The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Maximum-Variance Total Variation Denoising for Interpretable Spatial Smoothing

Wesley Tansey[*]
Columbia University
New York, NY 10027
wt2274@cumc.columbia.edu

Jesse Thomason,[†] **James G. Scott**[‡§]
University of Texas at Austin
Austin, TX 78712
jesse@cs.utexas.edu
james.scott@mccombs.utexas.edu

Abstract

We consider the problem of spatial regression where interpretability of the model is a high priority. Such problems appear frequently in a diverse set of fields from climatology to epidemiology to predictive policing. For cognitive, logistical, and organizational reasons, humans tend to infer regions or neighborhoods of constant value, often with sharp discontinuities between regions, and then assign resources on a per-region basis. Automating this smoothing process presents a unique challenge for spatial smoothing algorithms, which tend to assume stationarity and smoothness everywhere.

To address this problem, we propose Maximum Variance Total Variation (MVTV) denoising, a novel method for interpretable nonlinear spatial regression. MVTV divides the feature space into blocks of constant value and smooths the value of all blocks jointly via a convex optimization routine. Our method is fully data-adaptive and incorporates highly robust routines for tuning all hyperparameters automatically. We compare our approach against the existing CART and CRISP methods via both a complexity-accuracy tradeoff metric and a human study, demonstrating that that MVTV is a more powerful and interpretable method.

Introduction

Many modern machine learning techniques, such as deep learning and kernel machines, tend to focus on the "big data, big features" regime. In such a scenario, there are often so many features—and highly non-linear interations between features—that model interpretability is a secondary consideration. Instead, effort is focused solely on a measure of model performance such as root mean squared error (RMSE). Under that research paradigm, only a model that out-performs the previous champion method warrants an investigation into understanding its decisions.

By contrast, there is a parallel recent line of machine-learning research in interpretable low-dimensional regression, with relatively few features and with human intelligibility as a primary concern. For example, lattice regression with monotonicity constraints has been shown to perform well in video-ranking tasks where interpretability was a prerequisite (Gupta et al. 2016). The interpretability of the system enables users to investigate the model, gain confidence in its recommendations, and guide future recommendations. In the two- and three- dimensional regression scenario often found in spatiotemporal data, the Convex Regression via Interpretable Sharp Partitions (CRISP) method (Petersen, Simon, and Witten 2016) has recently been introduced as a way to achieve a good trade off between accuracy and interpretability by inferring sharply-defined 2-dimensional rectangular regions of constant value. Such a method is readily useful, for example, when making business decisions or executive actions that must be explained to a non-technical audience.

Data-adaptive, interpretable sharp partitions are also useful in the creation of areal data from a set of spatial point-referenced data—turning a continuous spatial problem into a discrete one. A common application of the framework arises when dividing a city, state, or other region into a set of contiguous cells, where values in each cell are aggregated to help anonymize individual demographic data and create well-defined neighborhoods for resource allocation. Ensuring that the number and size of grid cells remains tractable, handling low-data regions, and preserving spatial structure are all important considerations for this problem. Ideally, one cell should contain data points which all map to a similar underlying value, and cell boundaries should represent significant change points in the value of the signal being estimated. If a cell is empty or contains a small number of data points, the statistical strength of its neighbors should be leveraged to both improve the accuracy of the reported areal data and further aid in anonymizing the cell, which may otherwise be particularly vulnerable to deanonymization. Viewed through this lens, we can interpret the areal-data creation task as a machine learning problem, one focused on finding sharp partitions that still achieve acceptable predictive loss.[1]

In this paper we propose MVTV: a method for inter-

[*]Department of Systems Biology
[†]Department of Computer Science
[‡]Department of Information, Risk, and Operations Management
[§]Department of Statistics and Data Sciences

[1]We note that such a task will likely only represent a single step in a larger anonymization pipeline that may include other techniques such as additive noise and spatial blurring. While we provide no proofs of how strong the anonymization is for our method, we believe it is compatible with other methods that focus on adherence to a specified k-anonymity threshold (e.g., (Cassa et al. 2006)).

pretable, low-dimensional convex regression with sharp partitions. MVTV involves two main steps: (1) an aggregate variance maximization procedure that creates a data-adaptive grid over the feature space; and (2) smoothing over this grid using a fast total variation denoising algorithm (Barbero and Sra 2014). The resulting model displays a good balance between four key measurements: (1) interpretability, (2) average accuracy, (3) worst-region accuracy, and (4) degrees of freedom. Through a series of benchmarks against both a baseline CART model and the state-of-the-art CRISP model, we show both qualitatively and quantitatively that MVTV achieves superior performance. Finally, we conduct a human study on the predictive interpretability of each method. Our results show that humans are better able to understand the predictions made by MVTV, as measured by their ability to intuit the surrounding spatial context of a smoothed region to predict the true underlying data. The end result is thus a fast, fully auto-tuned approach to interpretable low-dimensional regression and classification.

