Robust Regression on Image Manifolds for Ordered Label Denoising

Hui Wu and Richard Souvenir
University of North Carolina at Charlotte
{hwu13,souvenir}@uncc.edu

Abstract

In this paper, we present a computationally efficient and non-parametric method for robust regression on manifolds. We apply our algorithm to the problem of correcting mislabeled examples from image collections with ordered (e.g., real-valued, ordinal) labels. Compared to related methods for robust regression, our method achieves superior denoising accuracy on a variety of data sets, with label corruption levels as high as 80%. For a diverse set of widely-used, large-scale, publicly-available data sets, our approach results in image labels that more accurately describe the associated images.

1. Introduction

Given the availability of images from the Web and increasingly cheap sensors and storage, amassing large image sets is relatively low-cost both in terms of effort and computational resources. However, obtaining the associated labels, necessary for supervised learning, is often a time-consuming, manual process that is becoming decreasingly viable with the staggering increase in the size of image collections. A recent trend is to acquire image labels via crowdsourcing or co-located sensors. These approaches effectively automate the label collection process, allowing for the rapid creation of labeled data sets at scales previously impossible. However, label accuracy often suffers. For example, Figure 1 shows representative images from two publicly-available computer vision data sets (AMOS [14] and Geofaces [13]) and the associated labels, including instances of mislabeled images. The goal of this paper is to correct mislabeled examples for image sets with ordered labels. While there has been work that addresses the classification variant of this problem (i.e., categorical labels or "tags"), there has not been much work for the problem of denoising real-valued or ordinal labels.

In this paper, we present a method to address the problem of denoising ordered labels from natural image sets. We take advantage of the fact that these data sets contain semantically-related images whose relationship can be exploited to learn a smooth function of the labels with respect to the images. From this point of view, the problem can be framed as robust regression in the high-dimensional domain of images. Unlike traditional robust regression methods, our method incorporates the observation that many natural image sets, although embedded in high-dimensional spaces, have only a few underlying causes of change that are usually semantically meaningful and correlated with the visual concept described by the image labels. We further combine this manifold assumption with sparse regularization, which allows our method to learn the underlying dependency between images and labels even with very high rate of label corruption. The contributions of this paper are:

- introducing the problem of ordered label denosing;
- an efficient, data-driven algorithm, based on the Hessian regularizer, for high-dimensional robust regression; and
- providing more accurate labels for widely-used image sets.

Original	-90	0	45	45
Output	-90	-52	43	90

Original	Clear	Clear	Cloudy	Cloudy
Output	Partly	Clear	Cloudy	Partly

Figure 1: Our method can be applied to image sets with ordered labels (top: head pose estimates, bottom: cloudiness estimates). For each image, we show the original label (top) and predicted value from our method (bottom). The examples in red highlight errors in the original labels.

2. Related Work

There has been a lot of work in the area of denoising categorical labels (e.g., [7, 26]) and the general problem of robust classification with mislabeled examples (e.g., [19]). Our work, to our knowledge, is the first to consider this problem in the context of regression, with ordinal or real-valued labels. While most regression techniques are somewhat tolerant to noise, they are generally not designed to handle large amounts of corruption found in the labels from real-world images sets. The problem, and our proposed approach are most closely related to robust regression and manifold regularization, specifically for high-dimensional ambient spaces.

Robust Regression The literature on robust regression is vast, spanning approaches from M-estimation to more recent methods designed to overcome the limitations of the commonly-used least squares error measure (e.g., sensitivity to noise and outliers). Robust substitutes have been investigated, including least median of squares [24] and least trimmed squares [2]. Least absolute deviation [27] has seen increased interest with the growing prominence of sparse representations and compressed sensing theory, with applications to vision and imaging problems, such as face recognition [29]. Our method also incorporates sparsity as means of discriminating between noisy and noise-free labels, but additionally correlates the labels to the underlying manifold structure commonly exhibited by natural image sets.

RANSAC Random Sample Consensus (RANSAC) and its variants [8, 22] have been successfully applied to a variety of geometric vision problems, such as 3D reconstruction from noisy feature matches [1, 25]. Most RANSAC methods are superlinear (and often exponential) in the number of iterations as a function of the number of model parameters. For geometric vision problems, the number of model parameters is usually small (e.g., 7 for the fundamental matrix). However, for our problem, the model parameters are derived from a high-dimensional image space and the relationship between the domain and range is unknown and, in most cases, nonlinear. In comparison, our algorithm is non-parametric, data-driven, and the time complexity is not a function of the ambient space dimension.

Manifold Regularization The manifold assumption is often used for natural image sets to sidestep the issue of high dimensionality in the image space. In addition to nonlinear dimensionality reduction, the image manifold assumption has been the basis for, among others, groupwise registration [30], semi-supervised learning [18, 15], and manifold denoising [9]. In this paper, we do not explicitly learn the underlying manifold parametrization, but leverage the manifold structure of the input images to learn a smoothly-varying label function using a corrupted set of examples.

Our method exploits the underlying manifold structure of natural image sets and integrates sparse regularization to formulate the problem of label denoising as an efficient convex optimization. In the next section, we describe the details of our approach and, in Section 4, show how our approach quantitatively and qualitatively outperforms representative regression methods on a variety of regression and ordered label denoising tasks.

