

CORRECTION ON SUCCESSIVE NORMALIZATION OF RECTANGULAR ARRAYS

BY RICHARD OLSHEN* AND BALAJI BRAJARATNAM†

*Stanford University** †

Ann. Statist. (38) (2010) 1638–1664

We report an error in Theorem 4.1 of the paper identified in the title. That theorem is not true in general. However, the main mathematical message in the paper as well as all subsidiary mathematical remarks and all simulations and numerical results are correct. As the title suggests, the paper concerns successive normalization of rectangular arrays. We learned the algorithm from Bradley Efron. Here we sketch the algorithm, state the mistake in our paper, sketch our approach to proving its main theorem, and give reference to a paper by the authors where many more details can be found.

For a given matrix, Efron’s algorithm involves the following four successive steps at each iteration:

1. Mean polish each of the J columns.
2. Standard deviation polish each column.
3. Mean polish each of the I rows.
4. standard deviation polish each row.

These four steps, which constitute one iteration, are repeated until “convergence.”

THEOREM 4.1 of the paper by Olshen and Rajaratnam [2] is false as it stands. Writing $\mathcal{G}_{2m-1}^{(i)}$ correctly does not change things. One reason is that, in the notation of the paper, typically

$$E\{X_{i\pi(1)}^{(2m-1)}\}^2 I_B I_Q\} \text{ is not}$$

$$E\{X_{i\pi(1)}^{(1)}\}^2 I_B I_Q\} .$$

Efforts towards finding a non-trivial sub- σ -field of $\mathcal{G}_{2m-1}^{(i)}$ for which there might be equality failed. However, the goal of the theorem does obtain; but

it is a corollary, not a tool, in proving the a.s. convergence of Efron's algorithm applied to $\mathbf{X}_{n \times k} \sim N_{n \times k}(\mathbf{0}, \mathbf{I})$. That is, if, as is the case, the argument from the bottom of page 1648 through the end of Section 4 (of [2]) can be proven without leaning on THEOREM 4.1, then not only does the purpose for THEOREM 4.1 hold, but in fact much more. Namely, write $\mathcal{F}_i^{(m)}$ for the σ -field of the first display on page 1647. For each fixed i , these σ -fields are nested (decreasing) in m . Furthermore, off the set of probability 0 implicit in (9), page 1646, $\sum_{i,j} (X_{ij}^{(q)})^2 = IJ$ for all (i, j, q) . We assume henceforth that \mathbf{x} lies outside this "cursed" set. Therefore, for any real-valued, continuous function f on $\mathcal{R}^{n \times k}$, $\lim_{m \rightarrow \infty} E\{f(X_{ij}^{(q)}) | \mathcal{F}_i^{(m)}\}$ exists almost surely for every fixed q (as m increases without bound), but also, as a consequence of Theorem 2 of the paper by Blackwell and Dubins [1], the same holds with q replaced by m .

Now we sketch an argument by which one shows directly that simultaneously in (i, j) , $X_{ij}^{(m)}$ converges a.s. as m increases without bound. First, we study $(S_j^{(2m-1)})^2$ as in the last display on page 1648, and we sketch the argument that $P\{\overline{\lim}_m (S_j^{(2m-1)})^2 > 0\} = 1$. As in [2], let $A = \{\overline{\lim}_m (S_j^{(2m-1)})^2 = 0\}$. For a more detailed argument, please see [3].

Since the entries of \mathbf{X} are independent, \mathbf{X} is row and column exchangeable. This property is inherited by $\mathbf{X}^{(q)}$ for every q . Because all entries (for $q \geq 1$) are bounded, $E\{X_{ij}^{(q)}\}$ and $E\{(X_{ij}^{(q)})^2\}$ exist and are finite (with fixed bound that applies to all q). Exchangeability implies that all $E\{X_{ij}^{(q)}\} = 0$, and all $E\{(X_{ij}^{(q)})^2\} = 1$. Bounded convergence implies that if $(S_j^{(2m-1)})^2$ tends to 0 along a subsequence as m increases, not only is the limit bounded as a function of \mathbf{x} , but also the limit random variable has expectation 0. Necessarily every almost sure subsequential limit in m of the random variables $\overline{X}_{\cdot j}^{(2m-1)}$ has mean 0. Likewise, every almost sure subsequential limit in m of the random variables $(X_{ij}^{(2m-1)})^2$ has expectation 1. All are bounded as functions of \mathbf{x} . One consequence of these things is that $P(A) = 0$. Please note that the statements about $\underline{\lim}_m X_{ij}^{(2m-1)}$ and $\overline{\lim}_m X_{ij}^{(2m-1)}$ are wrong.

Continue to the paragraph on page 1649 that begins, "Again, let ...". Define m_q as in the paper. The cardinality of E is at least 2. (In fact, the cardinality of E exceeds 2.) Now proceed to the next paragraph. We define the last three displays on page 1649 slightly differently; thus, we require that along m_q

$\lim_{m_q} X_{i_0j}^{(2m_q-1)}$ and $\lim_{m_q} X_{i_0j}^{(2m_q)}$ both exist;

$\lim_{m_q} X_{i_{1j}}^{(2m_q-1)}$ and $\lim_{m_q} X_{i_{1j}}^{(2m_q)}$ both exist; and

$\lim_{m_q} |X_{i_0j}^{(2m_q)} - X_{i_{1j}}^{(2m_q)}|$ exists and is positive.

If there were no such subsequence, our proof would be complete (though previous arguments show that this is not a possibility). Proceed to the first two displays on page 1650. The argument for $S_j^{(2m-1)}$ tending to 1 holds not because the cited difference tends to 0 (something we don't know yet), but instead because of the third display. The remainder stays the same. THEOREM 4.2 follows.

THEOREM 4.2. So long as I and J are at least 3, Efron's algorithm converges almost surely for \mathbf{X} , and therefore on a Lebesgue set of entries with complement a set of Lebesgue measure 0.

REFERENCES

1. BLACKWELL, D. and DUBINS, L. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* **33** (3) 882–886.
2. OLSHEN, R.A. and RAJARATNAM, B. (2010). Successive normalization of rectangular arrays. *Ann. Statist.* **38** (3) 1638–1664.
3. Olshen, R.A. and Rajaratnam, B. “Successive standardization of rectangular arrays,” *Algorithms* **5**(1) (29 February 2012), 98-112. doi: 10.3390/a5010098

RICHARD A. OLSHEN
DEPARTMENTS OF HEALTH RESEARCH AND POLICY,
ELECTRICAL ENGINEERING, AND STATISTICS
STANFORD UNIVERSITY
STANFORD, CA 94305-4065, USA
E-MAIL: olshen@stanford.edu

BALA RAJARATNAM
DEPARTMENTS OF STATISTICS
AND OF ENVIRONMENTAL
EARTH SYSTEM SCIENCE,
STANFORD UNIVERSITY
STANFORD, CA 94305-4065, USA
E-MAIL: brajarat@stanford.edu