Background

Both CRISP and MVTV formulate the interpretable regression problem as a regularized convex optimization problem. We first give a brief overview of the CRISP loss function and its computational complexity. We then give a brief preliminary overview of total variation denoising, the approach used by MVTV.

Convex Regression with Interpretable Sharp Partitions

Recently work introduced and motivated the problem of low-dimensional smoothing via constant plateaus (Petersen, Simon, and Witten 2016). As in our approach, their CRISP algorithm focuses on the 2d scenario and divides the (x_1, x_2) space into a grid via a data-adaptive procedure. For each dimension, they divide the space into q regions, where each region break is chosen such that a region contains $\frac{1}{q}$ of the data. This creates a $q \times q$ grid of differently-sized cells, some of which may not contain any observations. A prediction matrix $M \in \mathbb{R}^{q \times q}$ is then learned, with each element M_{ij} representing the prediction for all observations in the region specified by cell (i, j).

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:

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

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, and

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

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

By rewriting $\Omega(\cdot)$ as a sparse binary selector matrix and introducting slack variables for each row and column in the $P(M)$ term, CRISP solves (1) via the Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2011). The resulting algorithm requires an initial step of $\mathcal{O}(n + q^4)$ operations for n samples on a $q \times q$ grid, and has a per-iteration complexity of $\mathcal{O}(q^3)$. The authors recommend using $q = n$ when the size of the data is sufficiently small so as to be computationally tractable, and setting $q = 100$ otherwise.

In comparison to other interpretable methods, such as CART (Breiman et al. 1984) and thin-plate splines (TPS), CRISP is shown to yield a good tradeoff between accuracy and interpretability.

Graph-based Total Variation Denoising

Total variation (TV) denoising solves a convex regularized optimization problem defined generally over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{V}|}}{\text{minimize}} \quad \sum_{s \in \mathcal{V}} \ell(y_s, \beta_s) + \lambda \sum_{(r,s) \in \mathcal{E}} |\beta_r - \beta_s|\,, \tag{3}$$

where ℓ is some smooth convex loss function over the value at a given node β_s. The solution to (3) yields connected subgraphs (i.e. plateaus in the 2d case) of constant value. TV denoising has been shown to have attractive minimax rates theoretically (Sadhanala, Wang, and Tibshirani 2016) and is robust against model mispecification empirically, particularly in terms of worst-cell error (Tansey et al. 2016).

Many efficient, specialized algorithms have been developed for the case when ℓ is a Gaussian loss and the graph has a specific constrained form. For example, when $\mathcal{G}$ is a one-dimensional chain graph, (3) is the ordinary (1d) fused lasso (Tibshirani et al. 2005), solvable in linear time via dynamic programming (Johnson 2013). When $\mathcal{G}$ is a d-dimensional grid graph, (3) is typically referred to as total variation denoising (Rudin, Osher, and Faterni 1992) or the graph-fused lasso, for which several efficient solutions have been proposed (Chambolle and Darbon 2009; Barbero and Sra 2011; 2014). For scenarios with a general smooth convex loss and an arbitrary graph, the GFL method (Tansey et al. 2017) is efficient and easily extended to non-Gaussian losses such as the binomial loss.

The TV denoising penalty was investigated as an alternative to CRISP in past work (Petersen, Simon, and Witten 2016). They note anecdotally that TV denoising over-smooths when the same q was used for both CRISP and TV denoising. We next present a maximum-variance criterion for choosing q in a data-adaptive way that prevents such over-smoothing and leads to a superior fit in terms of the accuracy-complexity tradeoff.

The MVTV Algorithm

We first note that we can rewrite (1) as a weighted least-squares problem,

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{q^2}}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^{q^2} \eta_i (\tilde{y}_i - \beta_i)^2 + \lambda g(\boldsymbol{\beta})\,, \tag{4}$$

where $\boldsymbol{\beta} = \text{vec}(M)$ 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 $\boldsymbol{\beta}$ rather than a matrix M.

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

$$g(\boldsymbol{\beta}) = \sum_{(r,s)\in\mathcal{E}} |\beta_r - \beta_s| \,, \tag{5}$$

where $\mathcal{E}$ is the set of edges defining adjacent cells on the $q \times q$ grid graph.[2] Having formulated the problem as a graph TV denoising problem, we can now use an existing convex minimization algorithm (Barbero and Sra 2014) (or any other suitable algorithm) to efficiently solve (4).