3. Method

We are given inputs of N examples, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^\intercal$ and associated (noisy) labels, $\mathbf{y} = [y_1, y_2, \cdots, y_N]^\intercal$. For clarity, in this section, the labels are treated as one-dimensional real values; in Section 4, we present extensions for ordinal and multi-dimensional labels. For images, each example, $\mathbf{x}_i \in \mathcal{R}^D$, corresponds to the D-dimensional feature representation (e.g., raw pixel values, bag of words, HOG) of image i. We assume that the images are samples drawn from (or near) a low-dimensional manifold, $\mathcal{M}$, embedded in $\mathcal{R}^D$; the labels are samples of a function defined on the manifold; and the set of labels is contaminated by outliers. The goal is to learn a smooth function on the image manifold that recovers denoised labels, $\hat{\mathbf{y}}$.

3.1. Framework

Learning an unknown function on a manifold only defined by samples is an ill-posed problem. We follow the framework of regularized empirical risk minimization and start with the following general optimization:

$$\underset{f}{\operatorname{argmin}} \quad R(f; \mathbf{X}) + \lambda L(f; \mathbf{y}) \tag{1}$$

where R regularizes the function on the manifold defined by the samples, $\mathbf{X}$, L is a loss function, and λ is the trade-off parameter. Many choices are possible for the two terms and the space of possible functions. In this section, we describe our choices, motivated by the problem of ordered label denoising for natural image collections.

3.1.1 Manifold Regularization

Many approaches to manifold regularization have been proposed, which extend some notion of local linearity to a global model of the manifold. One such approach is based on the Hessian regularizer, which has been applied to, for example, nonlinear dimensionality reduction [6] and semi-supervised regression [6].

For a point on the manifold, the local Hessian functional is defined on its associated tangent space as the Frobenius norm of the Hessian matrix. This provides a coordinate

system that is isometric to the manifold intrinsic coordinate. The local measure is then averaged over the entire manifold to provide a global measurement, which is an extension of the average Frobenius norm of the Hessian of a function in Euclidean space to manifolds. Minimizing this term leads to locally linear functions. Several properties of the Hessian functional make it useful in our case: (1) it provides a data-driven way for manifold function regularization that enables non-parametric regression; (2) it can handle extrapolation better than other proposed manifold regularizers (e.g., Laplacian) [11]. However, unlike [6], our goal is not to explicitly learn a low-dimensional parametrization of the manifold, but to estimate a function of the labels over the images sampled from the manifold.

For an input point, $\mathbf{x}_i$, let $\mathcal{N}_i$ represent the neighborhood of K nearest neighbors and $\mathbf{z}_j^{(i)}$ represent the coordinates of $\mathbf{x}_j \in \mathcal{N}_i$ in the d-dimensional tangent space of $\mathbf{x}_i$, where $\mathbf{z}_i^{(i)}$ is defined as the origin. The local Hessian functional estimates a second-order polynomial, f, near $\mathbf{x}_i$ of the form:

$$f = \hat{y}_i + \mathbf{J}^{(i)}\mathbf{z}^{(i)} + \frac{1}{2}\mathbf{z}^{(i)\intercal}\mathbf{H}^{(i)}\mathbf{z}^{(i)} \tag{2}$$

where $\mathbf{J}^{(i)}$ and $\mathbf{H}^{(i)}$ are the local Jacobian and Hessian matrices, respectively, $\hat{y}_i$ is the predicted label, and $\mathbf{z}^{(i)}$ is the d-dimensional tangent space coordinate. Equation 2 is linear with respect to $\mathbf{J}^{(i)}$ and $\mathbf{H}^{(i)}$. Let $\mathbf{P}^{(i)}$ denote the design matrix on neighborhood, $\mathcal{N}_i$, and each row of $\mathbf{P}^{(i)}$ corresponds to a neighboring point, $\mathbf{x}_j$:

$$[\mathbf{z}_{j1}, \cdots, \mathbf{z}_{jd}, \mathbf{z}_{j1}\mathbf{z}_{j1}, \mathbf{z}_{j1}\mathbf{z}_{j2}, \cdots, \mathbf{z}_{jd}\mathbf{z}_{jd}] \tag{3}$$

where $\mathbf{z}_{jd}$ (superscript omitted for space) represents the d-th dimension of $\mathbf{z}_j^{(i)}$. Substituting the predicted values of the labels at the local neighborhood, denoted by $\hat{\mathbf{y}}^{(i)}$, for the unknown function, f, the least-squares solution for the parameters of the local Jacobian and Hessian matrices is given by:

$$\mathbf{P}^{(i)} \begin{bmatrix} | \\ \mathbf{J}^{(i)} \\ | \\ \breve{\mathbf{H}}^{(i)} \\ | \end{bmatrix} = \hat{\mathbf{y}}^{(i)} - \hat{y}_i \cdot \mathbf{1} \tag{4}$$

where $\breve{\mathbf{H}}^{(i)}$ represents the upper triangular portion of $\mathbf{H}^{(i)}$, $\mathbf{1}$ is a K-length column vector of 1, and $\mathbf{J}^{(i)}$ and $\breve{\mathbf{H}}^{(i)}$ are converted to column vectors. Let $\tilde{\mathbf{P}}^{\dagger}$ represent the bottom $d + d(d+1)/2$ rows of the pseudo-inverse of the design matrix and $\tilde{\mathbf{p}}_r^{\dagger}$ denote the r-th row of $\tilde{\mathbf{P}}^{\dagger}$.[1] We get the following expression for the approximation of local Hessian

[1] $\tilde{\mathbf{P}}^{\dagger}$ includes contributions from the $\hat{y}_i$ term in Equation 4, and rows in $\tilde{\mathbf{P}}^{\dagger}$ corresponding to the diagonal elements of $\mathbf{H}^{(i)}$ are scaled by 2 and those corresponding to off-diagonal elements are scaled by $\sqrt{2}$. These details are omitted for clarity and space.

