Interpretable Low-Dimensional Regression via Data-Adaptive Smoothing

Wesley Tansey, Jesse Thomason, James G. Scott

University of Texas at Austin

ICML

Human Intelligible Low-Dimensional Regression

We consider the problem of estimating a regression function in the common situation where the number of features is small, where interpretability of the model is a high priority, and where simple linear or additive models fail to provide adequate performance. To address this problem, we present Maximum Variance Total Variation denoising (MVTV). MVTV divides the feature space into blocks of constant value and fits the value of all blocks jointly via a convex optimization routine. Our method is fully data-adaptive, in that it incorporates highly robust routines for tuning all hyperparameters automatically.

State of the Art: CRISP

Petersen et al. (2016) propose Convex Regression with Interpretable Sharp Partitions (CRISP). They focus on the 2d scenario and divide the (x_1, x_2) space into a $q \times q$ grid via a data-adaptive procedure. CRISP applies a Euclidean penalty on the differences between adjacent rows and columns of M. The final estimator is then learned by solving the convex optimization problem,

$$\underset{M \in \mathbb{R}^{q \times q}}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - \Omega(M, x_{1i}, x_{2i}))^2 + \lambda P(M) \, , \qquad (1)$$

where Ω is a lookup function mapping (x_{1i}, x_{2i}) to the corresponding element in M. $P(M)$ is the group-fused lasso penalty on the rows and columns of M,

$$P(M) = \sum_{i=1}^{q-1} \left[||M_{i\cdot} - M_{(i+1)\cdot}||_2 + ||M_{\cdot i} - M_{\cdot (i+1)}||_2 \right] \, , \qquad (2)$$

where $M_{i\cdot}$ and $M_{\cdot i}$ are the i^{th} row and column of M, respectively.

Maximum Variance Total Variation Denoising

Problem reformulation. We rewrite (1) as a weighted least-squares problem,

$$\underset{\beta \in \mathbb{R}^{q^2}}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{q^2} \eta_i (\tilde{y}_i - \beta_i)^2 + \lambda g(\beta) \, , \qquad (3)$$

where $\beta = \text{vec}(M)$ is the vectorized form of M, η_i is the number of observations in the i^{th} cell, and $\tilde{y}_i$ is the empirical average of the observations in the i^{th} cell. $g(\cdot)$ is a penalty term that operates over a vector β rather than a matrix M.

Graph TV. We choose $g(\cdot)$ to be a graph-based total variation penalty,

$$g(\beta) = \sum_{(r,s) \in \mathcal{E}} |\beta_r - \beta_s| \, , \qquad (4)$$

where $\mathcal{E}$ is the set of edges defining adjacent cells on the $q \times q$ grid graph. Having formulated the problem as a graph TV denoising problem, we can now use the convex minimization algorithm of Barbero and Sra (2014) to efficiently solve (3).

Maximum variance q selection. The main challenge in our problem is to adaptively choose q to fit the appropriate level of overall data sparsity. We do this by choosing the grid which maximizes the sum of variances of all cells:

$$q = \underset{q}{\text{argmax}} \sum_{c \in \mathcal{C}(q)} \widehat{\text{var}}(\mathbf{y}_c) \, , \qquad (5)$$