We auto-tune the two hyperparameters: q, the granularity of the grid, and λ, the regularization parameter. We take a pipelined approach by first choosing q and then selecting λ under the chosen q value.

Choosing bins via a maximum variance heuristic

The recommendation for CRISP is to choose $q = n$, assuming the computation required is feasible. Doing so creates a very sparse grid, with $q - 1 \times q$ empty cells. However, by tying together the rows and columns of the grid, each CRISP cell actually draws statistical strength from a large number of bins. This compensates for the data sparsity problem and results in reasonably good fits despite the sparse grid.

Choosing $q = n$ does not work for our TV denoising approach. Since the graph-based TV penalty only ties together adjacent cells, long patches of sparsity overwhelm the model and result in over-smoothing. If one instead chooses a smaller value of q, however, the TV penalty performs quite well. The challenge is therefore to adaptively choose q to fit the appropriate level of overall data sparsity. We do this by choosing the grid maximizing the sum of variances of all cells:

$$q = \underset{q}{\text{argmax}} \sum_{c\in\mathcal{C}(q)} \hat{\text{var}}(\mathbf{y}_c) \,, \tag{6}$$

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

Choosing the grid is a tradeoff between each cell's fit to the data and the total number of cells. Each sample y_i is assumed to be IID conditioned on being in the same cell. We find that maximizing the sum of variances as in (6) serves as a useful heuristic for finding cells that fit well to the distribution of the data and prevent overfitting by using too many cells.

A clear connection also exists between our heuristic and principal components analysis (PCA). Since we are dealing with univariate observations, maximizing the variance corresponds to finding the approximation to the first principal component of the data. The TV penalty then helps to smooth over these principal components by incorporating the spatial adjacency information. Such a connection presents the possiblity for future extensions to multivariate observations and smoothing using group TV methods like the network lasso (Hallac, Leskovec, and Boyd 2015).

[2]Though our goal in this work is not to increase the computational efficiency of existing methods, we do note that CRISP can be solved substantially faster via the reformulation in (4). The weighted least squares loss enables a much more efficient solution to (1) via a simpler ADMM solution similar to the network lasso (Hallac, Leskovec, and Boyd 2015).

Classification extension

The optimization problem in (4) focuses purely on the Gaussian loss case. When the observations are binary labels, as in classification, a binomial loss function is a more appropriate choice. The binomial loss case specifically has been derived in previous work (Tansey et al. 2016) and shown to be robust to numerous types of underlying spatial functions. Therefore, unlike CRISP, the inner loop of our method immediately generalizes to the non-Gaussian scenario, with only minor modifications. Extensions to any other smooth, convex loss are similarly straightforward.

Choosing the TV penalty parameter

Once a value of q has been chosen, λ can be chosen by following a solution path approach. For the regression scenario with a Gaussian loss, as in (4), determining the degrees of freedom is well studied (Tibshirani and Taylor 2011). Thus, we could select λ via an information criterion such as AIC (Sakamoto, Ishiguro, and Kitagawa 1986). We choose to select λ via cross-validation as we found empirically that it produces better results both in terms of subjective interpretability and overall AIC.

Experiments

We compare results on a suite of both synthetic and real-world datasets. We first compare MVTV against two benchmark methods with sharp partitions, CART and CRISP, on a synthetic dataset with varying sample sizes. We also compare against CRISP with q fixed at the maximum variance solution in a method we call MV-CRISP. We show that the MVTV method leads to better Akaike information criterion (AIC) scores. We then demonstrate the advantage of the maximum variance criterion by showing that it chooses grid sizes that offer a good trade-off between average and worst-cell accuracy. Finally, we test all four methods against two real-world datasets of crime reports for Austin and Chicago. A human evaluation on the results for Austin shows that the MV* methods are most interpretable.

Synthetic Benchmark

We generated 100 independent 100×100 grids, each with six 1000-point plateaus. Each plateau was generated via a random walk from a randomly chosen start point and the means of the plateaus were -5, -3, -2, 2, 3, and 5; all points not in a plateau had mean zero. For each grid, we sampled points uniformly at random with replacement and added Gaussian noise with unit variance. Figure 1 shows an example ground truth for the means. Sample sizes explored for each grid were 50, 100, 200, 500, 1K, 2K, 5K, and 10K. For each trial, we evaluate the CART method from the R package `rpart`, CRISP, and the MV* methods. For CRISP, we use $q = \max(n, 100)$ as per the suggestions in (Petersen, Simon, and Witten 2016); for the MV* methods, we use the maximum variance criterion to choose from $q \in [2, 50]$. For both CRISP and the

