SENSITIVITY OF LEVERAGE SCORES AND COHERENCE FOR RANDOMIZED MATRIX ALGORITHMS*

ILSE C. F. IPSEN† AND THOMAS WENTWORTH‡

The sampling strategies in many randomized matrix algorithms are, either explicitly or implicitly, controlled by statistical quantities called *leverage scores*. We present bounds for the sensitivity of leverage scores.

Leverage Scores. *Statistical leverage scores* were introduced in 1978 by Hoaglin and Welsch [8] to detect outliers when computing regression diagnostics, see also [3, 12]. To be specific, consider the least squares problem $\min_x \|Ax - b\|_2$, where A is a real $m \times n$ matrix with $\text{rank}(A) = n$. The so-called *hat matrix* $H \equiv A(A^TA)^{-1}A^T$ is the orthogonal projector onto range(A), and determines the *fit* $\hat{b} \equiv Hb$.

The diagonal elements of the hat matrix are called *leverage scores* of A,

$$\ell_j(A) \equiv H_{jj}, \qquad 1 \leq j \leq m,$$

because $\ell_j(A)$ reflects the leverage of the jth point b_j on the corresponding fit $\hat{b}_j$. To see this, suppose that $\ell_k(A) = 1$ for some k. Then $\hat{b}_k = b_k$. Because b_k has maximal leverage, it completely determines the corresponding element of the fit. That is, a degree of freedom has been sacrificed to completely fit b_k. In contrast, if $\ell_k(A) = 0$ then b_k has zero leverage on the fit $\hat{b}_k$.

Leverage scores can be stably computed from a thin QR decomposition $A = QR$, where Q is $m \times n$ with orthonormal columns, via $\ell_j(A) = \|e_j^TQ\|_2^2$. Leverage scores can also be expressed in terms of those left singular vectors of A that are associated with the non-zero singular values.

Leverage scores are the basis for many sampling strategies in randomized matrix computations [10], including low rank approximations [6], CUR decompositions [7], subset selection [1], Nyström approximations [11], least squares [5], and matrix completion [2].

Coherence. The largest leverage score is called *coherence* of A,

$$\mu(A) \equiv \max_{1 \leq j \leq m} \ell_j(A).$$

The general notion of *mutual coherence between two bases* was introduced in 2001 by Donoho and Huo [4], to capture the difficulty of recovering a matrix from sampling.

The above quantity $\mu(A)$ reflects the mutual coherence between an orthonormal basis for range(A) and a canonical basis. In the extreme case of $\mu(A) = 1$ at least one basis vector must be a canonical vector. At the other extreme, if $\mu(A) = n/m$, then all leverage scores are identical, and orthonormal bases for range(A) can be expressed in terms of columns of a Hadamard matrix.

*Both authors were supported in part by NSF grant CCF-1145383. The first author also acknowledges the support from the XDATA Program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323 FA8750-12-C-0323.

†Department of Mathematics, North Carolina State University, P.O. Box 8205, Raleigh, NC 27695-8205, USA, (ipsen@ncsu.edu, http://www4.ncsu.edu/~ipsen/)

‡Department of Mathematics, North Carolina State University, P.O. Box 8205, Raleigh, NC 27695-8205, USA (thomas_wentworth@ncsu.edu)

Low coherence is crucial for the effectiveness of sampling, because it means that all leverage scores are almost identical and the "mass" of an orthonormal basis is uniformly distributed. For instance, suppose we want to uniformly sample $c \geq n$ rows from a $m \times n$ matrix Q with orthonormal columns. What is the probability that the resulting $c \times n$ matrix Q_s also has columns that are close to orthonormal? In [9] we show that with probability at least δ, the two-norm condition number is $\|Q_s\|_2 \|Q_s^\dagger\|_2 \leq 10$, provided the number of sampled rows is at least

$$c \geq 2.7m\, \mu(Q)\, \ln(2n/\delta).$$

In the case of minimal coherence $\mu(Q) = n/m$, we need to sample only $\mathcal{O}(n \ln n)$ rows. However, if a column of Q is a canonical vector then $\mu(Q) = 1$, and a randomized algorithm based on uniform sampling is unlikely to extract a full-rank matrix Q_s.

Sensitivity of Leverage Scores. In order to gauge the sensitivity of leverage scores, we consider a real $m \times n$ matrix B of full column rank, and compare the leverage scores of B to those of A. Since the leverage scores of A are the diagonal elements of the orthogonal projector H, they are basis-independent. Therefore we bound $\ell_j(B)$ in terms of the principal angles between range(A) and range(B).

We start with the simple case where both A and B have orthonormal columns. In the SVD $A^T B = U \Sigma V^T$ the matrices U and V are $n \times n$ orthogonal matrices, and $\Sigma = \mathrm{diag}\begin{pmatrix}\cos\theta_1 & \dots & \cos\theta_n\end{pmatrix}$. The singular values $\cos\theta_1 \geq \cdots \geq \cos\theta_n \geq 0$ are the canonical correlations, and $0 \leq \theta_1 \leq \cdots \leq \theta_n \leq \pi/2$ are the principal angles between range(A) and range(B).