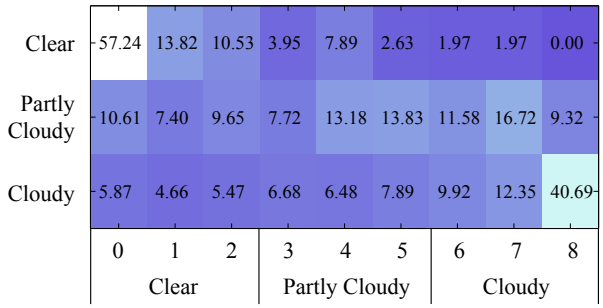


Figure 2: Distribution of the mislabeled examples. For a data set of outdoor images with cloudiness metadata (measured in okta from 0-8), the confusion matrix shows the distribution of the input label (columns) with manual annotations (rows).

functional:

$$\begin{aligned} \|\mathbf{H}^{(i)}\|_F^2 &= \sum_r \left(\tilde{\mathbf{p}}_r^{\dagger}\hat{\mathbf{y}}^{(i)}\right)^2 \\ &= \hat{\mathbf{y}}^{(i)\intercal}\mathbf{B}^{(i)}\hat{\mathbf{y}}^{(i)} \end{aligned} \tag{5}$$

where

$$\mathbf{B}^{(i)} = \sum_r \left(\tilde{\mathbf{p}}_r^{\dagger}\right)^{\intercal}\left(\tilde{\mathbf{p}}_r^{\dagger}\right) \tag{6}$$

The global Hessian estimator is the sum of the local estimators over all the input points. Let $\tilde{\mathbf{B}}^{(i)}$ denote the sparse $N \times N$ version of $\mathbf{B}^{(i)}$ where $\tilde{\mathbf{B}}^{(i)} = \mathbf{B}^{(i)}$ at the locations corresponding to points in $\mathcal{N}_i$ and 0 otherwise. So,

$$\mathbf{B} = \sum_{i=1}^{N} \tilde{\mathbf{B}}^{(i)} \tag{7}$$

and the global regularizer of the manifold function can be obtained in the quadratic form, $\hat{\mathbf{y}}^{\intercal}\mathbf{B}\hat{\mathbf{y}}$.

3.1.2 Loss Function

Modeling the noise of labels associated with large image collections can be difficult. Labels can be obtained from automated algorithms, co-located sensors, or crowdsourcing; each of which introduces different types of error. In our work, we observed that much of this data was corrupted nearly uniformly and not necessarily biased toward the ground truth. For example, consider the AMOS data set [14], which provides weather metadata associated with images captured from globally-distributed webcams. One label is cloud okta, a cloudiness measure that ranges from clear (0) to cloudy (8). Figure 2 shows a confusion matrix of the cloudiness values between the AMOS labels and manual annotations for a representative subset of 1000 images. This pattern of roughly uniformly distributed noise is consistent with research into labels obtained via crowdsourcing (e.g.

Amazon Mechanical Turk) where "bad" users tend to provide information uncorrelated with the correct answer [23].

This suggests that the commonly-used $L2$ error measure is not well-suited to the problem, as it often results in poor performance for non-normal noise distributions [27]. The $L1$ norm, however, is robust to high variance in noise and implicitly promotes sparsity in the residual error. This is the desired behavior, since sparsity in the residual error allows for soft subset selection of "good" labels and de-emphasizes the contribution of labels with extreme noise. In Section 4.1, we compare the performance of our method using both $L1$ and $L2$ loss.

3.2. Optimization

Combining the Hessian regularization term with the $L1$ loss, we are left with the following optimization:

$$\underset{\hat{\mathbf{y}}}{\operatorname{argmin}} \quad \hat{\mathbf{y}}^\intercal \mathbf{B} \hat{\mathbf{y}} + \lambda \|\hat{\mathbf{y}} - \mathbf{y}\|_1 \tag{8}$$

In order to efficiently solve Equation (8) for the denoised labels, $\hat{\mathbf{y}}$, we show that the global Hessian estimator, $\mathbf{B}$, is positive semidefinite (PSD).