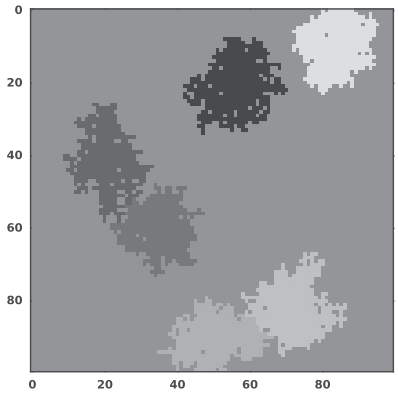


Figure 1: An example 100×100 grid of ground truth means ranging from -5 to 5. Each grid has six randomly-generated plateaus of raised or lowered means from the background mean (zero); darker colors correspond to regions of higher value.

MV* methods, we chose λ via 5-fold cross validation across a log-space grid of 50 values.

In order to quantify model complexity, we calculate the number of constant-valued plateaus in each model. Intuitively, this captures the notion of "sharpness" of the partitions by penalizing smooth partitions for their visual blurriness. Statistically, this corresponds directly to the degrees of freedom of a TV denoising model in the unweighted Gaussian loss scenario (Tibshirani and Taylor 2011). Thus for all of our models this is only an approximation to the degrees of freedom. Nonetheless, we find the plateau-counting heuristic to be a useful measurement of the *visual* degrees of freedom. Finally, to quantify the trade-off of accuracy and complexity, we use the AIC with the plateau count as the degrees of freedom surrogate.

Figure 2 shows qualitative results for the four smoothing methods as the sample size grows from 100 to 2000. CART (a-c) tends to over-smooth, leading to very sharp partitions that are too coarse grained to produce accurate results even as the sample size grows large. On the other hand, CRISP (d-f) under-smooths by creating very blurry images. The MV-based version of CRISP (g-i) alleviates this in the low-sample cases, but tying across entire rows and columns causes the image to blur as the data increases. The MVTV method (j-l) achieves a reasonable balance here by producing large blocks in the low-sample setting and progressively refining the blocks as the sample size increases, without substantially compromising the sharpness of the overall image.

Finally, Figure 3 shows the quantitative results of the experiments, averaged over the 100 trials. The CRISP and MV* methods perform similarly in terms of RMSE (Figure 3a), but both CRISP methods create drastically more plateaus. In the case of the original CRISP method, it quickly approaches one plateau per cell (i.e., completely smooth) as denoted by the dotted red horizontal line in Figure 3b. MVTV also presents a better trade-off point as measured by AIC (Figure 3c). Using the data-adaptive q value chosen by our maximum variance criterion helps improve the AIC scores in the low-sample regime, but as samples grow the MV-CRISP method begins to under-smooth by creating too many plateaus. This demonstrates that it is not merely the size of the grid, but also our choice of TV-based smoothing that leads to strong results.

Maximum Variance Criterion Evaluation

To understand the effect of the maximum variance criterion, for each MVTV trial and sample size, we exhaustively solved the graph TV problem for a finely discretized grid of values of q in the range $[2, 50]$. Figure 4 shows how the choice of q impacts the average RMSE and maximum point error for three different sample sizes; the dotted vertical red line denotes the value selected by the MV criterion. As expected, when the sample size is small, the MV criterion selects much smaller values; as the sample size grows, the MV criterion selects progressively larger q values. This enables the model to smooth over increasingly finer-grained resolutions.

