Regression with Sparse Approximations of Data

Noorzad, Pardis; Sturm, Bob L.

Published in:
Proceedings of the European Signal Processing Conference

Publication date:
2012

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Users may download and print one copy of any publication from the public portal for the purpose of private study or research.

Users may not further distribute the material or use it for any profit-making activity or commercial gain

Users may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: February 13, 2020
Introduction

- We propose SPARRe-apprOximation W眼界ed regression (SPARRW), a new method for locally polynomial regression function estimation, and an extension of sparse representation classification to the regression setting.
- To estimate the regression function at a point, SPARROW uses Taylor polynomial expansion around that point, least-squares optimal parameter estimation, and sparse approximation in terms of a dictionary of regressors and regressands.
- Our results show that locally constant SPARROW performs competitively, but the locally linear form, with and without regularization, does not.

What is Local Regression?

Consider we have a dataset of $N$ observations indexed by $\Omega := \{1, \ldots, N\}$:

$$D := \{(x_i, y_i) : x_i \in \mathbb{R}^m, y_i \in \mathbb{R}, i \in \Omega\}$$

We wish to estimate the regression function $f(x) : \mathbb{R}^m \to \mathbb{R}$ at a point $z \in \mathbb{R}^m$.

1. Approximating this function by a Taylor polynomial about $z$,

$$f(x) \approx f(z) + (x - z)^T \theta_2 + \frac{1}{2}(x - z)^T H_\alpha(x - z)$$

where $\theta_2$ and $H_\alpha$ are the gradient and Hessian of $f(x)$, evaluated at $z$.

2. We can solve for $(f(z), \theta_2, \text{and } H_\alpha) by

$$\min_{f(z), \theta_2, H_\alpha} \sum_{i \in \Omega} \alpha_i(z) \left| y_i - f(z) - (x_i - z)^T \theta_2 - \frac{1}{2}(x_i - z)^T H_\alpha(x_i - z) \right|^2$$

where $\alpha_i(z)$ is the $i$th observation weight. We can pose this as

$$\min_{\theta_2, H_\alpha} \left\{ \left[ A_i^2 / (y_i - X_i \theta) \right] \right\}$$

where $[A_i]_{..} := \alpha_i(z)$ and zero else, $\Theta_2 := \{f(z), \theta_2, \text{vech}(H_\alpha)\}$, and

$$X_\alpha := \begin{bmatrix} (x_1 - z)^T \text{vech}[(x_1 - z)(x_1 - z)^T] \\ (x_2 - z)^T \text{vech}[(x_2 - z)(x_2 - z)^T] \\ ... \end{bmatrix}$$

The notation vech(B) is the superscript of half of the symmetric matrix B.

3. The first element of the solution $\hat{\Theta}_2 = (X_\alpha^T A_i X_\alpha)^{-1} X_\alpha^T A_i y$ gives the least-squares optimal locally polynomial estimate of $f(z)$

$$\hat{f}(z) = (X_\alpha^T A_i X_\alpha)^{-1} X_\alpha^T A_i y$$

Taking only the first column of $X_\alpha$ gives a locally constant estimate of $f(z)$:

$$\hat{f}(z) = (1^T a) 1^T A_i y = \sum_{i \in \Omega} \alpha_i(z) y_i$$

Taking the first two columns gives a locally linear estimate of $f(z)$.

We must now define the $N$ observation weights $\alpha_i(z) : i \in \Omega$.

- Weighted $k$-nearest neighbor regression (Wk-NNR) defines the weights by the reciprocal of their Euclidean distance to $z$.
- Nadaraya-Watson kernel regression (NWR) method defines the weights using a kernel function, e.g., Gaussian, evaluated with respect to $z$.
- SPARROW defines the weights using the sparse approximation of $z$ with respect to the observed points in $D$.

How SPARROW Defines the Observation Weights

We construct a dictionary matrix by concatenating normalized regressors

$$D := \begin{bmatrix} x_1 \bar{x}_{1,1} & \bar{x}_{1,2} & \cdots & \bar{x}_{1,p} \\ x_2 \bar{x}_{2,1} & \bar{x}_{2,2} & \cdots & \bar{x}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_N \bar{x}_{N,1} & \bar{x}_{N,2} & \cdots & \bar{x}_{N,p} \end{bmatrix}.$$ 

For a given point $z$, SPARRW finds a solution to $z \approx D s$ such that $s$ has many zero elements by solving the basis pursuit denoising (BPDN) problem

$$\min_{s \in \mathbb{R}^N} |s|_1 \text{ subject to } \frac{\|z - D s\|_2^2}{\|z\|_2^2} \leq c^2$$

where $c > 0$. Defining $\Sigma$ as a diagonal matrix of the unbiased estimates of the variances observed in the dimensions of the regressors in $D$, SPARRW then defines the $i$th observation weight by

$$\alpha_i(z) := \frac{1}{\min_{j \in \Omega} \sum_j (z_j - x_{ij})^2} \frac{1}{\sum_j (z_j - x_{ij})^2} s_i$$

where $s_i$ is the $i$th element of $s$, $i \in \Omega$.

Experiments and Simulations

<table>
<thead>
<tr>
<th>Dataset</th>
<th># observations ($N$)</th>
<th># attributes ($M$)</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone</td>
<td>4,177</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Bodyfat</td>
<td>252</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>Housing</td>
<td>506</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>MPG</td>
<td>392</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

Our test datasets are described in the table above, with the last column showing the tuned parameter $k$ in our experiments with k-NNR and WA-NNR. The figures below compare locally-constant SPARROW (C-SPAR) and other methods. We use 100 independent trials of 10-fold cross-validation to estimate the mean squared error (MSE). Red lines mark median. Boxes delimit 25 to 75 percentiles. Extrema marked by whiskers, and outliers by pluses.

![Graphs showing MSE for different datasets](https://example.com/graphs)

For locally linear SPARROW (L-SPAR), we can employ regularization to solve

$$\min_{\Theta_\alpha, \lambda} \left\{ A_i^2 / (y_i - X_i \Theta) \right\}_2 + \lambda \| \Theta_\alpha \|_2^2$$

For the same datasets as above, the table below compares C-SPAR with L-SPAR with and without regularization. To estimate the MSE, we run 10 independent trials of 10-fold cross-validation. The last column denotes the ridge parameter to obtain the L-SPAR w/ R. estimate.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C-SPAR</th>
<th>L-SPAR w/ R.</th>
<th>L-SPAR</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone</td>
<td>5</td>
<td>16</td>
<td>988</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>Bodyfat</td>
<td>$5 \times 10^{-5}$</td>
<td>$35 \times 10^{-5}$</td>
<td>$960 \times 10^{-5}$</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>Housing</td>
<td>10</td>
<td>45</td>
<td>4304</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>MPG</td>
<td>7</td>
<td>8</td>
<td>6335</td>
<td>$10^{-3}$</td>
</tr>
</tbody>
</table>

Acknowledgments: B. L. Sturm is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsråd.