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):
We propose **SPARrow** to **approximate weighted regression** (SPARROW), 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 \( f(x) : \mathbb{R}^M \rightarrow \mathbb{R} \) at a point \( z \in \mathbb{R}^M \).

1. **Approximating this function by a Taylor polynomial about** \( z \), we have

\[
f(x) \approx f(z) + (x - z)^T \theta_z + \frac{1}{2} (x - z)^T H_z (x - z)
\]

where \( \theta_z \) and \( H_z \) are the gradient and Hessian of \( f(x) \), evaluated at \( z \).

2. **We can solve for** \( f(z) \), \( \theta_z \) and \( H_z \) **by**

\[
\min_{f(z), \theta_z, H_z} \sum_{i \in \Omega} \alpha_i (z) \left[ y_i - f(z) - (x_i - z)^T \theta_z - \frac{1}{2} (x_i - z)^T H_z (x_i - z) \right]^2
\]

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

\[
\min_{\theta_z, H_z} \left\| A_z [y - X z] \right\|_2
\]

where \( A_z = \alpha(z) \) and zero else, \( \theta_z := [f(z), \theta_z, \text{vech}(H_z)]^T \), and

\[
X_z := \begin{bmatrix}
1 & (x_1 - z)^T & \text{vech}[(x_1 - z)(x_1 - z)^T] \\
1 & (x_2 - z)^T & \text{vech}[(x_2 - z)(x_2 - z)^T] \\
\end{bmatrix}
\]

The notation \( \text{vech}(B) \) is the supervec of half of the symmetric matrix \( B \).

3. **The first element of the solution** \( \hat{\theta}_z = (X_z^T A_z X_z)^{-1} X_z^T A_z y \) **gives** the least-squares optimal locally polynomial estimate of \( f(z) \)

\[
\hat{f}(z) = (1^T A_{\Omega 1})^{-1} 1^T A_{\Omega y} = \sum_{i \in \Omega} \alpha_i(z) y_i \sum_{j \in \Omega} \alpha_j(z)
\]

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

\[
\hat{f}(z) = (1^T A_{\Omega 1})^{-1} 1^T A_{\Omega y}
\]

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

\[
D = \begin{bmatrix} x_1 & x_2 & \ldots & x_N \end{bmatrix}
\]

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

\[
\min_{s \in \mathbb{R}^N} \| s \|_1 \quad \text{subject to} \quad \frac{\| z - D s \|_2 \|_2}{\| z \|_2} \leq \epsilon^2
\]

where \( \epsilon^2 > 0 \). Defining \( \Sigma \) as a diagonal matrix of the unbiased estimates of the variances observed in the dimensions of the regressors in \( D \), SPARROW then defines the \( i \)th observation weight by

\[
\alpha_i(z) = \frac{(z - X_i)^T \Sigma^{-1} (z - X_i)}{\min_{j \neq i} \| (z - X_j)^T \Sigma^{-1} (z - X_j) \|_2^2}
\]

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.

---

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