Eigenvalue based regression, including Ridge (Tikhonov) and Differential regularization

Antoine Savine, April 2020

The exercise is to perform a classic least square linear regression $Y = \mu_Y + \left(\phi - \mu_\phi\right)\beta$, where the columns of $\phi = \phi\left(X\right)$ are basis functions (e.g. monomials, excluding constant) of known inputs X (also excluding constant, with examples in rows and inputs in columns), given a column vector Y of the corresponding targets, where μ_Y is the mean of Y and the row vector μ_ϕ contains the means of the columns of ϕ. To simplify notations, we denote $\phi \equiv \phi - \mu_\phi$ and $Y \equiv Y - \mu_Y$. Classic least squares finds β by minimization of the least square errors:

$$\beta = \arg\min_\beta \left\|Y - \phi\beta\right\|^2$$

The analytic solution, also called 'normal equation':

$$\beta = \left(\phi^T\phi\right)^{-1}\phi^T Y$$

is known to bear unstable results, the matrix $\phi^T\phi$ often being near singular (certainly so with monomials of high degree of correlated inputs). This is usually resolved with SVD regression. We prefer the (very similar) eigenvalue regression, which we recall first, and then, extend to Ridge (Tikhonov) regularization and differential regression (in the sense of Huge and Savine, Risk 2020).

Parts 1 and 2 are summaries of classic results. Part 3 is new.

After β is learned, the value approximation for an input row vector x is given by $y = \mu_Y + \left[\phi\left(x\right) - \mu_\phi\right]\beta$ and the derivative approximations are given by:

$$y_j = \phi_j\left(x\right)\beta$$

where subscripts denote partial derivatives to input number j.

1. Standard eigenvalue regression

Perform the eigenvalue decomposition $\phi^T\phi = PDP^T$.

Denote $D^{-\frac{1}{2}}$ the diagonal matrix whole diagonal elements are the elements of the diagonal matrix D, raised to power -0.5, when they exceed a threshold (say, 10^{-8} times the mean trace of D), and zero otherwise.

Denote $\phi = \phi P D^{-\frac{1}{2}}$ and perform the least square minimization in the orthonormal basis:

$$\beta = \arg\min_{\beta} \left\| Y - \phi\beta \right\|^2$$

The normal equation is stable in the orthonormal basis:

$$\beta = \left(\phi^T \phi \right)^{-1} \phi^T Y$$

It is easy to see that $\phi^T \phi$ is a diagonal matrix with diagonal elements 1 corresponding to significant eigenvalues, and 0 corresponding to insignificant ones. With the convention $\left(\phi^T \phi \right)^{-1} = \phi^T \phi$ (invert the significant diagonal elements and zero the insignificant ones), we get:

$$\beta = D^{-\frac{1}{2}} P^T \phi^T Y$$

(notice, $D^{-\frac{1}{2}}$ zeroes the lines corresponding to insignificant eigenvalues so there is no need to left multiply by $\phi^T \phi$.)

Hence: $Y = \phi\beta = \phi P D^{-1} P^T \phi^T Y = \beta Y$ where $D^{-1} = \left(D^{-\frac{1}{2}} \right)^2$ has diagonal elements inverse of the significant eigenvalues in D, zero for insignificant eigenvalue, and:

$$\boxed{\beta = P D^{-1} P^T \phi^T Y}$$

2. <u>Ridge (Tikhonov) eigenvalue regression</u>

Classic regression works best with regularization, the most common classic form of which is ridge regression, also called Tikhonov regularization, which adds a penalty on the norm of β in the objective cost.

