Thursday, 10/8/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 14

p. 1

Lecture 14	EE 660	Oct 8, 2020

Announcements

- Homework 5 is due tomorrow
- Final project assignment will be posted soon

Today's Lecture

- Sparsity and regularization
- Bridge regression and p-norm

L1 Regularization and Sparsity [Murphy 13.3.1]

Minimization problem:

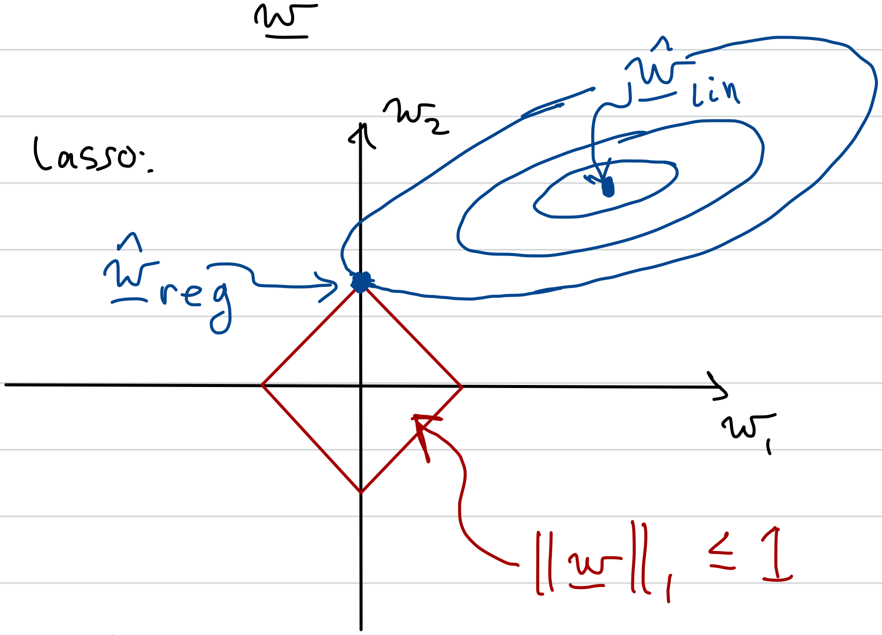
$$\underset{\underline{w}}{\operatorname{argmin}} \; J(\underline{w})$$

(i) $\Rightarrow$ $$\min_{\underline{w}} \left[RSS(\underline{w}) \right] \text{ s.t. } \|\underline{w}\|_1 \le B \qquad \text{(Lasso)}$$

Compare with:

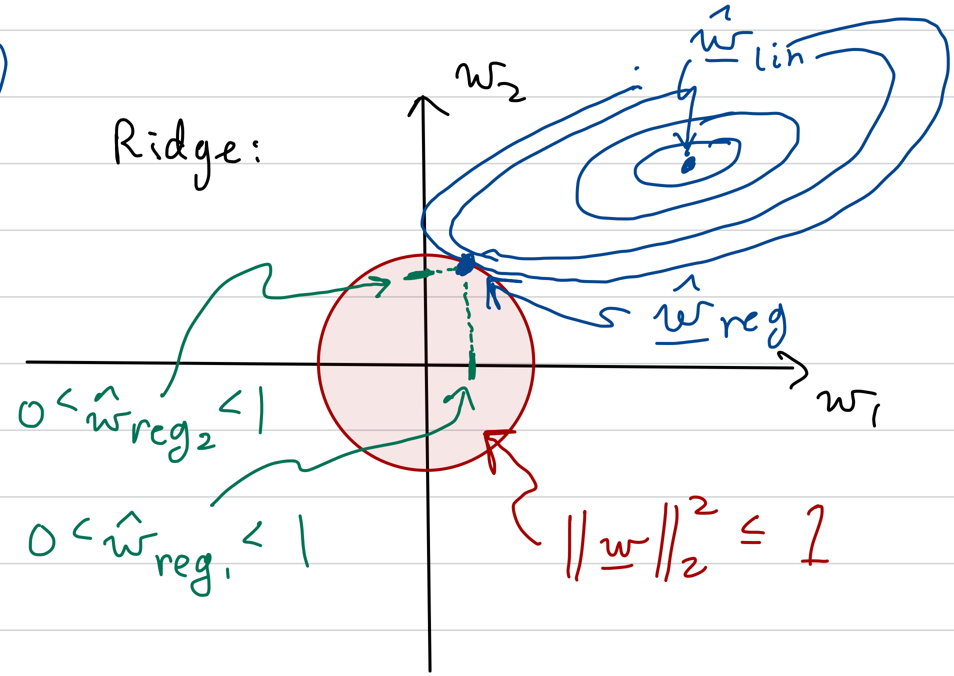
(ii) $$\min_{\underline{w}} \left[RSS(\underline{w}) \right] \text{ s.t. } \|\underline{w}\|_2^2 \le B' \qquad \text{(Ridge)}$$

