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

Regression with Sparse Approximations of Data

Pardis Noorzad † and Bob L. Sturm ‡

† Dept. Computer Engineering and Information Technology, Amirkabir University of Technology, Iran;
‡ Dept. Architecture, Design and Media Technology, Aalborg University Copenhagen, Denmark

Introduction

- We propose **SPAR**se app**RO**ximation **W**eighted 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 $\Omega := \{1, \dots, N\}$:

$$\mathcal{D} := \left\{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^M, y_i \in \mathbb{R}, i \in \Omega\right\}$$

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

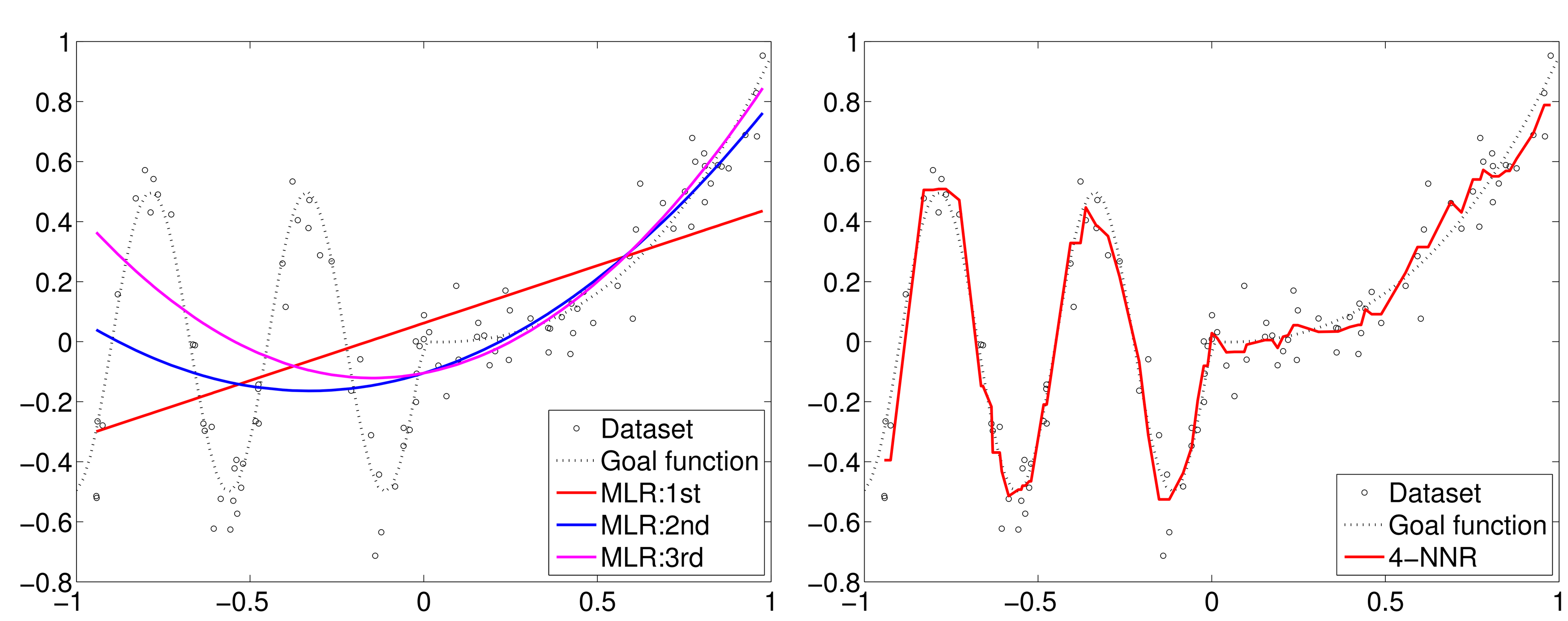
1. Approximating this function by a Taylor polynomial about $\mathbf{z}$, we have
$$f(\mathbf{x}) \approx f(\mathbf{z}) + (\mathbf{x}-\mathbf{z})^\top \boldsymbol{\theta}_\mathbf{z} + \frac{1}{2}(\mathbf{x}-\mathbf{z})^\top \mathbf{H}_\mathbf{z}(\mathbf{x}-\mathbf{z})$$
where $\boldsymbol{\theta}_\mathbf{z}$ and $\mathbf{H}_\mathbf{z}$ are the gradient and Hessian of $f(\mathbf{x})$, evaluated at $\mathbf{z}$.