Proof. First, the local Hessian estimator, $\mathbf{B}^{(i)}$, is PSD. In Equation 6, each term in the summation can be represented as the product of a matrix and its transpose, which is PSD. Next, we show that the sparse variant of the local estimator, $\tilde{\mathbf{B}}^{(i)}$, is PSD. Let $\mathbf{v}$ be a column vector of length N, so

$$\mathbf{v}^\intercal \tilde{\mathbf{B}}^{(i)} \mathbf{v} = \sum_{l=1}^{N} \sum_{m=1}^{N} \tilde{\mathbf{B}}^{(i)}_{lm} v_l v_m$$

Since $\tilde{\mathbf{B}}^{(i)}$ is a sparse matrix that contains the same elements of $\mathbf{B}^{(i)}$ at the intersections of rows and columns corresponding to $\mathcal{N}_i$, the above equation is reduced to a sum of K^2 terms:

$$\begin{aligned} \mathbf{v}^\intercal \tilde{\mathbf{B}}^{(i)} \mathbf{v} &= \sum_{l \in \mathcal{N}_i} \sum_{m \in \mathcal{N}_i} \tilde{\mathbf{B}}^{(i)}_{lm} v_l v_m \\ &= \hat{\mathbf{v}}^\intercal \mathbf{B}^{(i)} \hat{\mathbf{v}} \geq 0 \end{aligned}$$

where $\hat{\mathbf{v}}$ is a K-length column vector of elements from $\mathbf{v}$ at positions $\mathcal{N}_i$. Therefore, $\tilde{\mathbf{B}}^{(i)}$ is PSD. Finally, we get the global Hessian estimator, $\mathbf{B}$, is PSD since it is the sum of the N sparse local PSD matrices, $\{\tilde{\mathbf{B}}^{(i)}\}_{i=1}^{N}$. □

Therefore, Equation (8) is a convex quadratic program with $L1$ regularization. Performing Cholesky decomposition on $\mathbf{B}$, we get $\mathbf{B} = \Delta^\intercal \Delta$ and are left with:

$$\underset{\hat{\mathbf{y}}}{\operatorname{argmin}} \quad \|\Delta \hat{\mathbf{y}}\|_2^2 + \lambda \|\hat{\mathbf{y}} - \mathbf{y}\|_1 \tag{9}$$

where Δ is a sparse $N \times N$ upper triangular matrix. This convex optimization can be solved using standard algorithms, or using more efficient solvers specialized for large-scale, sparse $L1$-regularized least squares problems [16].

3.3. Algorithm

Given a set of images and (noisy) labels, our method, *Hessian-Regularized Robust Regression (H3R)*, outlined in Algorithm 1, returns denoised labels.

Algorithm 1 Hessian-Regularized Robust Regression

Input: images, $\mathbf{X}$; labels, $\mathbf{y}$;
Output: denoised labels, $\hat{\mathbf{y}}$

1: Estimate subspace dimension, d, and neighborhood size, K
2: **for all** $\mathbf{x}_i \in \mathbf{X}$ **do**
3: Find $\mathcal{N}_i$, the K-nearest neighbors of $\mathbf{x}_i$
4: Perform PCA on neighborhood, $\mathcal{N}_i$, to obtain d-dimensional tangent space coordinates
5: Construct design matrix, $\mathbf{P}^{(i)}$ (Eq. 3)
6: Compute local Hessian estimator (Eq. 6)
7: Construct global Hessian estimator, $\mathbf{B}$ (Eq. 7)
8: Solve for $\hat{\mathbf{y}}$ (Eq. 9)

For our method, the intrinsic dimension of the method, d, and the neighborhood size, K, can be provided using prior knowledge or estimated directly from the data. In Section 4.1, we describe the implementation details for H3R.

4. Evaluation

We evaluate the performance of H3R for ordered label denoising on a diverse set of labeled image collections and compare the results against the following regression methods.

- K**-NN**: The label of each point is estimated as the average labels of its K nearest neighbors in the data set, where K is set to the same value used by our method.
- Radial basis function network **(RBFN)** [20]: The neural network contains $\sqrt{N}$ hidden layer nodes with kernel width equal to the average distance to the 2-nearest cluster centers.
- **RANSAC** [22]:[2] The threshold for inliers is set to the 10% of the label dynamic range and maximum number of iterations is set to 10^7.
- ϵ support vector regression **(SVR)** [4] with the radial basis kernel. The kernel width is set to the average Euclidean distances of the input, and the inlier threshold, ϵ, is set to the 10% of the label dynamic range.
- Kernel Supervised PCA **(KSPCA)** [3]: for both the input data and labels, the radial basis kernel is used with the kernel width set to the average Euclidean distance.

[2]The linear model of RANSAC learns $D + 1$ parameters, where D is the dimensionality of the input. To make the problem tractable, for image data, we applied PCA to preserve 80% of the variation, which resulted in an input dimensionality of ∼20 across the data sets. Higher-order models were computationally prohibitive.

Data Sets We used labeled data sets with known ground truth. For each, the labels are normalized to $[0, 1]$.

- *Swiss Roll*, commonly used to evaluate machine learning algorithms, consists of 5000 points randomly sampled from a 2D manifold embedded in 3D. For each 3D example, the real-valued label is defined as sine of the geodesic distance to the center point on the manifold.
- *Paper Boy Statue* [21] consists of 840 images of a rigid object on a turntable platform captured from a camera on an elevating arm. The images are captured every 6 degrees of rotation from 0–354 and every 6 degrees of elevation from 6–84. Each image is cropped and subsampled to 32×20, represented as a pixel intensity vector, and noise was added to the elevation and rotation angles. To account for the cyclic rotation parameter, we take the values as cylindrical coordinates ($r = 1$) and convert to a 3D Euclidean parametrization. Results are reported as rotation and elevation angles.
- *Digit* [15] consists of 10,000 images of the digit "1", with four degrees of variations: horizontal translation, vertical translation, rotation and thickness. Each image is represented as a vector of raw pixel values.