$$\beta = \arg\min_{\beta} \left[\left\| Y - \phi\beta \right\|^2 + \lambda^2 \left\| \beta \right\|^2 \right] = \arg\min_{\beta} \left[\left\| Y - \left(\phi P D^{-\frac{1}{2}} \right) \left(D^{\frac{1}{2}} P^T \beta \right) \right\|^2 + \lambda^2 \left\| \beta \right\|^2 \right]$$

$$= P D^{-\frac{1}{2}} \arg\min_{\gamma} \left[\left\| Y - \phi\gamma \right\|^2 + \lambda^2 \left\| P D^{-\frac{1}{2}} \gamma \right\|^2 \right] = P D^{-\frac{1}{2}} \arg\min_{\gamma} \left[\left\| Y - \phi\gamma \right\|^2 + \lambda^2 \left\| D^{-\frac{1}{2}} \gamma \right\|^2 \right]$$

$$= P D^{-\frac{1}{2}} \left[\phi^T \phi + \lambda^2 D^{-1} \right]^{-1} \phi^T Y = P D^{-\frac{1}{2}} \left[\phi^T \phi + \lambda^2 D^{-1} \right]^{-1} D^{-\frac{1}{2}} P^T \phi^T Y = P \Lambda^{-1} P^T \phi^T Y$$

where Λ^{-1} has diagonal elements $\frac{1}{D_{ii} + \lambda^2}$ where D_{ii} is significant, zero otherwise. And we get:

$$\boxed{\beta\left(\lambda\right)=P\Lambda\left(\lambda\right)^{-1}P^T\phi^TY}$$

The Tikhonov parameter λ can be found e.g. by cross validation:

$$\lambda=\arg\min_{\lambda}\left\|Y_V-\phi_V\beta\left(\lambda\right)\right\|^2$$

where $\phi_V=\phi\left(X_V\right)$, $\left(X_V,Y_V\right)$ form a validation set of independent examples and $\beta\left(\lambda\right)$ is the result of a ridge regression over the training set with Tikhonov parameter λ, obtained with the boxed formula above. The objective function can be expanded:

$$\begin{aligned}
f\left(\lambda\right)&=\left\|Y_V-\phi_V\beta\left(\lambda\right)\right\|^2\\
&=\left(Y_V-\phi_V\beta\left(\lambda\right)\right)^T\left(Y_V-\phi_V\beta\left(\lambda\right)\right)\\
&=\left(Y_V-\phi_VP\Lambda\left(\lambda\right)^{-1}P^T\phi^TY\right)^T\left(Y_V-\phi_VP\Lambda\left(\lambda\right)^{-1}P^T\phi^TY\right)\\
&=\left(Y_V{}^T-Y^T\phi P\Lambda\left(\lambda\right)^{-1}P^T\phi_V{}^T\right)\left(Y_V-\phi_VP\Lambda\left(\lambda\right)^{-1}P^T\phi^TY\right)\\
&=Y_V{}^TY_V-Y^T\phi P\Lambda\left(\lambda\right)^{-1}P^T\phi_V{}^TY_V-Y_V{}^T\phi_VP\Lambda\left(\lambda\right)^{-1}P^T\phi^TY\\
&\quad+Y^T\phi P\Lambda\left(\lambda\right)^{-1}P^T\phi_V{}^T\phi_VP\Lambda\left(\lambda\right)^{-1}P^T\phi^TY
\end{aligned}$$

Since $Y_V{}^TY_V$ doesn't depend on λ, we minimize:

$$\begin{aligned}
g\left(\lambda\right)&=Y^T\phi P\Lambda\left(\lambda\right)^{-1}P^T\phi_V{}^T\phi_VP\Lambda\left(\lambda\right)^{-1}P^T\phi^TY\\
&\quad-Y^T\phi P\Lambda\left(\lambda\right)^{-1}P^T\phi_V{}^TY_V\\
&\quad-Y_V{}^T\phi_VP\Lambda\left(\lambda\right)^{-1}P^T\phi^TY\\
&=K^T\Lambda\left(\lambda\right)^{-1}M\Lambda\left(\lambda\right)^{-1}K-K^T\left(\lambda\right)^{-1}L-L^T\Lambda\left(\lambda\right)^{-1}K\\
&=K^T\Lambda\left(\lambda\right)^{-1}M\Lambda\left(\lambda\right)^{-1}K-2K^T\left(\lambda\right)^{-1}L\\
&=K^T\Lambda\left(\lambda\right)^{-1}\left[M\Lambda\left(\lambda\right)^{-1}K-2L\right]
\end{aligned}$$

where $K=P^T\phi^TY\left\{n\times1\right\},L=P^T\phi_V{}^TY_V\left\{n\times1\right\},M=P^T\phi_V{}^T\phi_VP\left\{n\times n\right\}$

Optimization may be efficiently performed by a classic one-dimensional minimization procedure.