2. We can solve for $f(\mathbf{z}), \theta_\mathbf{z}$ and $\mathbf{H}_\mathbf{z}$ by
$$\min_{f(\mathbf{z}),\theta_\mathbf{z},\mathbf{H}_\mathbf{z}} \sum_{i\in\Omega} \alpha_i(\mathbf{z})\left[y_i - f(\mathbf{z}) - (\mathbf{x}_i-\mathbf{z})^\top\boldsymbol{\theta}_\mathbf{z} - \frac{1}{2}(\mathbf{x}_i-\mathbf{z})^\top\mathbf{H}_\mathbf{z}(\mathbf{x}_i-\mathbf{z})\right]^2$$
where $\alpha_i(\mathbf{z})$ is the ith *observation weight*. We can be posed this as
$$\min_{\Theta_\mathbf{z}} \left\|\mathbf{A}_\mathbf{z}^{1/2}\left[\mathbf{y} - \mathbf{X}_\mathbf{z}\Theta_\mathbf{z}\right]\right\|_2^2$$
where $[\mathbf{A}_\mathbf{z}]_{ii} := \alpha_i(\mathbf{z})$ and zero else, $\Theta_\mathbf{z} := \left[f(\mathbf{z}), \boldsymbol{\theta}_\mathbf{z}, \text{vech}(\mathbf{H}_\mathbf{z})\right]^\top$, and
$$\mathbf{X}_\mathbf{z} := \begin{bmatrix} 1 & (\mathbf{x}_1-\mathbf{z})^\top & \text{vech}^\top[(\mathbf{x}_1-\mathbf{z})(\mathbf{x}_1-\mathbf{z})^\top] \\ \vdots & \vdots & \vdots \\ 1 & (\mathbf{x}_N-\mathbf{z})^\top & \text{vech}^\top[(\mathbf{x}_N-\mathbf{z})(\mathbf{x}_N-\mathbf{z})^\top] \end{bmatrix}.$$
The notation vech($\mathbf{B}$) is the supervector of half of the symmetric matrix $\mathbf{B}$.

3. The first element of the solution $\widehat{\Theta}_\mathbf{z} = \left(\mathbf{X}_\mathbf{z}^T\mathbf{A}_\mathbf{z}\mathbf{X}_\mathbf{z}\right)^{-1}\mathbf{X}_\mathbf{z}^T\mathbf{A}_\mathbf{z}\mathbf{y}$ gives the least-squares optimal locally polynomial estimate of $f(\mathbf{z})$
$$\hat{f}(\mathbf{z}) = \mathbf{e}_1^\top\left(\mathbf{X}_\mathbf{z}^\top\mathbf{A}_\mathbf{z}\mathbf{X}_\mathbf{z}\right)^{-1}\mathbf{X}_\mathbf{z}^\top\mathbf{A}_\mathbf{z}\mathbf{y}.$$
Taking only the first column of $\mathbf{X}_\mathbf{z}$ gives a locally constant estimate of $f(\mathbf{z})$:
$$\hat{f}(\mathbf{z}) = (\mathbf{1}^\top\mathbf{A}_\mathbf{z}\mathbf{1})^{-1}\mathbf{1}^\top\mathbf{A}_\mathbf{z}\mathbf{y} = \frac{\sum_{i\in\Omega}\alpha_i(\mathbf{z})y_i}{\sum_{k\in\Omega}\alpha_k(\mathbf{z})}$$
Taking the first two columns gives a locally linear estimate of $f(\mathbf{z})$.

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

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



Above we see the ability of local regression methods to model data locally.

How SPARROW Defines the Observation Weights

We construct a dictionary matrix by concatenating normalized regressors

$$\mathbf{D} := \left[\frac{\mathbf{x}_1}{\|\mathbf{x}_1\|_2}, \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|_2}, \dots, \frac{\mathbf{x}_N}{\|\mathbf{x}_N\|_2}\right].$$

For a given point $\mathbf{z}$, SPARROW finds a solution to $\mathbf{z} \approx \mathbf{Ds}$ such that $\mathbf{s}$ has many zero elements by solving the **basis pursuit denoising** (BPDN) problem

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

where $\epsilon^2 > 0$. Defining $\boldsymbol{\Sigma}$ as a diagonal matrix of the unbiased estimates of the variances observed in the dimensions of the regressors in $\mathcal{D}$, SPARROW then defines the ith observation weight by