Lasso:



$\hat{w}_{reg_1} = 0, \quad \hat{w}_{reg_2} = 1$

Ridge:



$\Rightarrow$ Lasso provides a sparser solution $\underline{\hat{w}}$ reg. (Generally true)

Example

Prostate cancer data (regression)
Goal: predict PSA from common features (8 features)
$N = 97$ patients

[Hastie Table 3.3]

[Murphy Fig. 13.7 – regularization paths]

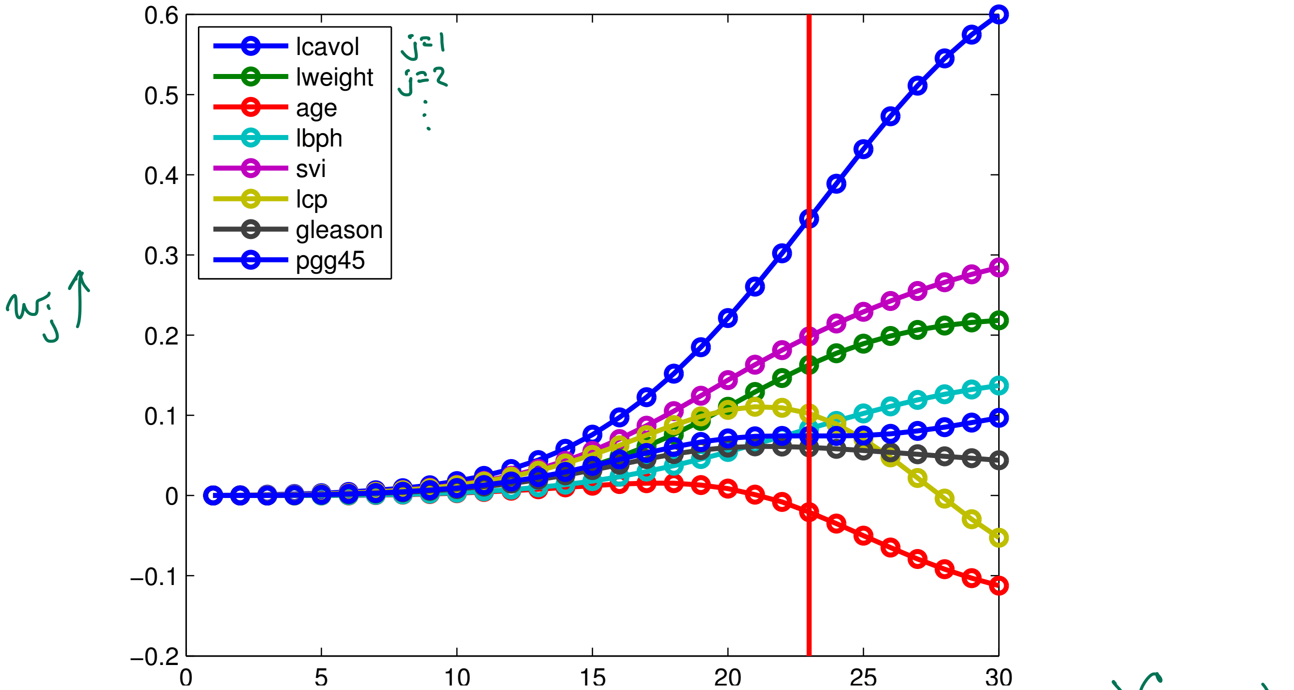
Comment: Bayesian feature selection and ℓ_0 regularization are covered in Murphy 13.2 (optional N.R.F.)