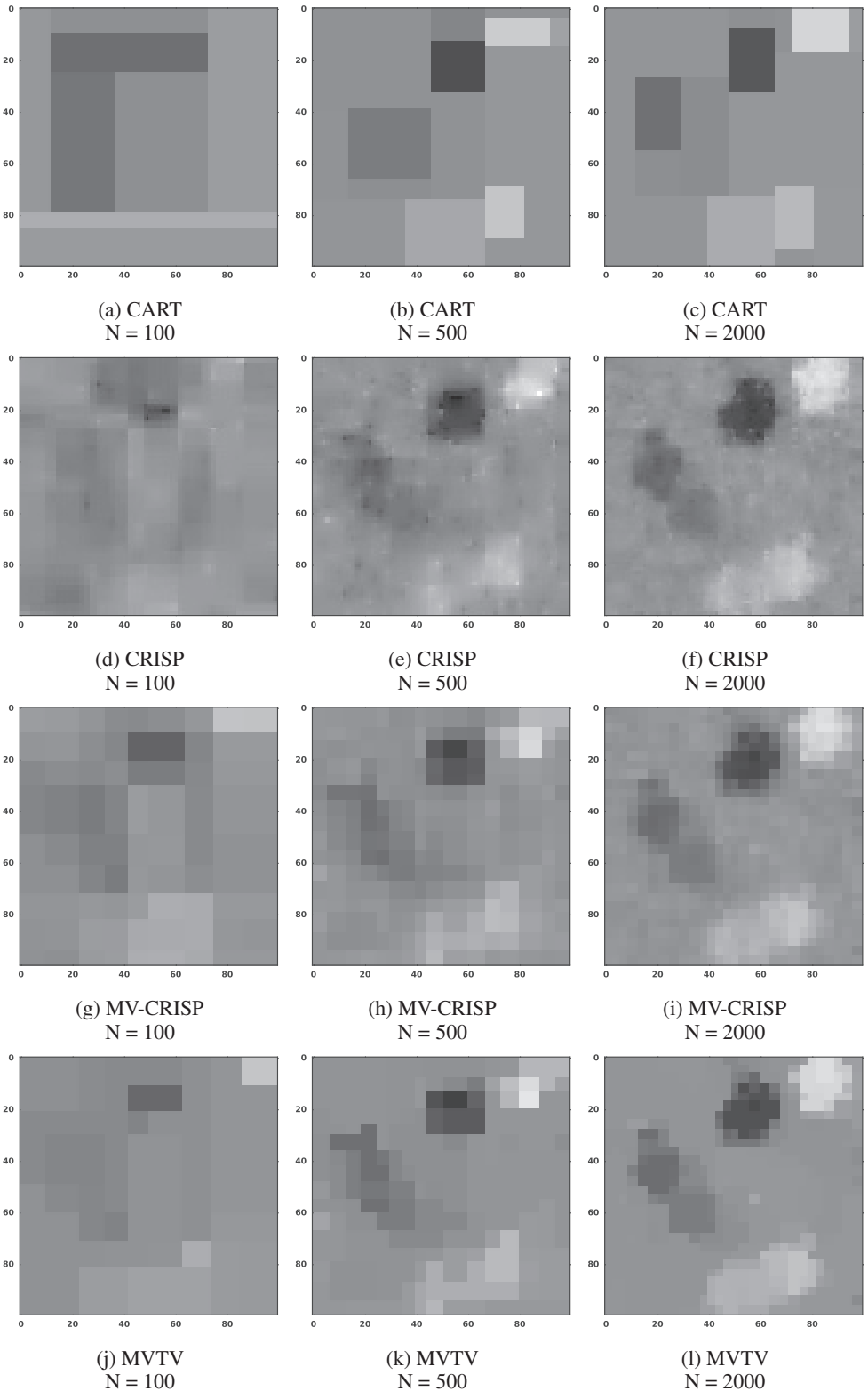
Perhaps counter-intuitively, the MV criterion is *not* choosing the q value which will simply minimize RMSE. As the middle panel shows, the MV criterion may actually choose one of the worst possible q values from this perspective. Instead, the resulting model is identifying a good trade-off between average accuracy (RMSE) and worst-case accuracy (max error). In small-sample scenarios like Figure 4a, RMSE is not substantially impacted by having a very coarse-grained q. Thus this trade-off helps prevent over-smoothing in the small sample regime—a problem observed previously when using TV with a large q (Petersen, Simon, and Witten 2016). But as the data grows (Figure 4b), both overly fine and overly coarse grids may have problems, with the latter creating the potential for the TV method to under-smooth, similar to how CRISP performed in the synthetic benchmarks. Once sample sizes become relatively large (Figure 4c), a very fine-grained grid poses less risk of under-smoothing. The MV criterion here prevents selecting low q, which would lead to a much higher-variance estimate.

Austin and Chicago Crime Data

We applied all four methods to a dataset of publicly-available crime report counts[3] in Austin, Texas in 2014 and Chicago, Illinois in 2015. To preprocess the data, we binned all observations into a fine-grained 100×100 grid based on latitude and longitude, then took the log of the total counts in each cell. Points with zero observed crimes were omitted from the dataset as it is unclear whether they represented the absence of crime or a location outside the boundary of the local police department. Figure 5 (a) shows raw data for Austin.