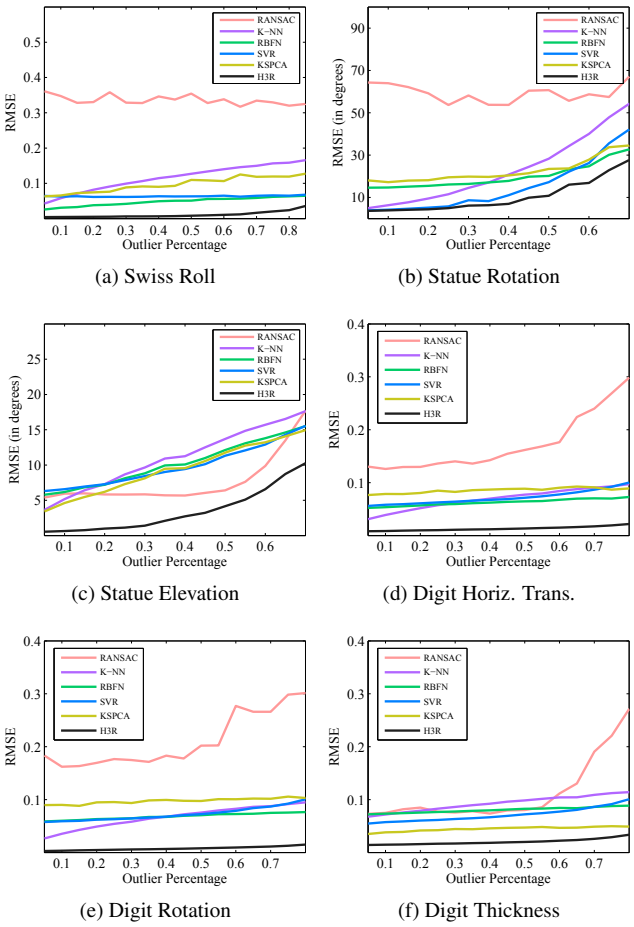
4.1. Implementation Details

For H3R, the intrinsic manifold dimensionality, d, neighborhood size, K, and trade-off parameter, λ can be provided if prior knowledge is available. However, these values can be directly estimated from the data, leaving no free parameters to the system. To estimate the intrinsic manifold dimensionality, d, we apply local PCA on a neighborhood of 20 points from a small set of randomly selected examples and set d as the value corresponding to the 'elbow point' of the residual variance curve. The number of nearest neighbors, K, is loosely related to the manifold intrinsic dimension. We found that the method was robust to the value of K, and empirically determined that $K = 5d$. For the regularization parameter, λ, we use the L-curve method [10] to select a value in the range $[10^{-10}, 10^{5}]$. The algorithm is implemented in Matlab and we use *l1-ls* package [16] for $L1$-regularized least squares optimization. The computation of the algorithm is dominated by the optimization step. On a standard PC, with an input of 1,000 samples, our method takes less than 5 seconds, on average.

Loss Function To evaluate the choice of loss function in our method, we performed manifold regression using the Swiss Roll data set (Figure 4) corrupted by commonly-used artificial noise models. Table 1 shows the root-mean-square error (RMSE) values of the predicted output from H3R. The order of noise models (left to right) represents moderate to high noise levels, and across all of the settings, the $L1$ norm outperforms the choice of $L2$, often by a wide margin. For the remaining experiments, we use the $L1$ loss with H3R.

	Lap.	Gauss.	Unif.	S&P
$L1$	**0.067**	**0.090**	**0.013**	**0.014**
$L2$	0.068	0.136	0.112	0.197

Table 1: RMSE of H3R on the Swiss Roll data using $L1$ or $L2$ loss. Noise was generated using the Laplacian ($b = 0.05$), Gaussian ($\sigma = 0.5$), uniform additive ($[-1, 1]$, 50% corruption), and salt & pepper (50%) noise models.



(a) Swiss Roll (b) Statue Rotation

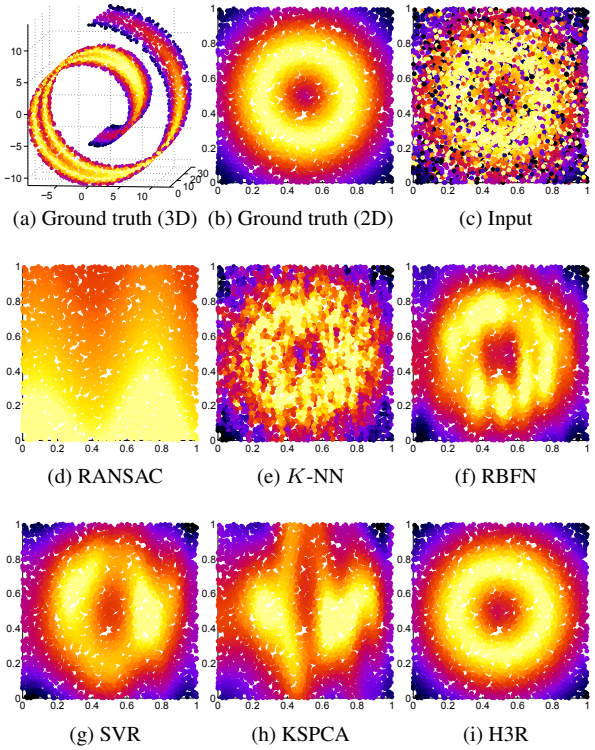
(c) Statue Elevation (d) Digit Horiz. Trans.