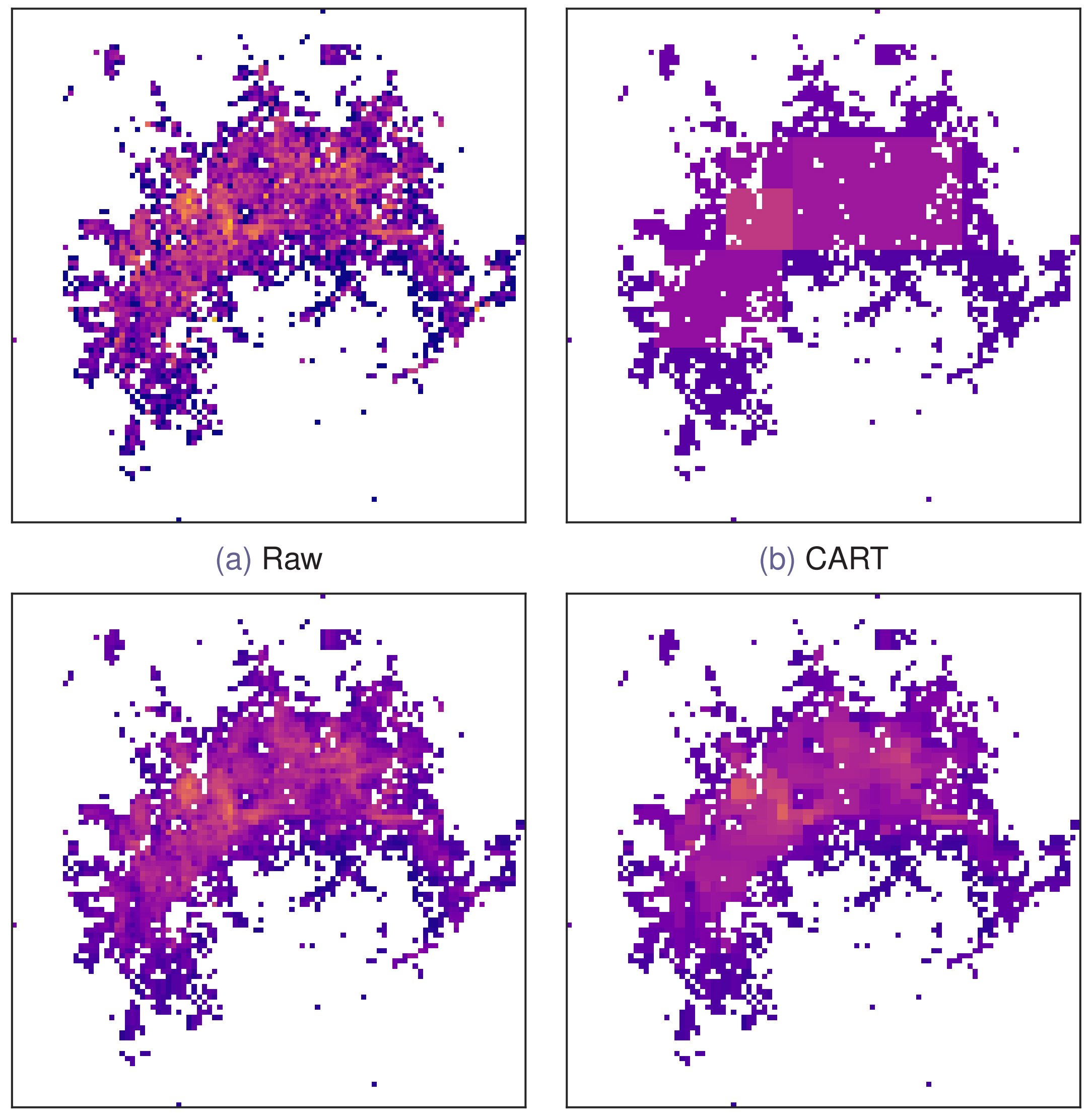
where $\mathcal{C}(q)$ is the set of cells in the $q \times q$ grid and $\text{var}(\emptyset) = 0$.

References

Petersen, Ashley, Simon, Noah, and Witten, Daniela. Convex regression with interpretable sharp partitions. Journal of Machine Learning Research, 17(94):1–31, 2016.

Barbero, Álvaro and Sra, Suvrit. Modular proximal optimization for multidimensional total-variation regularization. arXiv:1411.0589, 2014.

Case Study: Austin Crime Data



(a) Raw

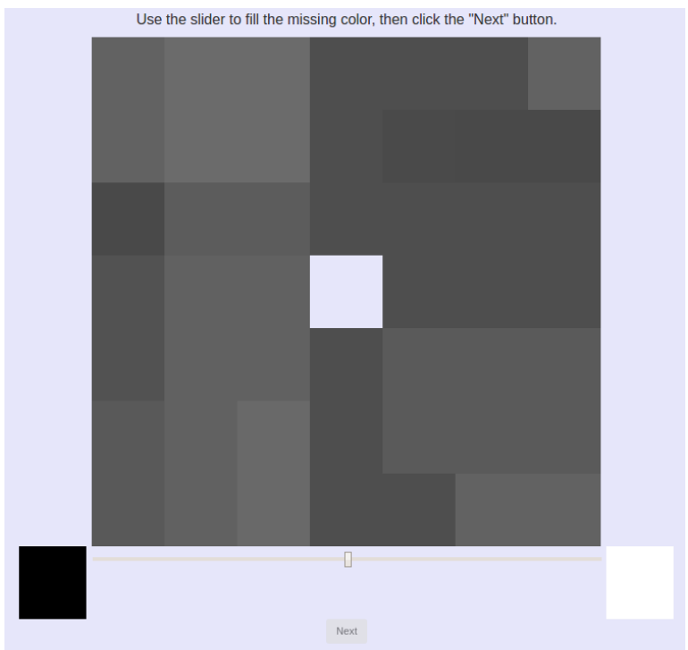
(b) CART

(c) CRISP

(d) MVTV

Quantitative Evaluation.
MVTV outperforms both CART and CRISP in terms of AIC, where degrees of freedom is the number of plateaus of constant value.

	Austin Crime Data	
	AIC	Human error $\times 10^{-2}$
CART	11139.29	3.24±0.341
CRISP	18326.33	3.99±0.664
MVTV	**10327.58**	**2.75**±0.334



Use the slider to fill the missing color, then click the "Next" button.

Human Interpretability Evaluation.
Mechanical Turk study with human annotators asked to choose a grayscale value for a held-out cell in the center of a 7 × 7 patch of data. Each annotator was shown a patch as rendered by MVTV, CART, CRISP, and as raw data; each task involved two randomly sampled patches from the Austin crime dataset ($4 \times 2 = 8$ patches per annotator, shown in random order).