3. Differential regression (Huge and Savine, Risk 2020)

In addition to inputs X and labels Y, we have derivatives labels Z whose columns Z_j are the differentials of Y to X_j. Denote ϕ_j the matrix of derivatives of the basis functions ϕ to X_j.

Linear regression makes value predictions $Y = \phi\beta$ and derivatives predictions $Z_j = \phi_j\beta$. We now minimize a cost combining value and derivatives errors:

$$\beta = \arg\min_\beta \left[\left\| Y - \phi\beta \right\|^2 + \sum_j \lambda_j \left\| Z_j - \phi_j \beta \right\|^2 \right]$$

where $\lambda_j = \lambda^2 \frac{\left\| Y \right\|^2}{\left\| Z_j \right\|^2}$ (norms are computed across examples) ensures that the components of the cost are of similar magnitudes. The parameter λ has little effect and generally left to 1.

It is not hard to see that this minimization is analytically solved with the adjusted normal equation:

$$\beta = \left(\phi^T \phi + \sum_j \lambda_j \phi_j^T \phi_j \right)^{-1} \left(\phi^T Y + \sum_j \lambda_j \phi_j^T Z_j \right)$$

This is, again, a theoretical equation, unstable in practice. As before, we change basis by eigenvalue decomposition of:

$$\phi^T \phi + \sum_j \lambda_j \phi_j^T \phi_j = PDP^T$$

-- beware, notations have changed so P and D denote different (respectively unitary and diagonal) matrices than before.

Changing basis as before: $\phi = \phi P D^{-\frac{1}{2}}$ (where, as previously, $D^{-\frac{1}{2}}$ has zero diagonal elements where the eigenvalues in D are insignificant) we notice that:

$$\phi_j \equiv \frac{\partial \phi}{\partial X_j} = \frac{\partial \left(\phi P D^{-\frac{1}{2}} \right)}{\partial X_j} = \frac{\partial \phi}{\partial X_j} P D^{-\frac{1}{2}} = \phi_j P D^{-\frac{1}{2}}$$

Performing the minimization in the ‘tilde’ basis:

$$\beta = \arg\min_\beta \left[\left\| Y - \phi\beta \right\|^2 + \sum_j \lambda_j \left\| Z_j - \phi_j \beta \right\|^2 \right]$$

We have the normal equation:

$$\beta = \left(\phi^T \phi + \sum_j \lambda_j \phi_j{}^T \phi_j\right)^{-1} \left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right)$$

$$= \left[\left(D^{-\frac{1}{2}} P^T\right)\left(\phi^T \phi + \sum_j \lambda_j \phi_j{}^T \phi_j\right)\left(P D^{-\frac{1}{2}}\right)\right]^{-1} \left[\left(D^{-\frac{1}{2}} P^T\right)\left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right)\right]$$

$$= D^{-\frac{1}{2}} P^T \left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right)$$

Predicted values are given by:

$$Y = \phi\beta = \phi P D^{-\frac{1}{2}} D^{-\frac{1}{2}} P^T \left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right) = \phi P D^{-1} P^T \left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right) = \phi\beta$$

where $D^{-1} = \left(D^{-\frac{1}{2}}\right)^2$ is defined as previously (with zeroes on insignificant eigenvalues) and:

$$\boxed{\beta = P D^{-1} P^T \left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right)}$$

Note that this is all consistent, in particular, derivatives predictions are given by:

$$Z_j = \phi_j \beta = \phi_j P D^{-1} P^T \left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right) = \left(\phi_j P D^{-\frac{1}{2}}\right)\left[D^{-\frac{1}{2}} P^T \left(\phi^T Y + \sum_j \lambda_j \phi_j{}^T Z_j\right)\right] = \phi_j \beta$$

Hence, just like Tikhonov regularization, differential regularization is analytic and extremely effective, as seen in the demonstration notebook. Unlike Tikhonov, differential regularization is *unbiased*, as demonstrated in an appendix of Huge & Savine, Risk 2020.