_p^p \end{aligned}$$

- $\hat{q}(\cdot, y_{cf})$ - estimate for the density of target class y_{cf} in $\mathscr{X}$
- $\mathscr{L}_f$ - differentiable classification loss
- $h(x, y_{cf}) := \|x - x_f\|_2^2 + \gamma \mathscr{L}_f(x, y_{cf}) - \tau \hat{q}(x, y_{cf})$
  $\hookrightarrow$ Smooth non-convex function of Lipschitz constant L
  $\hookrightarrow$ Use differentiable density estimates such as $\hat{q}_{KDE}(x, y_{cf})$ and $\hat{q}_{GMM}(x, y_{cf})$ to compute the gradient
- $g_p(x) := I_{\mathscr{A}}(x) + \beta \|x - x_f\|_p^p$
- Solution is computed by solving

$$x_{cf}^{t+1} := \arg\min_{x \in \mathbb{R}^d} \frac{L}{2} \left\| x - \left(x^t - \frac{1}{L} \nabla_{x^t} h(x^t, y_{cf}) \right) \right\|_2^2 + g_p(x)$$

- Closed-form for $p \in \{0, 1/2, 2/3, 1\}$

Constraining the Sparsity

- Regularize sparsity using the indicator function

$$I_{\|x - x_f\|_p^p \leq m}(x) := \begin{cases} 0, & \text{if } \|x - x_f\|_p^p \leq m \\ +\infty, & \text{otherwise} \end{cases}$$

- Reframe the problem

$$\begin{aligned} x_{cf} := \ & \arg\min_{x \in \mathbb{R}^d} \|x - x_f\|_2^2 + I_{\mathscr{A}}(x) + \gamma \mathscr{L}_f(x, y_{cf}) \\ & - \tau \hat{q}(x, y_{cf}) + \beta I_{\|x - x_f\|_p^p \leq m}(x) \end{aligned}$$

- $g_p(x) := I_{\mathscr{A}}(x) + \beta I_{\|x - x_f\|_p^p \leq m}(x)$ is an indicator function
  $\hookrightarrow$ Solution for $p = 0$ coincides with the projection onto the intersection $\{\|x - x_f\|_0 \leq m\} \cap \mathscr{A}$
  $\hookrightarrow$ Convergence of APG to a critical point can be assured under some mild conditions