(e) Digit Rotation (f) Digit Thickness

Figure 3: RMSE values for the predicted labels on the Swiss Roll (a), Paper Boy Statue (b,c), and Digit (d-f) data sets with varying label corruption rate. (The results for vertical translation for the Digit data set closely followed that for horizontal translation.)

4.2. Robust Regression

For randomly-selected subsets of examples of varying size, the labels are corrupted by adding uniform noise in the range $[-1, 1]$. Each method is provided the (corrupted) labeled data as input. For multi-dimensional labels, each label is predicted independently for consistency across the



(a) Ground truth (3D) (b) Ground truth (2D) (c) Input

(d) RANSAC (e) K-NN (f) RBFN

(g) SVR (h) KSPCA (i) H3R

Figure 4: For the Swiss Roll (50% label corruption), the color in each plot indicates the manifold function value. For clarity, (b) to (i) are plotted using 2D manifold coordinates.

methods. Figure 3 shows the results of these experiments, reported as the average RMSE in the predicted values from 10 repeated trials.

Experiment Summary Across all of the experiments, H3R returns the closest predicted values, even at corruption rates as high as 80%. In all but two cases, RANSAC performed poorly. This is expected as RANSAC requires a pre-specified model and the linear model is not a reasonable choice for these experiments, as the relationship between the input and labels is nonlinear in most cases. While the remaining methods should be better-suited to nonlinear regression, for the problem of estimating the elevation of the camera for the Paper Boy Statue, RANSAC returned the next best predictions. In aggregate, SVR, KSPCA, K-NN, and RBFN showed similar performance with little consistency in relative performance across the experiments.

Manifold Regression The Swiss Roll data allows for both the manifold and function defined on the manifold to be easily visualized. Figure 4 shows the ground truth, corrupted input, and regression results from each method for an trial with 50% corruption rate. This is a 3D problem (Figure 4(a)); however, for clarity, the graphs in Figure 4 are plotted using 2D manifold coordinates. K-NN locally smooths the label noise, but the global model remains discontinuous and noisy. RBFN, SVR and KSPCA all learn smooth functions on the manifold, however, the recovered function deviates substantially from ground truth. The result from H3R closely matches the ground truth (RMSE $\approx$ 0), and remains nearly perfect up to corruption rates of 60%.

Image Manifold Regression Figure 5 shows the output from each method at 50% label corruption for the Paper Boy Statue data set. As shown in the top (ground truth) row, the images are sampled from a smoothly varying function of rotation and elevation. Most of the other methods include misplaced images, which indicate incorrect predictions for rotation, elevation, or both. While H3R returned the best predictions for both rotation and elevation, there were differences in the patterns of results. The change in elevation appears to be approximately linear, as RANSAC outperformed the nonlinear approaches (except for H3R) and was able to achieve low error rates (RMSE $\approx 5°$) up to 50% corrupted labels. This was not the case for the nonlinear transformation represented by turntable rotation, where RANSAC was the worst performer. However, for these different transformations, our method learned different accurate smooth functions on the same image manifold.

Similar results are observed with the Digit data. Figure 6 shows results for an experiment with 50% label corruption. Each group shows the images sorted by the listed parameter, with the remainder fixed. So, in the ideal case, there should only be a single smoothly varying transformation (e.g., rotation) across each row. Non-smooth changes from left to right or auxiliary changes from other transformations indicate an inaccurate prediction. The visual results align with the quantitative results. This is a challenging 4-dimensional prediction; H3R is the top-performer for each of the modes of image variability and returns low errors (RMSE < 0.05) at corruption rates up to 80%. This demonstrates the ability of our method to learn a variety of different functions on image manifolds.

4.3. Ordered Label Denoising

We consider the problem of denoising ordered labels from real-world, large-scale, publicly available data sets used as computer vision benchmarks. Due to space constraints, we only include the top three related methods (RBFN, SVR and KSPCA) for comparison. As opposed to quantitative measures of error, we interpret these results by visual inspection as ground truth is unavailable.

Weather from Images The Archive of Many Outdoor Scenes (AMOS) [14, 12] is a repository of millions of images captured from globally-distributed webcams. In addi-