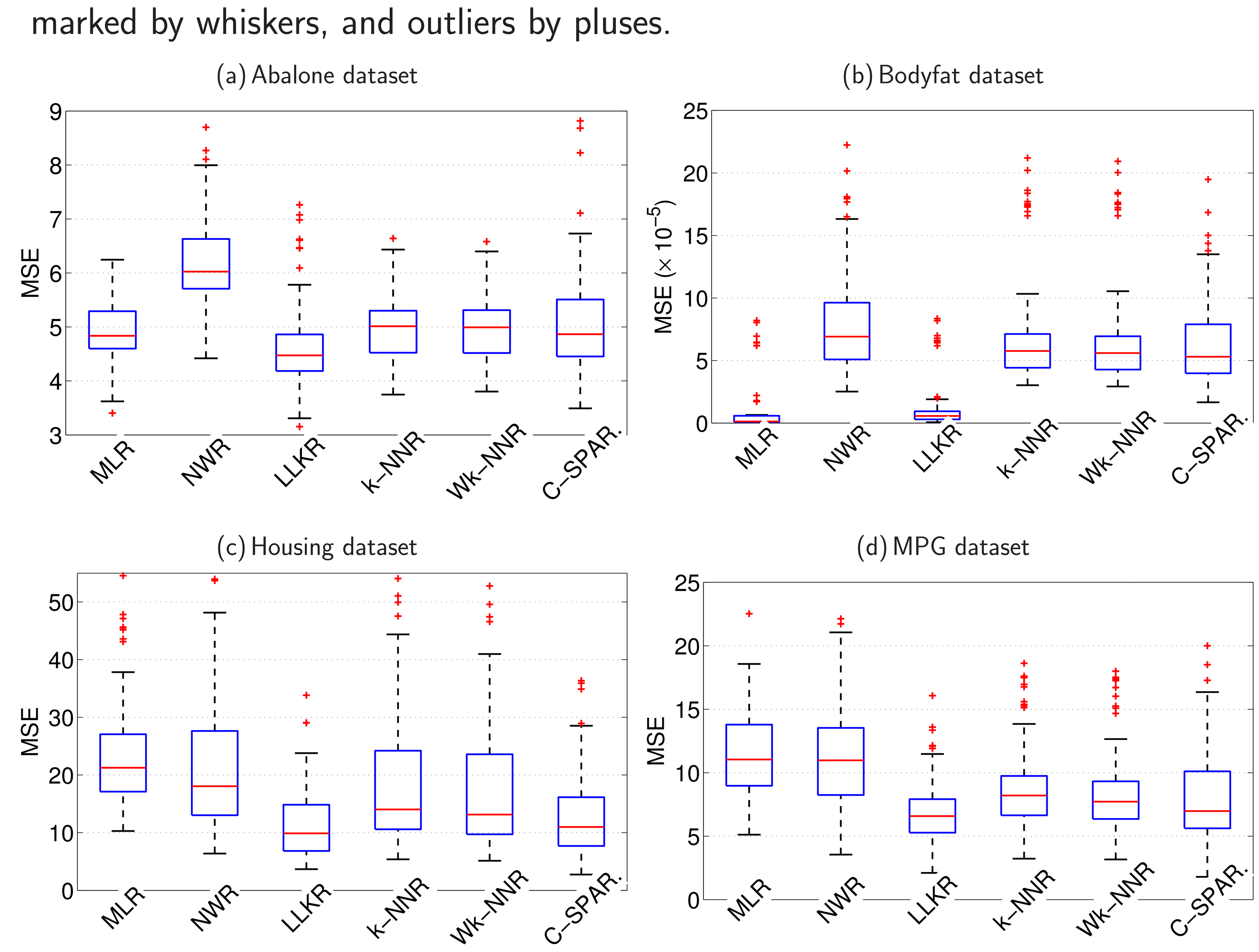
$$\alpha_i(\mathbf{z}) := \left[\frac{(\mathbf{z}-\mathbf{x}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\mathbf{x}_i)}{\min_{j\in\Omega}(\mathbf{z}-\mathbf{x}_j)^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\mathbf{x}_j)}\right]^{-1}\frac{s_i}{\|\mathbf{z}\|_2}$$

where s_i is the ith element of $\mathbf{s}$, $i \in \Omega$.

Experiments and Simulations

Dataset	# observations (N)	# attributes (M)	k
Abalone	4,177	8	9
Bodyfat	252	14	4
Housing	506	13	2
MPG	392	7	4

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 Wk-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.



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

$$\min_{\Theta_\mathbf{z},\lambda} \left\|\mathbf{A}_\mathbf{z}^{1/2}\left[\mathbf{y} - \mathbf{X}_\mathbf{z}\Theta_\mathbf{z}\right]\right\|_2^2 + \lambda\|\Theta_\mathbf{z}\|_2^2.$$

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.

Dataset	C-SPAR.	L-SPAR. w/ R.	L-SPAR.	λ
Abalone	5	16	988	10^{-3}
Bodyfat	5×10^{-5}	35×10^{-5}	960×10^{-5}	10^{-6}
Housing	10	45	4304	10^{-4}
MPG	7	8	6335	10^{-3}

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