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

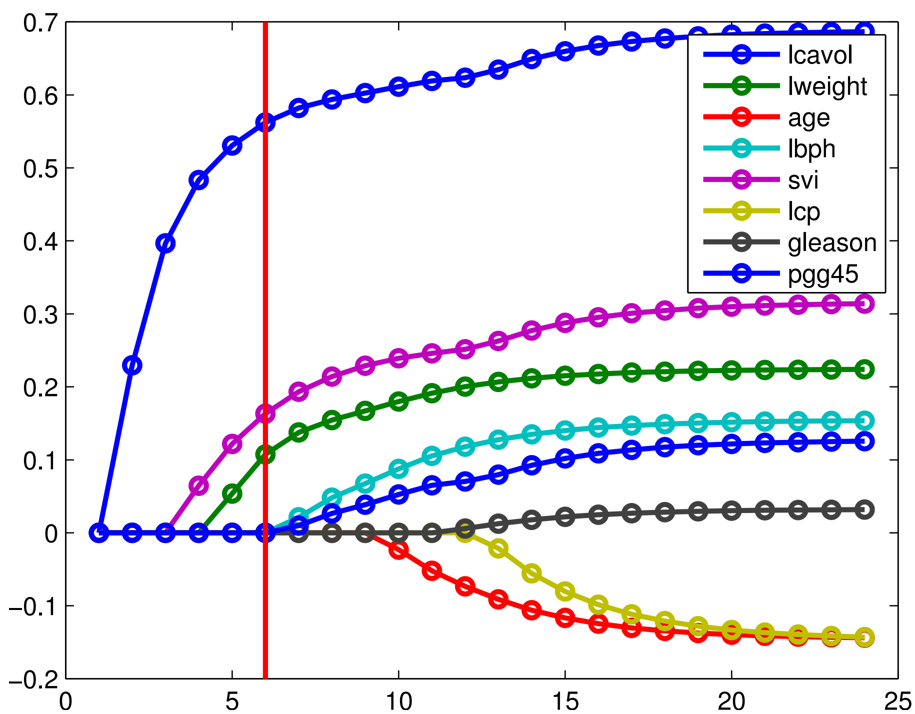
$f = RSS(\underline{w})$
no regularization

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

From: T. Hastie, et al., <u>The Elements of Statistical Learning</u>, 2nd Ed. (Springer, 2013).



Murphy Fig. 13.7(a). w_j vs. bound on L2 norm of $\underline{w}$



Murphy Fig. 13.7(b). w_j vs. bound on L1 norm of $\underline{w}$

Bridge Regression [Murphy 13.6.1]

→ Generalize l_2, l_1 regularization and constraints.

Estimate $\underline{w}$ using MAP with an exponential power distribution as prior.

$$\text{ExpPwr}(w_j \mid \mu_j, a, b) \triangleq \frac{b}{2a\Gamma(\frac{1}{b})} \exp\left\{ -\frac{|w_j - \mu_j|^b}{a^b} \right\}$$

[Murphy Eq. (13.132) ← may have errors]

Leads to an est.:

$$\hat{\underline{w}} = \underset{\underline{w}}{\arg\min} \left\{ NLL(\underline{w}) + \underbrace{\lambda \sum_{j=1}^{D} |w_j|^b}_{} \right\}, \quad b \geq 0$$

Like an "l_b" regularizer or "p-norm" based regularizer.

$$\text{p-norm} \triangleq \left[\sum_{j=1}^{D} |w_j|^p \right]^{1/p}$$

Note:

1. p-norm is a norm for $p \geq 1$
2. $\displaystyle\sum_{j=1}^{D} |w_j|^b$ is convex for $b \geq 1$
3. $\underbrace{\qquad\qquad}_{''}$ promotes sparsity for $0 \leq b \leq 1$.