The maximum variance methods used q values in the range $[2, 100]$ and the CRISP method used $q = 100$. We ran a 20-fold cross-validation to measure RMSE and calculated plateaus with a fully-connected grid (i.e., as if all adjacent pixels were connected) which we then projected back to the real data for every non-missing point. Figure 5 shows the qualitative results for CART (b), CRISP (c), and MVTV (d). MV-CRISP is omitted as it adds little insight. The CART model clearly over-smooths by dividing the entire city into

[3] https://www.data.gov/open-gov/



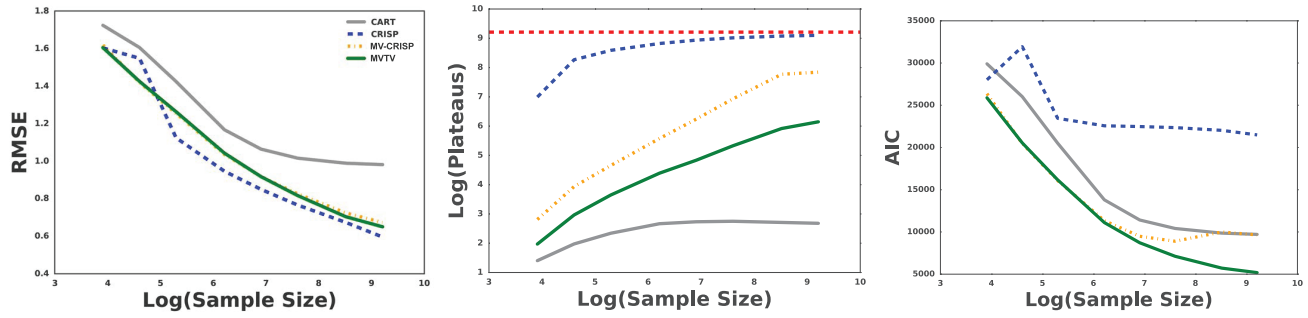
(a) CART N = 100 (b) CART N = 500 (c) CART N = 2000

(d) CRISP N = 100 (e) CRISP N = 500 (f) CRISP N = 2000

(g) MV-CRISP N = 100 (h) MV-CRISP N = 500 (i) MV-CRISP N = 2000

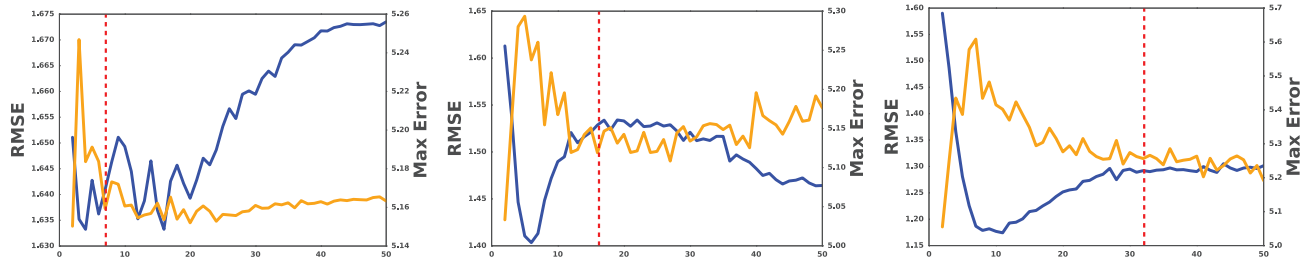
(j) MVTV N = 100 (k) MVTV N = 500 (l) MVTV N = 2000

Figure 2: Qualitative examples of each method's performance as the sample size increases on the ground truth from Figure 1. The MVTV method achieves a good subjective balance between limiting the number of plateaus and flexibly modeling the data, as supported quantitatively in Figure 3.



(a) RMSE (b) Plateau Count (c) Approximate AIC

Figure 3: Performance of the four methods as the sample size increases for the example grid in Figure 2 (a). While CRISP, MV-CRISP, and MVTV achieve similar sample efficiency in terms of RMSE scores (a), CRISP and MV-CRISP do so with drastically more change points (b); the dashed red horizontal line marks the maximum number of plateaus possible. Using AIC as a trade-off measurement (c), both MV* methods initially perform similarly but as the sample size (and thus the size of q) grows, the MVTV method continues to improve while the MV-CRISP method begins to over-smooth.



(a) N = 100 (b) N = 500 (c) N = 2000

Figure 4: RMSE (blue) and maximum error (orange) for the MVTV method for different sizes of the grid (q^2) for three different sample sizes; the dashed vertical red line indicates the value of q chosen by the maximum variance criterion. The results demonstrate that the MV criterion chooses models which provide a balance between average and worst-case error.

huge blocks of constant plateaus; conversely, CRISP under-smooths and creates too many regions. The MVTV method finds an appealing visual balance, creating flexible plateaus that partition the city well. These results are confirmed quantitatively in Table 1, where MVTV outperforms the three other methods in terms of AIC on both datasets.

