Interpretable Low-Dimensional Regression via Data-Adaptive Smoothing

Wesley Tansey [1] **Jesse Thomason** [1] **James G. Scott** [1]

Abstract

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 GapTV, an approach that is conceptually related both to CART and to the more recent CRISP algorithm (Petersen et al., 2016), a state-of-the-art alternative method for interpretable nonlinear regression. GapTV 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. We compare our approach against CART and CRISP via both a complexity-accuracy tradeoff metric and a human study, demonstrating that that GapTV is a more powerful and interpretable method.[1]

1. Introduction

A recent line of research in interpretable machine learning focuses on low-dimensional regression, where the feature set is relatively small and 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, the Convex Regression via Interpretable Sharp Partitions (CRISP) method (Petersen et al., 2016) has recently been introduced as a way to achieve a good trade off between accuracy and interpretability by inferring sharply-defined 2d 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. 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.[2]

To this end, and motivated by the success of CRISP, we present GapTV, a method for interpretable, low-dimensional convex regression with sharp partitions. GapTV involves two main steps: (1) a non-standard application of the gap statistic (Tibshirani et al., 2001) to create a data-adaptive grid over the feature space; and (2) smoothing over this grid using a fast total variation denoising algorithm (Barbero & Sra, 2014). The resulting model displays a good balance between interpretability, average accuracy, and degrees of freedom. We conduct a human study on the predictive interpretability of each method, showing both qualitatively and

[1] University of Texas at Austin, Austin, Texas, USA. Correspondence to: Wesley Tansey <tansey@cs.utexas.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[1] A full version of this paper is currently in submission to *NIPS'17*.

[2] 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)).

quantitatively that GapTV achieves superior performance over CART and CRISP.

2. Background

2.1. Convex Regression with Interpretable Sharp Partitions

Petersen et al. (2016) propose the CRISP algorithm. As in our approach, they focus on the 2d scenario and divide 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 $1/q$ 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,

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

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 ADMM. 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 and thin-plate splines (TPS), CRISP is shown to yield a good tradeoff between accuracy and interpretability.

2.2. 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 $\mathcal{E}$,

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

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 (Wang et al., 2014) 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 et al., 1992) or the graph-fused lasso, for which several efficient solutions have been proposed (Chambolle & Darbon, 2009; Barbero & Sra, 2011; 2014).

The TV denoising penalty was investigated as an alternative to CRISP in (Petersen et al., 2016). They note anecdotally that TV denoising over-smooths when the same q was used for both CRISP and TV denoising. We present a principled approach to choosing q in a data-adaptive way that prevents over-smoothing and leads to a superior fit in terms of the accuracy-complexity tradeoff.

3. The GapTV Algorithm

We 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. Having formulated the problem as a graph TV denoising problem, we can now use the convex minimization algorithm of Barbero & 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.

3.1. Choosing bins via the gap statistic

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 via a novel use of the gap statistic (Tibshirani et al., 2001).

In a typical clustering algorithm, such as K-means, one would have unlabeled data $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, some distance metric $\delta(\mathbf{x}_i, \mathbf{x}_j)$, and a specified number of K clusters to find. In K-means, cluster assignment is based on the nearest centroid,

$$a_i = \underset{k}{\operatorname{argmin}} \quad \delta(\mathbf{x}_i, \mathbf{c}_k) \,, \tag{6}$$

where $\mathbf{c}_k = \frac{1}{|A_k|} \sum_{i \in A_k} \mathbf{x}_i$ is the cluster centroid and $A_k = \{i : a_i = k, \forall i\}$.

The gap statistic is an approach to choosing the value of K for a generic clustering algorithm by comparing it against a suitable null distribution. The best clustering is the one which minimizes the gap term:

$$\mathbb{E}_n \left[\log(W_1^*)\right] - log(W_K) \,, \tag{7}$$

where W_K is the sum of average pairwise distances in each cluster for a clustering with K clusters. To use the gap statistic, one must define a suitable null distribution over W_1.

In our case, the "clusters" are defined by a quantile grid over (x_1, x_2). The number of cells is specified by the choice of q, which means choosing the value of q corresponds directly to choosing K. However, unlike typical clustering, a cluster centroid is defined by the y_i values corresponding to the $\mathbf{x}_i$ points in the cell. Therefore, our distance metric for computing the gap statistic is actually between pairs of (y_i, y_j).

In the regression case, we assume each $y_i \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown. For a distance metric, we use Euclidean distance, $\delta(y_i, y_j) = (y_i - y_j)^2$. Since each y_i is assumed to be IID normal, the null distribution over pairwise distances is $W_1 \sim 2\sigma^2 \chi^2_\nu$, where $\nu = \frac{n^2}{2} - n$ is the degrees of freedom. The expectation of the log of a χ^2 distribution can be calculated exactly (Walck, 2007) as

$$\mathbb{E}\left[\log(\chi^2_\nu)\right] = \log 2 + \psi\left(\frac{\nu}{2}\right) \,, \tag{8}$$

where ψ is the digamma function. Thus, up to an additive constant, we can calculate the reference distribution exactly without knowing the mean or variance.

The procedure for choosing q is now straightforward. We first partition the points on a grid for a series of candidate q values in the range $2 \leq q \leq q_{\text{max}} \leq n$. For each candidate partitioning, we calculate the gap statistic

$$\text{gap}(q) = \psi(\frac{\nu}{2}) - \sum_{k=1}^{q^2} \frac{1}{\eta_k} \sum_{i \in A_i} \sum_{j \in A_i, j > i} \delta(y_i, y_j) \,. \tag{9}$$

We then choose the q that minimizes gap(q) and smooth using the TV denoising algorithm.