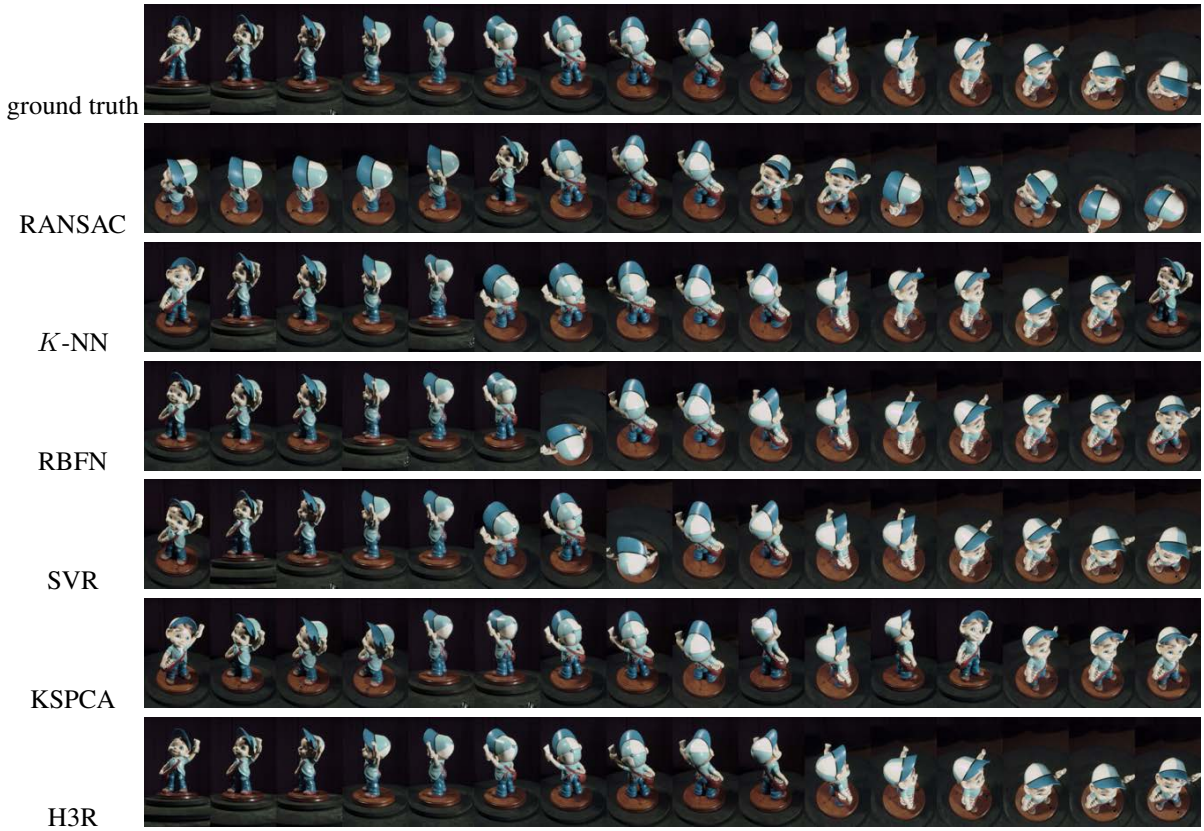


Figure 5: Each row shows images evenly sampled along a smoothly varying function of rotation and elevation. The top row shows the desired result; mismatches in subsequent rows indicate errors in denoising.

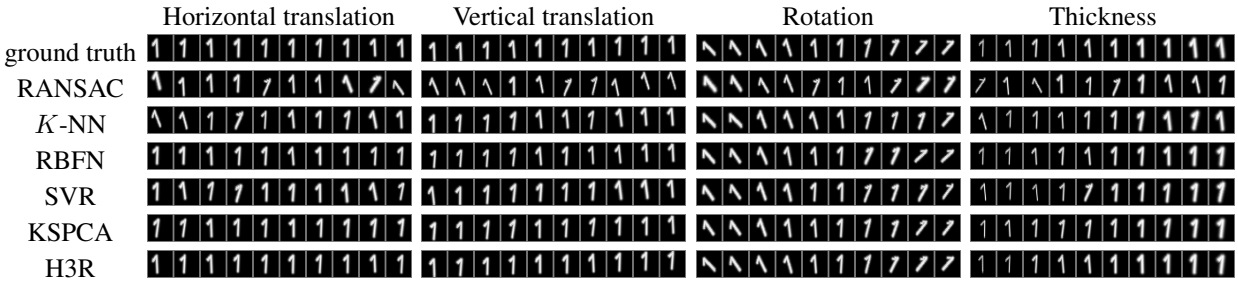


Figure 6: Each row shows images sorted by the predicted label. For each group, the specified (normalized) transformation should smoothly vary from 0 to 1, with the other labels fixed to 0.5. Non-smooth changes from left to right or auxiliary changes from other transformations indicate an incorrect prediction.

tion to images, AMOS provides associated weather metadata. While some of these parameters (e.g., air pressure, wind velocity) do not affect the appearance of the images, others, such as measures of cloud cover, can be important for methods in outdoor scene analysis. Some algorithms use clouds as a visual cue, while others assume cloudless imagery. Cloud okta, collected with AMOS images, is a measure of cloudiness from clear (0) to cloudy (8). These weather values are estimated from the closest weather stations, which may be far enough to be under different weather conditions from where the image is captured. This results in inaccurate labels, rendering cloudiness-based filtering unreliable.

Each AMOS image is represented using the 16-dimensional bag-of-colors feature [28], and the images are grouped by the originating webcam. Figure 7 shows representative results from various scenes. Those boxed in red are examples where the original label does not appear to

RBFN	0.3	5.9	2.0	0.7	2.5	4.6	3.2	7.0	8.0	4.7
SVR	0.8	3.9	2.7	0.6	3.0	3.9	2.2	7.0	7.8	7.0
KSPCA	0.4	5.3	2.3	0.6	2.0	4.3	2.4	6.3	8.0	6.6
H3R	0.8	1.3	2.0	0.8	2.5	3.6	4.7	6.4	8.0	7.3

Figure 7: Each image shows the original cloudiness label, which ranges, from 0 (clear) to 8 (cloudy). For each method, the predicted value is shown. Clearly mislabeled (input or predicted) values are indicated by the red text and boxes.

