Relating Leverage Scores and Density using Regularized Christoffel Functions

Edouard Pauwels, Francis Bach and Jean-Philippe Vert

Toulouse University / INRIA / ENS / Google

Outline

1. Leverage scores
2. Main result
3. Technical comments

Outline

Classical statistical leverage scores

- Goal: characterize how points "stick out" and affect the results of a statistical procedure
- Linear regression model:

$$y = X\beta + \epsilon$$

- Ordinary least squares

$$\hat{y} = Hy \quad \text{with} \quad H = X(X^\top X)^{-1}X^\top$$

- Leverage scores:

$$\ell = diag(H)$$

- Property

$$\forall i = 1, \ldots, n \quad \ell_i = \frac{\partial \hat{y}_i}{\partial y_i}$$

λ-ridge leverage scores

- (Kernel) ridge regression

$$\hat{y} = H(\lambda)y \quad \text{with} \quad H(\lambda) = X(X^\top X + n\lambda I_p)^{-1}X^\top = K(K + n\lambda I_n)^{-1}$$

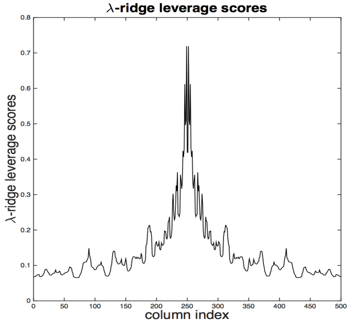
- Leverage scores:

$$\ell(\lambda) = diag(H(\lambda))$$

Use of leverage scores

- Diagnosis tool for linear regression (Hoaglin and Welsch, 1978; Velleman and Welsch, 1981; Chatterjee and Hadi, 1986)
- Matrix sketching and column sampling (Mahoney and Drineas, 2009; Mahoney, 2011; Drineas et al., 2012; Wang and Zhang, 2013)
- Low rank matrix approximation (Clarkson and Woodruff, 2013; Bach, 2013)
- Regression (Alaoui and Mahoney, 2015; Rudi et al., 2015; Ma et al., 2015)
- Random feature learning (Rudi and Rosasco, 2017)
- Quadrature (Bach, 2017).

Open questions: Link between leverage score and density?



"In this experiment, the data points $x_i \in (0, 1)$ have been generated with a distribution symmetric about 1, having a high density on the borders of the interval $(0, 1)$ and a low density on the center of the interval. [...] We can see that there are few data points with high leverage, and those correspond to the region that is underrepresented in the data sample (i.e. the region close to the center of the interval since it is the one that has the lowest density of observations)." (Alaoui and Mahoney, 2015)

Outline

Main result

- For a class of translation-invariant kernels K on $\mathbb{R}^d$
  - E.g., Sobolev space of functions with squared integrable derivatives of order up to $s > d/2$
- For the population λ-ridge leverage score

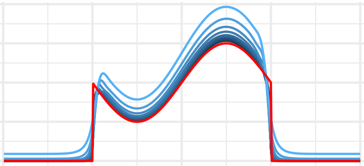
$$\forall z \in \mathbb{R}^d, \quad L_\lambda(z) = \left\langle k(z,\cdot), (\Sigma + \lambda I)^{-1} k(z,\cdot) \right\rangle_{\mathcal{H}_K}$$

- We have, for any $z \in \mathbb{R}^d$ with $p(z) > 0$:

$$L_\lambda(z) \underset{\lambda\to 0,\, \lambda>0}{\sim} L_0 \lambda^{-d/(2s)} p(z)^{d/2s-1}$$

Remarks

$$L_\lambda(z) \underset{\lambda \to 0,\, \lambda > 0}{\sim} L_0 \lambda^{-d/(2s)} p(z)^{d/2s-1}$$



- Explicit relationship between leverage score and density
- Leverage score can be used for density estimation and outlier detection
- May suggest new ways to estimate the leverage score
- Not valid for all kernels (e.g., Gaussian is too smooth)

Outline

Regularized Christoffel function

- Christoffel function, for $l \in \mathbb{N}$:

$$\Lambda_l(z) = \min_{P \in \mathbb{R}_l[X]} \int (P(x))^2 p(x) dx \quad \text{such that} \quad P(z) = 1\,,$$

- NEW: Regularized Christoffel function, for $\lambda > 0$

$$C_\lambda(z) = \inf_{f \in \mathcal{H}} \int_{\mathbb{R}^d} f(x)^2 p(x) dx + \lambda \|f\|_{\mathcal{H}}^2 \quad \text{such that} \quad f(z) = 1\,.$$

- Link with leverage score

$$\forall z \in \mathbb{R}^d\,, \quad C_\lambda(z) = \frac{1}{L_\lambda(z)}$$

Proof sketch

- We study the asymptotics of C_λ
- We show, under some assumptions on the kernel $K(x, y) = q(x - y)$, that:

$$C_\lambda(z) \underset{\lambda \to 0,\, \lambda > 0}{\sim} p(z) D\left(\frac{\lambda}{p(z)}\right),$$

where

$$\begin{aligned} D(\lambda) &:= \min_{f \in \mathcal{H}} \int_{\mathbb{R}^d} f(x)^2 dx + \lambda \|f\|_{\mathcal{H}}^2 \text{ subject to } f(0) = 1 \\ &= \frac{(2\pi)^d}{\int_{\mathbb{R}^d} \frac{\hat{q}(\omega)}{\lambda + \hat{q}(\omega)} d\omega} \end{aligned}$$

Conclusion

- Leverage scores are classical tools in statistics, which gained importance in ML for sketching, sampling, approximating
- We propose a variational formulation of leverage scores, that is an extension of Christoffel functions
- This allows to prove that, under some assumptions on the kernel, leverage scores and proportional to a negative power of the density
- This can suggest new ways to estimate leverage scores, and clarifies why they can be used for density estimation and outlier detection

THANK YOU

References

A. Alaoui and M. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

F. R. Bach. Sharp analysis of low-rank kernel matrix approximations. In *COLT*, volume 30, pages 185–209, 2013.

S. Chatterjee and A. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

K. Clarkson and D. Woodruff. Low rank approximation and regression in input sparsity time. In *ACM symposium on Theory of computing*, pages 81–90. ACM, 2013.

P. Drineas, M. Magdon-Ismail, M. Mahoney, and D. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

D. Hoaglin and Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32 (1):17–22, 1978.

P. Ma, M. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, 16(1):861–911, 2015.

M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

References (cont.)

M. W. Mahoney and P. Drineas. Cur matrix decompositions for improved data analysis. *Proc Natl Acad Sci U S A*, 106(3):697–702, Jan 2009. doi: 10.1073/pnas.0803205106. URL http://dx.doi.org/10.1073/pnas.0803205106.

A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3218–3228, 2017.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

P. Velleman and R. Welsch. Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242, 1981.

S. Wang and Z. Zhang. Improving cur matrix decomposition and the nyström approximation via adaptive sampling. *Journal of Machine Learning Research*, 14(1):2729–2769, 2013.