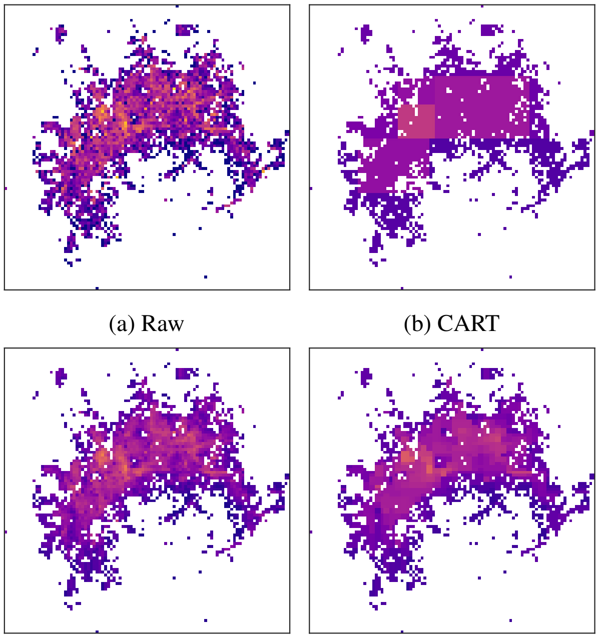
Once a value of q has been chosen, λ can be chosen by following a solution path approach. For the regression scenario with a Gaussian loss, as in (4), determining the degrees of freedom is well studied (Tibshirani & Taylor, 2011). Thus, we could select λ via an information criterion such as AIC or BIC. We select λ via cross-validation because we found empirically that it produces better results.

4. Case Study: Austin Crime Data

We applied CART, CRISP, and GapTV to a dataset of publicly-available crime report counts[3] in Austin, Texas in 2014. 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 1 (Panel A) shows the raw data for Austin.

The GapTV method 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 pixels were connected) which we then projected back to the real data for every non-missing point. Figure 1 shows the qualitative results for CART (Panel B), CRISP (Panel C), and GapTV (Panel D). The CART model clearly over-smooths by dividing the entire city into huge blocks of constant plateaus; conversely, CRISP under-smooths and creates too many regions. The

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



(a) Raw

(b) CART

(c) CRISP

(d) GapTV

Figure 1. Areal data results for the Austin crime data. The maps show the raw fine-grained results (Panel A) and the results of the three main methods. Qualitatively, CART (Panel B) over-smooths and creates too few regions in the city; CRISP (Panel C) under-smooths, creating too many regions; and GapTV (Panel D) provides a good balance that yields interpretable sections.

GapTV method finds an appealing visual balance, creating flexible plateaus that partition the city well. These results are confirmed quantitatively in Table 1, where GapTV outperforms the other methods in terms of AIC.

To evaluate the interpretability of the GapTV method against the benchmark CART and CRISP methods, we ran a Mechanical Turk study with human annotators. The annotation task was 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 GapTV, CART, CRISP, and as raw data; each task involved two randomly sampled patches from the Austin crime dataset ($5 \times 2 = 10$ patches per HIT, shown in random order).

We added two additional uniform validation patches, throwing out data from annotators who were not within 10% of the uniform value in the solid-colored patch. We gathered information from 207 annotators for 190 patches, throwing out 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).

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

Table 1. Results for the three methods on crime data for Austin. The GapTV method achieves the best trade-off between accuracy and the number of constant regions, as measured by AIC. Human annotator predictions are also statistically significantly closer than when annotators are shown raw data, which neither CART nor CRISP achieve.

The raw data is noisy and has high local variance, and so annotators do poorly at the prediction task without any smoothing (0.0471 ± 0.00539, not shown in Table 1). The oversmoothed 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. GapTV provides a good balance of smoothing and flexibility.

According to a Tukey's range test comparing pairwise human annotations across methods, GapTV statistically significantly outperforms the raw data for the human prediction task; by contrast, CART and CRISP fail to outperform the raw data. No methods were shown to outperform one another with significance.

References

Barbero, Álvaro and Sra, Suvrit. Fast newton-type methods for total variation regularization. In Getoor, Lise and Scheffer, Tobias (eds.), *ICML*, pp. 313–320. Omnipress, 2011.

Barbero, Álvaro and Sra, Suvrit. Modular proximal optimization for multidimensional total-variation regularization. 2014. URL http://arxiv.org/abs/1411.0589.

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

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

Gupta, Maya, Cotter, Andrew, Pfeifer, Jan, Voevodski, Konstantin, Canini, Kevin, Mangylov, Alexander, Moczydlowski, Wojciech, and van Esbroeck, Alexander. Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research*, 17(109):1–47, 2016. URL http://jmlr.org/papers/v17/15-243.html.

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

Petersen, Ashley, Simon, Noah, and Witten, Daniela. Convex regression with interpretable sharp partitions. *Journal of Machine Learning Research*, 17(94):1–31, 2016. URL http://jmlr.org/papers/v17/15-344.html.

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

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

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

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

Tibshirani, Robert, Walther, Guenther, and Hastie, Trevor. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Walck, Christian. Handbook on statistical distributions for experimentalists, 2007.

Wang, Yu-Xiang, Sharpnack, James, Smola, Alex, and Tibshirani, Ryan J. Trend filtering on graphs. *arXiv preprint arXiv:1410.7690*, 2014.