Sparse Coding for Motions

Leif Johnson & Joseph Cooper

2012-05-09

Outline

Basic least squares regression

Suppose we have some noisy measurements $\mathbf{y}$ that were generated by an unobserved state $\mathbf{x}$ from a space spanned by the k columns of D:

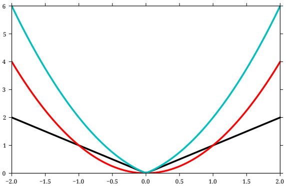
$$\mathbf{y} \sim \mathcal{N}(D\mathbf{x}, \sigma^2 I) \sim \mathcal{N}(\sum_{j=1}^{k} x_j \mathbf{d}_{\cdot j}, \sigma^2 I)$$

We can compute the most likely $\hat{\mathbf{x}}$ by minimizing squared error:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2} ||\mathbf{y} - D\mathbf{x}||_2^2$$

Least squares by itself is prone to modeling outliers and noise

Regularized least squares regression



To prevent overfitting, we introduce a **regularization** term:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2} ||\mathbf{y} - D\mathbf{x}||_2^2 + \lambda ||\mathbf{x}||_\zeta$$

Different values of ζ induce different priors on $\mathbf{x}$:

- $[\zeta = 0]$ — "L0-norm," unsolvable
- $[\zeta = 1]$ — lasso, Laplacian prior (Tibshirani, 1996)
- $[\zeta = 2]$ — ridge, Gaussian prior (Hoerl & Kennard, 1970)
- $[\zeta = 1] + [\zeta = 2]$ — elastic net (Zou & Hastie, 2005)

Why do we care about sparsity ?

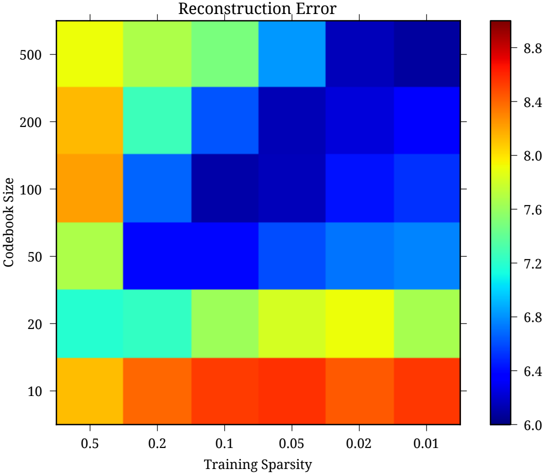
Suppose k is large ; sparsity limits "active" columns of D

- Helps make models easier for humans to understand
- Enables better compression

Sparsity seems to be a useful way of representing statistical properties of the natural world

So we'd like to keep ζ small to encourage sparse solutions

Sparse codes represent natural statistics efficiently



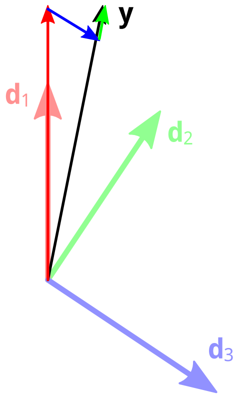
Forward feature selection (Mallat & Zhang 1993)

Repeat for $t = 1 \ldots T$:

- Compute correlations $\mathbf{c} = D^T \mathbf{r}_t$
- Find $i = \arg\max_j c_j$
- Add c_i to the model
- Define $\mathbf{r}_{t+1} \leftarrow \mathbf{r}_t - c_i \mathbf{d}_{\cdot i}$

Features are selected greedily based on current residual

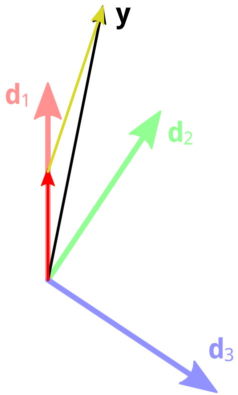
This is basically Matching Pursuit (Mallat & Zhang, 1993)



Least Angle Regression (Efron et al. 2004)

Repeat for $t = 1 \dots T$:

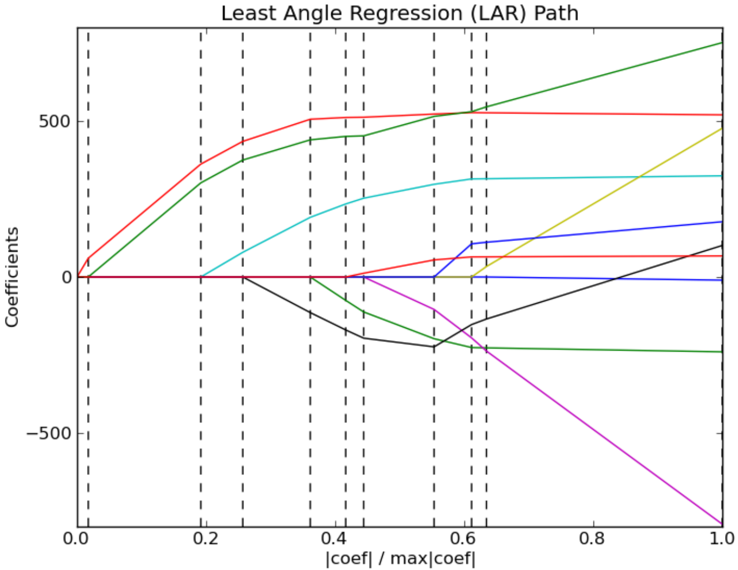
- Compute correlations $\mathbf{c} = D^T \mathbf{r}_t$
- Identify "active" columns $\mathcal{A} = \{j : |c_j| = \max_j\{|c_j|\}\}$
- Compute "equiangular" vector $\mathbf{u}$ such that $\mathbf{u}^T \mathbf{d}_{\cdot\mathcal{A}_1} = \mathbf{u}^T \mathbf{d}_{\cdot\mathcal{A}_2} = \dots$
- Compute largest γ such that $\mathbf{r}_t - \gamma\mathbf{u}$ admits one additional active column
- Define $\mathbf{r}_{t+1} \leftarrow \mathbf{r}_t - \gamma\mathbf{u}$



Developed by Efron, Hastie, Johnstone & Tibshirani (2004)

Same runtime complexity as OLS !

Regularization paths



Learning a sparse basis

With Matching Pursuit, dictionary is updated based on residual

- Multiple codebook vectors cannot "share" a residual

Another way to learn is through coordinate descent

- First, compute encoding(s) given a fixed dictionary
- Then, optimize the dictionary given a fixed set of encodings
- Somewhat similar in spirit to EM
- Provable convergence, no learning rate parameter

Developed by Mairal, Bach, Ponce & Sapiro (2009)

Learning via coordinate descent (Mairal et al. 2009)

Repeat for $t = 1 \ldots T$:

- Draw a sample $\mathbf{x}_t \sim p(\mathbf{x})$, and compute a sparse code:

$$\alpha_t = \arg\min_{\alpha} \frac{1}{2} \left|\left|\mathbf{x}_t - D_{t-1}\alpha\right|\right|_2^2 + \lambda \left|\left|\alpha\right|\right|_1$$

- Update running correlations:

$$A_t \leftarrow A_{t-1} + \alpha_t \alpha_t^T \qquad B_t \leftarrow B_{t-1} + \mathbf{x}_t \alpha_t^T$$

- Then optimize D given all previous α:

$$D_t = \arg\min_{D} \sum_{i=1}^{t} \frac{1}{2} \left(Tr(D^T D A_t) - Tr(D^T B_t) \right)$$