Austin Crime Data Human Evaluation

To evaluate the interpretability of the MV* methods against the benchmark CART and CRISP methods, we ran a Mechanical Turk study with human annotators. To smooth noisy data in an interpretable way, huamns should be able to extrapolate information from local patterns in the smoothed data. To test this, we explore holding out information and determining to what degree humans can guess that missing information from smoothed neighboring data; that is, the ability to guess the true underlying raw value given surrounding smoothed values. We hypothesized that, in the absence of smoothing, neighboring information would be too noisy to predict the missing data. Further, we believed over-smoothing methods like CART would provide too little neighboring information to inform missing data, while under-smoothing methods like CRISP would create problems similar to trying to guess held-out information from raw neighbors.

To this end, we designed an annotation task: choose a grayscale value for a held-out cell in the center of a 7×7 patch of smoothed (or raw) data (Figure 6). Each annotator was shown a patch as rendered by MVTV, MV-CRISP, CART, CRISP, and as raw data. Each annotator saw two randomly sampled patches from the Austin crime dataset under each method ($5 \times 2 = 10$ patches per annotator, shown in random order). We added two additional uniformly-colored validation patches, placed randomly among the 10 real patches ($5 \times 2 + 2 = 12$ real plus validation patches per HIT). We discarded data from annotators who were not within 10% of the uniform value in these validation patches, because they either did not understand the task, were trying to complete it minimally (just touching the slider and clicking 'Next'), or were automated agents advancing through the task.

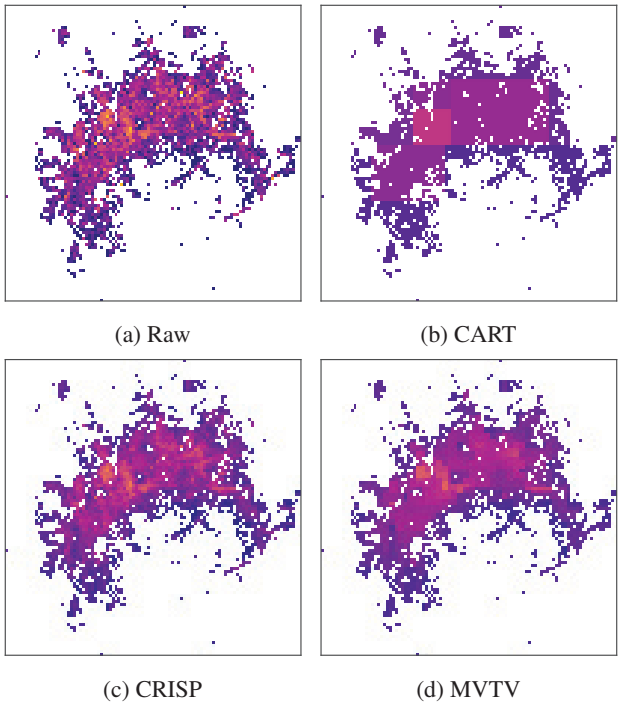
In this way, we gathered information from 207 annotators for 190 patches, throwing out an additional 37 annotators who failed validation. We measured the squared difference between the average annotators' predictions per (patch, method) combination against the true value in the raw data, shown in Table 1 (rightmost column).

All of our hypotheses were borne out in the resulting an-

	Chicago Crime Data		
	RMSE	Plateaus	AIC
(raw)	-	-	-
CART	1.04	**9.25**	43804.69
CRISP	**0.84**	9330.60	47245.57
MV-CRISP	0.85	8278.90	45314.71
MVTV	0.86	2270.15	**34016.60**

	Austin Crime Data			
	RMSE	Plateaus	AIC	Human error $\times 10^{-2}$
(raw)	-	-	-	4.71±0.539
CART	1.05	**10.40**	11139.29	3.24±0.341
CRISP	**0.94**	4699.15	18326.33	3.99±0.664
MV-CRISP	0.96	1361.75	12064.25	**2.72**±0.355**
MVTV	0.97	384.35	**10327.59**	**2.75**±0.334**

Table 1: Results for the four methods on crime data for Chicago and Austin. The MVTV method achieves the best trade-off between accuracy and the number of constant regions, as measured by AIC. The MV* methods also produce human annotator predictions that are statistically significantly closer than when annotators are shown raw data for the Austin human experiment, which neither CART nor CRISP achieve.



