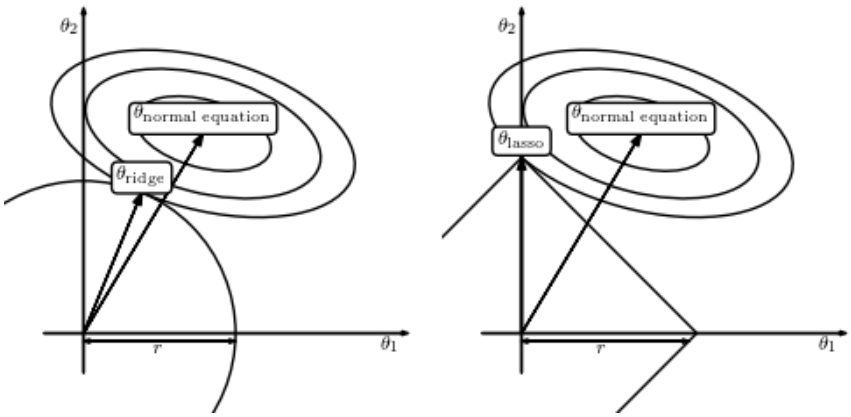

Regarding the L1 vs L2, I can't think of an exact article, but I can give you a quick intro:

ridge=L2, lasso = L1

- Both of them make sure that your coefficients in a linear regression do not grow too quickly. This is usually a good idea because if you have small datasets the coefficients would tend to over fit, and this over fit is (usually) given by large weights. Another view would be that large coefficients imply that a small change in your predictive variable creates large changes in your predicted variable, and this is often wrong (and associated to overfitting).
- Now, the pic that I sent explains what the difference is. If you do a regression, without regularization, your weights would go to the point theta-normal equation. When you add a regulaization, you penalize large weights, so you bring your coefficients closer to zero. Graphically, the idea is simple: adding a regularization simply says that you get two quantities now: the error (sum of square distances between fit and points), and the distance to the centre. Say that you fix the error. Then, you have an error-line (ellipses in the plot) within you can choose the coefficients. And you would choose the ones closer to the centre.
- The difference between L1 and L2 is on what "close" means.
    - In L2, "close" is your typical euclidean distance, so all the points in a radius around the origin are equally close. Then you would pick the coefficients that are at the point where your error-line touches (but does not cross) a circle centered at zero. The touch-but-not-cross comes from the fact that if the lines cross, then you could either reduce your error within the same distance-to-center or reduce your distance-to-center with the same error.
    - In L1, the distance is the sum of absolute values of your coefficient. This gives you a linear equation, as L1Dist((0,0), (x,y)) = |x| + |y|. But your coefficients still optimize the trade-off between (L1)-distance to the center and error, so you would still have them at a point where the error line intersects with a square.
- In practice, L2 is usually preferred if you have few variables and you know that probably all of them encode some independent information. However, if you have many variables and you think that some of them may be redundant, then use L1, which would take take the redundant information away by setting many coefficients to zero. In the plot, what you see is that the lasso regression puts theta-1 to zero, which means that in the trade-off error-distance, the error that can be avoided by theta_1 is smaller than the "regularization cost" of the coefficient theta_1.