"Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization" by Peter D. Hoff

Gustaw Matulewicz

May 23, 2017

Abstract

We provide a summary and analysis of the results from [Hof16]. The author shows that the problem of finding the Lasso estimator is equivalent by Hadamard reparametrization to a double ridge regression. The method can be extended to ℓ^q norm penalizations with $q \leq 1$ and to structured sparse estimation. The statement leads to an alternative algorithm which is compared to a few alternatives in terms of accuracy and speed.

1 The theory

We consider the Gaussian linear regression, where we wish to estimate $\beta \in \mathbb{R}^p$ given a matrix of predictor variables $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and the observation $y \sim \mathcal{N}\left(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}\right)$. This leads to the minimization of the square error $\|y - \boldsymbol{X}\beta\|_2^2$. Introducing $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{X}^\top$ and $l = \boldsymbol{X}^\top y$ this quantity is equal to:

$$h(\beta) = \beta^\top Q \beta - 2\beta^\top l.$$

1.1 Lasso and its Hadamard parametrization

The Lasso estimate for some $\lambda > 0$ is:

$$\hat{\beta}_\lambda := \arg\min_\beta \beta^\top Q \beta - 2\beta^\top l + \lambda \|\beta\|_1.$$

We reparametrize the problem by taking $u, v \in \mathbb{R}^p$ and writing $\beta = u \circ v$ (i.e. $\forall i, \beta_i = u_i v_i$) and use ℓ^2 penalities:

$$g(u, v) = (u \circ v)^\top Q (u \circ v) - 2(u \circ v)^\top l + \frac{\lambda}{2}(u^\top u + v^\top v).$$

Remark 1. *To see why this penalisation is chosen, observe that if $u^2 = v^2 = |\beta|$ (in the Hadamard sense), then the penalization is the same.*

The equivalence statement is given in Lemma 1:

Lemma 1. *For any function h, define $f(\beta) = h(\beta) + \lambda\|\beta\|_1$ and $g(u, v) = h(u \circ v) + \frac{\lambda}{2}(u^\top u + v^\top v)$. Then:*

1. $\inf_\beta f(\beta) = \inf_{u,v} g(u, v)$;

2. *if (u, v) is a local minimizer of g, then $\beta = u \circ v$ is a local minimizer of f.*

Remark 2. 1. *The infimum for g is unconstrained.*

2. *Both points together state that a global minimizer is the same for both problems and gives the same minimal value.*

3. *The Lemma applies to any function h.*

4. *The proof shows that the infima of g verify $u^2 = v^2 = |\beta|$ (there is a liberty in the choice of the signs).*

5. *We effectively replace a non-smooth problem by a smooth one. Intuitively, this is possible because when $(u, v) \to 0$, the variations of $\beta \sim \sqrt{u}$ are much higher.*

Proof. Main point: in dimension 1, the minimum of $u^2 + v^2$ constrained by $uv = \beta$ is attained when $u^2 = v^2 = |\beta|$. □

1.2 Fractional norm and multiple Hadamard parametrization

What happens if instead of 2 we use K parameters $\beta = u_1 \circ \dots u_K := \bigcirc_{i=1}^K u_i$? Lemma 2 shows that the problem is equivalent to penalizing h by a ℓ^q norm, where $q = 2/K$.

Lemma 2. *For any function h, define $f(\beta) = h(\beta) + \lambda \|\beta\|_q^q$ and $g((u_i)_{i \le K}) = h(\bigcirc_{i=1}^K u_i) + \frac{\lambda}{K} \sum_{i=1}^K \|u_i\|_2^2$. Then:*

1. $\inf_\beta f(\beta) = \inf_{(u_i)_{i \le K}} g((u_i)_{i \le K})$;

2. *if $(u_i)_{i \le K}$ is a local minimizer of g, then $\beta = \bigcirc_{i=1}^K u_i$ is a local minimizer of f.*

1.3 Structured sparsity and Bayesian interpretation

The Lasso penalization can is equivalent in Bayesian terms by a Laplace prior on the parameters. The fact that it treats all coordinates in the same way corresponds to prior independence between them. Using a penalization as group-lasso (which creates group-sparsity) is equivalent to prior dependence between the parameters. By analogy, the ℓ^2 penaliazations on the u parameters are equivalent to Gaussian priors. Taking a non-diagonal covariance creates dependence for β. For instance:

$$g(u, v) = (u \circ v)^\top Q (u \circ v) - 2(u \circ v)^\top l + u^\top \mathbf{\Sigma}_u^{-1} u + v^\top \mathbf{\Sigma}_v^{-1} v$$

is the log-posterior density of (u, v) under the model $y \sim \mathcal{N}\left(\boldsymbol{X}^\top \beta, \sigma^2 \boldsymbol{I}\right)$, $\beta = u \circ v$ and independent prior distributions $u/\sigma \sim \mathcal{N}(0, \mathbf{\Sigma}_u), v/\sigma \sim \mathcal{N}(0, \mathbf{\Sigma}_v)$. This gives a prior for β that verifies $\mathbb{C}\mathrm{ov}\left(\beta/\sigma^2\right) = \mathbf{\Sigma}_u \circ \mathbf{\Sigma}_v$.

Reverse engineering: if we want some covariance structure for β, a positive definite matrix, there exists a decomposition into a Hadamard product of positive definite matrices. We can use them as the prior covariances of u and v.

2 Numerical results

The equivalence suggests a simple algorithm, called HPP: use successive ridge regressions to minimize in one parameter u_i all other kept constant. An iteration consists in K ridge regressions.

Remark 3.

- *Each ridge regression requires the inversion of a $p \times p$ matrix.*
- *The algorithm never assigns 0 to any coordinate, hence after any number of iterations, the result will never be sparse. However, the sparsity is a property of the Lasso estimator, hence the result of the algorithm will have very small entries, which can be set to zero further down the line. Moreover, if any coordinate is set to zero, it will be kept at zero.*

The algorithm is easily comparable to the Local Quadratic Approximation, which solves successive ridge regression that approximate the Lasso problem. For $K = 2$, HPP performs marginally better than LQA. For larger K, LQA performs marginally better.

It is difficult to compare APP to a coordinate descent algorithm (CCD), because they use different tools, hence are more language/implementation specific.

The article suggests combining APP and CCD: in a loop, follow one iteration of APP by an iteration of CCD. The method had very strong numerical results. Furthermore, CCD creates sparsity, which simplifies the problem of APP as zero coordinates don't have to be recalculated. Hence, instead of inverting a $p \times p$ matrix, one needs to invert only a $s \times s$ matrix where s is the current sparsity of the estimator. Also, CCD can revert from a zero coordinate to a non-zero one, which improves the stability of the procedure.

Critical conclusion

The equivalence of the Lasso problem and a double-ridge regression can be interesting for some proofs. The resulting algorithm has performances that are marginally better than other techniques, but has the advantage of simplicity. It is also a rare example of an algorithm that works with some non-convex penalizations ℓ^q with $q = 2/K \leq 1$.

A similar method has been used in [JOB09]. The authors also use an additional variable in order to transform a non-convex problem into a series of convex problems.

References

[Hof16] Peter D. Hoff. Lasso, fractional norm and structured sparse estimation using a hadamard product parametrization. 11 2016.

[JOB09] Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. 09 2009.