Example:

$$-\log p(\underline{w} \mid \mathcal{D}) \propto \| \underline{y} - \underline{\underline{X}}\, \underline{w} \|^2 + \lambda \sum_{j=1}^{D} |w_j|^b$$

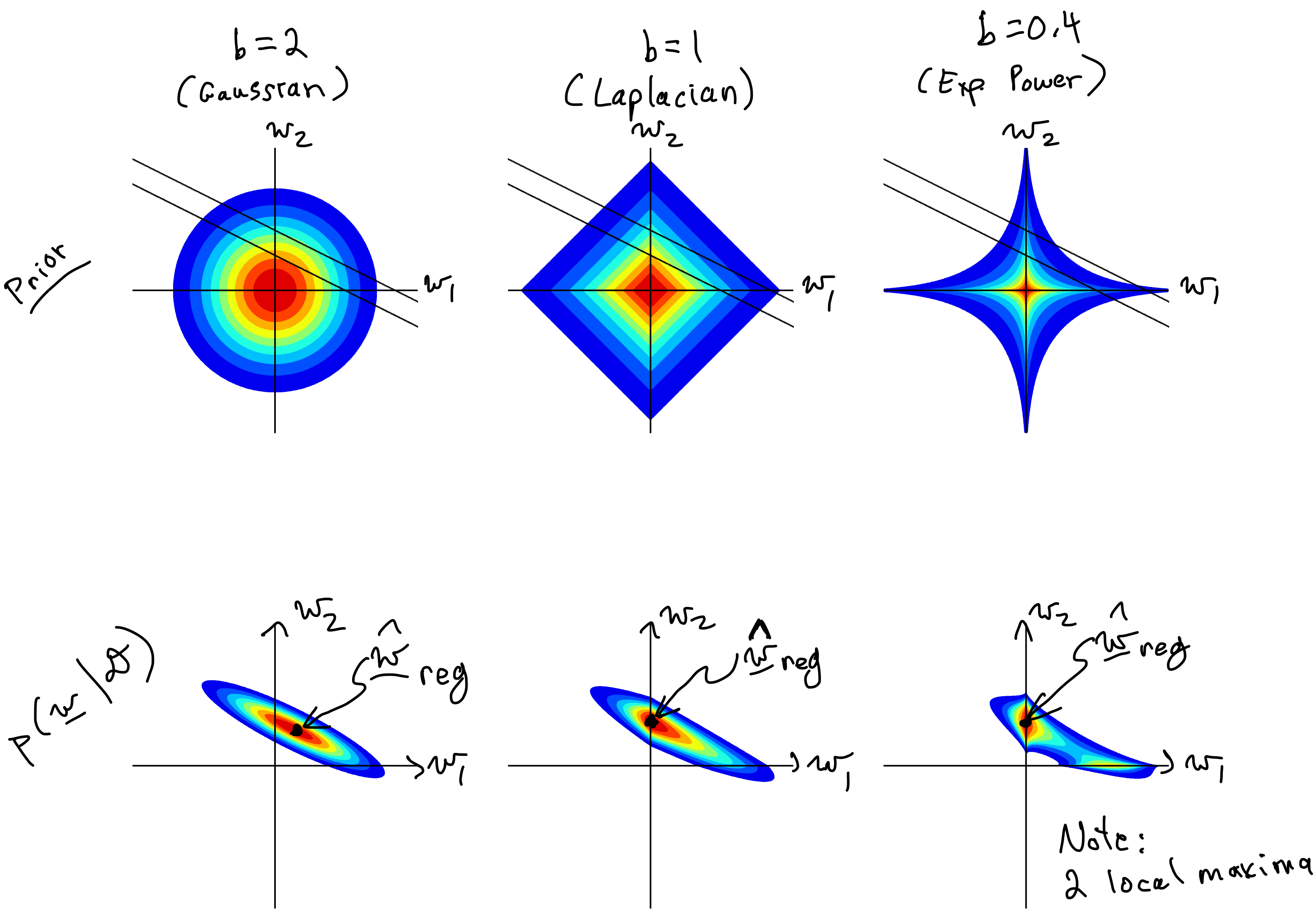
one observation: $N=1$, $(\underline{x}_1, y_1)$; $D=2$.

$$-\log p(\underline{w} \mid \mathcal{D}) \propto \left[y_1 - (w_1 x_1 + w_2 x_2) \right]^2 + \lambda \left[|w_1|^b + |w_2|^b \right]$$

Given

[Murphy Fig. 13.17]

N=1 throughout



Murphy Fig. 13.17.