The SVD implies $\hat{A}^T \hat{B} = \Sigma$ where $\hat{A} \equiv AU$ and $\hat{B} \equiv BU$ are the canonical vectors. Right multiplication by orthogonal matrices does not change the leverage scores, hence $\ell_j(\hat{A}) = \ell_j(A)$ and $\ell_j(\hat{B}) = \ell_j(B)$. Thus it suffices to work with the leverage scores of the canonical vectors. Then we can show that

$$\ell_j(B) \leq \left(\cos\theta_1 \sqrt{\ell_j(A)} + \sin\theta_n \sqrt{1 - \ell_j(A)}\right)^2, \qquad 1 \leq j \leq m$$

$$\ell_j(A) \leq \left(\cos\theta_1 \sqrt{\ell_j(B)} + \sin\theta_n \sqrt{1 - \ell_j(B)}\right)^2.$$

In the special case where range(B) = range(A), the two inequalities imply that $\ell_j(B) = \ell_j(A)$, $1 \leq j \leq m$. This suggests that the leverage scores of two orthonormal matrices are close if and only if the angles between their column spaces are small.

For the coherence the above inequalities imply

$$\mu(A)/\gamma_1 \leq \mu(B) \leq \gamma_1\, \mu(A), \qquad \text{where} \qquad \gamma_1 \equiv \left(\cos\theta_1 + \sin\theta_n \sqrt{\tfrac{m}{n} - 1}\right)^2.$$

This suggests that the sensitivity of the coherence increases in proportion to the largest angle between the column spaces, and the aspect ratio of the matrix dimensions.

In the case of large leverage scores, where $\ell_k(A) \geq 1/2$ and $\ell_k(B) \geq 1/2$ for some k, we obtain

$$\ell_k(A)/\gamma_2 \leq \ell_k(B) \leq \gamma_2\, \ell_k(A), \qquad \text{where} \qquad \gamma_2 \equiv (\cos\theta_1 + \sin\theta_n)^2.$$

This suggests that large leverage scores tend to be less sensitive than small ones.

At the other extreme, if $\ell_k(A) = 0$ for some k, then $\ell_k(B) \leq \sin^2\theta_n$, which suggests that small leverage scores are sensitive in the presence of large principal angles between the column spaces.

Conclusions. The above bounds imply that leverage scores of two matrices with orthonormal columns are close if the principal angles between the column spaces are small, and that large leverage scores tend to be less sensitive than small ones. Furthermore, the coherence tends to be more sensitive for tall and skinny matrices, and in the presence of perturbations that cause a large rotation of the column space.

It is not clear what the above bounds imply for the numerical stability of importance sampling strategies based on leverage scores. On the one hand, since the sampling strategies favour rows with large leverage scores, they should be numerically stable when the matrix contains a well-defined set of large leverage scores.

On the other hand, though, importance sampling strategies are most efficient for tall and skinny matrices with large aspect ratios m/n, but they amplify the sensitivity of the coherence. Furthermore many randomized algorithms preprocess the matrix with a fast transform whose purpose is to uniformize all the leverage scores, and in particular remove all the large ones.

To gain more insight, we will consider bounds for almost uniform leverage scores whose values vary in a small neighbourhood of n/m; and also for a set of well-defined set of large leverage scores separated by a definite gap from the smaller ones.

We will also extend the above results to general $m \times n$ full-column rank matrices A and $A + E$, and express the bounds in terms of the perturbation E. Furthermore, given a general rank parameter k, we will determine the sensitivity of leverage scores computed from the best rank-k approximation of A.

REFERENCES

[1] C. Boutsidis, M. W. Mahoney, and P. Drineas, *An improved approximation for the column subset selection problem.* arXiv:0812.4293v2, 12 May 2010.

[2] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[3] S. Chatterjee and A. S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Statist. Sci., 1 (1986), pp. 379–393.

[4] D. L. Donoho and X. Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2845–2862.

[5] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Sampling algorithms for ℓ_2 regression and applications*, in Proc. 17th Annual ACM-SIAM Symp. Discrete Alg. (SODA), 2006, pp. 1127–1136.

[6] ———, *Subspace sampling and relative error matrix approximation: Column-based methods*, in Proc. APPROX-RANDOM, 2006, pp. 316–326.

[7] ———, *Relative-error CUR matrix decompositions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 844–881.

[8] D. C. Hoaglin and R. E. Welsch, *The Hat matrix in regression and ANOVA*, Amer. Statist., 32 (1978), pp. 17–22.

[9] I. C. F. Ipsen and T. Wentworth, *The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems*, (2012). arXiv:1203.4809v1.

[10] M. W. Mahoney, *Randomized Algorithms for Matrices and Data*, Now Publishers Inc., 2011.

[11] A. Talwalkar and R. Rostamizadeh, *Matrix coherence and the Nyström method*, in Proc. 26th Conf. Uncertainty in Artificial Intelligence (UAI-10), AUAI Press, Corvalis, Oregon, 2010, pp. 572–579.

[12] P. F. Velleman and R. E. Welsch, *Efficient computing of regression diagnostics*, Amer. Statist., 35 (1981), pp. 234–242.