(a) Raw (b) CART

(c) CRISP (d) MVTV

Figure 5: Areal data results for the Austin crime data. The maps show the raw fine-grained results (a) and the results of the three main methods. Qualitatively, CART (b) over-smooths and creates too few regions in the city; CRISP (c) under-smooths, creating too many regions; and MVTV (d) provides a good balance that yields interpretable sections.

notator predictions. The raw data is noisy and has high local variance, and so annotators do poorly at the prediction task without any smoothing. The over-smoothed CART values create too many uniform plateaus where the annotators cannot reasonably predict anything other than the missing uniform value, which has low accuracy. The CRISP method fails to sufficiently smooth the data, resulting in overly noisy patches which again makes the prediction task difficult. MVTV and MV-CRISP provide a good balance of smoothing and flexibility.

According to a Tukey's range test comparing pairwise human annotations across methods, both MVTV and MV-CRISP statistically significantly outperform the "raw" data

Figure 6: The Mechanical Turk interface used to gather human annotations. Annotators used the slider to fill in the missing value from a 7x7 patch of smoothed data from MVTV, MV-CRISP, CART, CRISP, and raw (unsmoothed) data.

for the human prediction task. By contrast, CART and CRISP fail to provide sufficient evidence to reject the null hypothesis that they are indistinguishable from the "raw" data.. No smoothing methods were shown to outperform one another with significance.

Conclusion

This paper presented MVTV, a new method for interpretable low-dimensional regression. Through a novel maximum variance criterion, our model divides the covariate space into a finite-sized grid in a data-adaptive manner. We then use a fast TV denoising algorithm to smooth over the cells, creating plateaus of constant value.

On a series of synthetic benchmarks, we demonstrated that our method produces superior results compared to a baseline CART model and the current state of the art (CRISP). Additionally, we provided additional evaluation through a real-world case study on crime rates in Austin, showing that MVTV discovers more interpretable spatial plateaus. Overall, we believe the speed, accuracy, interpretability, and fully auto-tuned nature of MVTV makes it a strong candidate for

low-dimensional regression when human understanding is a top priority.

Acknowledgments

The authors thank the reviewers for the comments and Subhashini Venugopalan for her helpful insights on experimental design. This work was generously supported by NSF CAREER grant DMS-1255187, and an NSF Graduate Research Fellowship to the second author.

References

Barbero, A., and Sra, S. 2011. Fast newton-type methods for total variation regularization. In Getoor, L., and Scheffer, T., eds., *ICML*, 313–320. Omnipress.

Barbero, A., and Sra, S. 2014. Modular proximal optimization for multidimensional total-variation regularization.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3(1):1–122.

Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. A. 1984. *Classification and regression trees*. CRC press.

Cassa, C. A.; Grannis, S. J.; Overhage, J. M.; and Mandl, K. D. 2006. A context-sensitive approach to anonymizing spatial surveillance data. *Journal of the American Medical Informatics Association* 13(2):160–165.

Chambolle, A., and Darbon, J. 2009. On total variation minimization and surface evolution using parametric maximum flows. *International journal of computer vision* 84(3):288–307.

Gupta, M.; Cotter, A.; Pfeifer, J.; Voevodski, K.; Canini, K.; Mangylov, A.; Moczydlowski, W.; and van Esbroeck, A. 2016. Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research* 17(109):1–47.

Hallac, D.; Leskovec, J.; and Boyd, S. 2015. Network lasso: Clustering and optimization in large-scale graphs. *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'15)*.

Johnson, N. A. 2013. A dynamic programming algorithm for the fused lasso and l 0-segmentation. *Journal of Computational and Graphical Statistics* 22(2):246–260.

Petersen, A.; Simon, N.; and Witten, D. 2016. Convex regression with interpretable sharp partitions. *Journal of Machine Learning Research* 17(94):1–31.

Rudin, L.; Osher, S.; and Faterni, E. 1992. Nonlinear total variation based noise removal algorithms. *Phys. D* 60(259–68).

Sadhanala, V.; Wang, Y.-X.; and Tibshirani, R. J. 2016. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. arXiv:1605.08400.

Sakamoto, Y.; Ishiguro, M.; and Kitagawa, G. 1986. Akaike information criterion statistics.

Tansey, W.; Athey, A.; Reinhart, A.; and Scott, J. G. 2016. Multiscale spatial density smoothing: an application to large-scale radiological survey and anomaly detection. *Journal of the American Statistical Association*.

Tansey, W.; Koyejo, O.; Poldrack, R. A.; and Scott, J. G. 2017. False discovery rate smoothing. *Journal of the American Statistical Association (JASA): Theory and Methods*.

Tibshirani, R. J., and Taylor, J. 2011. The solution path of the generalized lasso. *Annals of Statistics* 39:1335–71.

Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; and Knight, K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society (Series B)* 67:91–108.