Figure 8: For each row, the images are shown with the (input or predicted) head pose estimate. Clearly mislabeled examples are highlighted by red boxes.

match the cloud level depicted in the scene. The 2^{nd} example shows a case where RBFN, SVR and KSPCA incorrectly changed a seemingly accurate label. The 4^{th} example shows a challenging scene that was both originally mislabeled and not corrected by any of the approaches. Overall, each of the methods improved upon the original labels, with H3R providing predictions that most closely matched visual appearance of the scene.

Face Pose Estimation Many widely used data sets for face analysis, including PubFig [17] and GeoFaces [13], rely on the same algorithm to annotate faces extracted from images collected from the Web or social networking sites. One of the provided parameters is an estimate of the pose of the face as one of five quantized directions: -90, -45, 0, 45, 90. This parameter would be used to, for example, retain only front-facing subjects.

Figure 8 shows the results of an experiment with 1,000 randomly selected images from GeoFaces. Each facial image patch is represented using HOG [5] features with a cell size of 50×50 and 9 orientation bins. The first row shows sample faces with the associated pose estimate. Each of the subsequent rows show the same subset of images sorted by the denoised head pose estimate. The red boxes indicate examples where the pose estimate does not visually match the direction the subject is facing. RBFN, SVR and KSPCA all improved upon the original labels and performed similarly in terms of the number of mislabeled predictions, even though the errors occurred in different regions of the label space. H3R outperformed each of the competing approaches, resulting in no grossly mislabeled examples.

5. Conclusions and Future Work

We presented an algorithm for robust regression on image manifolds and applied it to the problem of ordered label denoising for natural image sets. While the bulk of the algorithms and data sets for supervised learning in computer vision address classification, or categorization problems, there are important problems that rely on ordered output, such as articulated pose estimation and biomedical shape variation analysis, in addition to the examples presented in this paper. Our work is one of the first to address this underserved area. Our non-parametric and computationally efficient algorithm implicitly allows for the interpretation of ordered labels as a perceptually meaningful organization of the associated images and outperforms related regression methods on a variety of denoising tasks, including image collections with complex, multidimensional labels and over 70% label corruption. In the future, we plan to investigate large-scale adaptations, such as hierarchical decompositions, to apply this approach to Internet-scale image collections.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building rome in a day. In *IEEE International Conference on Computer Vision*, pages 72–79, 2009. 2

[2] A. Alfons, C. Croux, and S. Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248, 2013. 2

[3] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Zolghadri Jahromi. Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011. 4

[4] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011. 4

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005. 8

[6] D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003. 2, 3

[7] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *Advances in Neural Information Processing Systems 22*, pages 522–530. 2009. 2

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[9] D. Gong, F. Sha, and G. G. Medioni. Locally linear denoising on image manifolds. In *International Conference on Artificial Intelligence and Statistics*, pages 265–272, 2010. 2

[10] P. C. Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM review*, 34(4):561–580, 1992. 5

[11] M. Hein and M. Maier. Manifold denoising. In *Advances in neural information processing systems*, pages 561–568, 2006. 3

[12] M. Islam, N. Jacobs, H. Wu, and R. Souvenir. Images+weather: Collection, validation, and refinement. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop on Ground Truth*, 2013. 6

[13] M. Islam, S. Workman, H. Wu, R. Souvenir, and N. Jacobs. Exploring the geo-dependence of human face appearance. In *IEEE Winter Conference on Applications of Computer Vision*, 2014. 1, 8

[14] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2007. 1, 3, 6

[15] K. I. Kim, F. Steinke, and M. Hein. Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In *Advances in Neural Information Processing Systems*, pages 979–987, 2009. 2, 5

[16] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale l1-regularized least squares. *Selected Topics in Signal Processing, IEEE Journal of*, 1(4):606–617, 2007. 4, 5

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, Oct 2009. 8

[18] W. Liu, J. Wang, and S.-F. Chang. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, 100(9):2624–2638, 2012. 2

[19] N. Natarajan, I. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204. 2013. 2

[20] J. Park and I. W. Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991. 4

[21] R. Pless and I. Simon. Using thousands of images of an object. In *International Conference on Computer Vision, Pattern Recognition and Image Processing*, 2002. 5

[22] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J. Frahm. Usac: A universal framework for random sample consensus. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):2022–2038, Aug 2013. 2, 4

[23] V. C. Raykar and S. Yu. Ranking annotators for crowdsourced labeling tasks. In *Advances in neural information processing systems*, pages 1809–1817, 2011. 4

[24] P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984. 2

[25] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008. 2

[26] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 223–232, 2009. 2

[27] L. Wang, M. D. Gordon, and J. Zhu. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In *IEEE Sixth International Conference on Data Mining*, pages 690–700, 2006. 2, 4

[28] C. Wengert, M. Douze, and H. Jégou. Bag-of-colors for improved image search. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1437–1440, 2011. 7

[29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009. 2

[30] S. Ying, G. Wu, Q. Wang, and D. Shen. Groupwise registration via graph shrinkage on the image manifold. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2323–2330, 2013. 2