



Published in final edited form as:

Ann Stat. 2010 June 1; 38(3): 1638–1664. doi:10.1214/09-AOS743.

SUCCESSIVE NORMALIZATION OF RECTANGULAR ARRAYS

Richard A. Olshen and

Depts. of Health Research and Policy, Electrical Engineering, and Statistics, Stanford, CA 94305-5405, U.S.A.

Bala Rajaratnam

Department of Statistics, Stanford, CA 94305-4065, U.S.A.

Richard A. Olshen: olshen@stanford.edu; Bala Rajaratnam: brajarat@stanford.edu

Abstract

Standard statistical techniques often require transforming data to have mean 0 and standard deviation 1. Typically, this process of “standardization” or “normalization” is applied across subjects when each subject produces a single number. High throughput genomic and financial data often come as rectangular arrays, where each coordinate in one direction concerns subjects, who might have different status (case or control, say); and each coordinate in the other designates “outcome” for a specific feature, for example “gene,” “polymorphic site,” or some aspect of financial profile. It may happen when analyzing data that arrive as a rectangular array that one requires BOTH the subjects and features to be “on the same footing.” Thus, there may be a need to standardize across rows and columns of the rectangular matrix. There arises the question as to how to achieve this double normalization. We propose and investigate the convergence of what seems to us a natural approach to successive normalization, which we learned from colleague Bradley Efron. We also study the implementation of the method on simulated data and also on data that arose from scientific experimentation.

Keywords and phrases

normalization; standardization; backwards martingale convergence theorem

1. Introduction

This paper is about a method for normalization, or regularization, of large rectangular sets of numbers. In recent years many statistical efforts have been directed towards inference on such rectangular arrays. The exact geometry of the array matters little to the theory that follows. Positive results apply to the situation where there are at least three rows and at least three columns. We explain difficulties that arise when either numbers only two. Scenarios to which methodology studied here applies tend to have many more rows than columns. Data can be from gene expression microarrays, SNP (single nucleotide polymorphism) arrays, protein arrays, alternatively from large scale problems in imaging. Often there is one column per subject, with rows consisting of real numbers (as in expression) or numbers 0, 1, 2 (as with SNPs). Subjects from whom data are gathered may be “afflicted” or not, with a condition that while heritable is far from Mendelian. A goal is to find rows, better groups of rows, by which to distinguish afflicted from other subjects. One can be led to testing many statistical hypotheses simultaneously, thereby separating rows into those that are “interesting” for further follow-up and those that seem not to be. Genetic data tend to be

analyzed by test “genes” (rows), beginning with their being “embedded” in a chip, perhaps a bead. There may follow a subsequent molecule that binds to the embedded “gene”/molecule. A compound that makes use of the binding preferences of nucleotides and to which some sort of “dye” is attached is then “poured.” The strength of binding depends upon affinity of the “gene” or attached molecule and the compound. Laser light is shined on the object into which the test “gene” has been embedded, and from its bending, the amount of bound compound is assessed, from which the amount of the “gene” is inferred. The basic idea is that different afflicted status may lead to different amounts of “gene”.

With the cited formulation and ingenious technology, data may still suffer from problems that have nothing to do with differences between groups of subjects or with differences between “genes” or groups of them. There may be differences in background, by column, or even by row. Perhaps also “primers” (compounds) vary across columns for a given row. For whatever reasons, scales by row or column may vary in ways that do not enable biological understanding. Variability across subjects could be unrelated to afflicted status.

Think now of the common problem of comparing variables that can vary in their affine scales. Because covariances are not scale-free, it can make sense to compare in dimensionless coordinates that are centered at 0, that is, where values of each variable have respective means subtracted off, and are scaled by respective standard deviations. That way, each variable is somehow “on the same footing”.

Standardization, or normalization, studied here is done precisely so that both “subjects” and “genes” are “on the same footing”. We recognize one might require only that “genes” (or some “genes”) be on the same footing, and the same for “subjects.” The successive transformations studied here apply when one lacks *a priori* opinions that might limit goals. Thus, “genes” that result from the standardization we study are transformed to have mean 0 and standard deviation 1 across all subjects, while the same is true for subjects across all “genes”. How to normalize? One approach is to begin with, say, row, though one could as easily begin with columns. Subtract respective row means and divide by respective standard deviations. Now do the same operation on columns, then on rows, and so on. Remarkably, this process tends to converge, even rapidly in terms of numbers of iterations, and to a set of numbers that have the described good limiting properties in terms of means and standard deviations, by row and by column.

In this paper we show by examples how the process works and demonstrate for them that indeed it converges. We also include rigorous mathematical arguments as to why convergence tends to occur. Readers will see that the process and perhaps especially the mathematics that underlies it are not as simple as we had hoped they would be. This paper is only about convergence, which is demonstrated to be exponentially fast (or faster) for examples. The mathematics here does not apply directly to “rates”. The Hausdorff dimension of the limit set seems easy enough to study. Summaries will be reported elsewhere.

2. Motivating Example

We introduce a motivating example to ground the problem that we address in this paper. Consider a simple 3-by-3 matrix with entries generated from a uniform distribution on $[0, 1]$. We standardize the initial matrix $X^{(0)}$ by row and column, first subtracting the row mean from each entry and then dividing each entry in a given row by its row standard deviation. The matrix is then column standardized by subtracting the column mean from each entry and then by dividing each entry by the respective column standard deviation. In this section, these four steps of row mean polishing, row standard deviation polishing, column mean polishing and column standard deviation polishing entail one iteration in the process of

attempting to row and column standardize the matrix. After one such iteration, the same process is applied to resulting matrix $X^{(1)}$ and the process repeated with the hope that successive renormalization will eventually yield a row and column standardized matrix. Hence these four steps are repeated until “convergence” - which we define as the difference in the Frobenius norm between two consecutive iterations being less than 10^{-8} .

In order to illustrate this numerically, we start with the following 3-by-3 matrix with independent entries generated from a uniform distribution on $[0,1]$ and repeat the process described above.

$$X^{(0)} = \begin{bmatrix} 0.1182 & 0.7069 & 0.4145 \\ 0.9884 & 0.9995 & 0.4648 \\ 0.5400 & 0.2878 & 0.7640 \end{bmatrix} \quad (1)$$

The successive normalization algorithm took 9 iterations to converge. The initial matrix, the final solution, and relative (and log relative) difference for the 9 iterations are given below (see also figure 1):

$$X^{(final)} = \begin{bmatrix} -1.2608 & 1.1852 & 0.0756 \\ 1.1852 & 0.0757 & -1.2608 \\ 0.0756 & -1.2608 & 1.1852 \end{bmatrix} \quad (2)$$

$$\text{Successive Difference} = \begin{bmatrix} \text{Iteration no.} & \text{difference} & \text{log(difference)} \\ 1 & 8.7908 & 2.1737 \\ 2 & 0.5018 & -0.6895 \\ 3 & 0.0300 & -3.5057 \\ 4 & 0.0019 & -6.2862 \\ 5 & 0.0001 & -9.0607 \\ 6 & 0.0000 & -11.8337 \\ 7 & 0.0000 & -14.6064 \\ 8 & 0.0000 & -17.3790 \\ 9 & 0.0000 & -20.1516 \end{bmatrix} \quad (3)$$

The whole procedure of 9 iterations takes less than 0.15 seconds on a standard modern laptop computer. We also note that the final solution has effectively 3 distinct entries. When other random starting values are used, we observe that convergence patterns can vary in the sense that convergence may not be monotonic. The plots below (see Figure 2) capture the type of convergence patterns that are observed in our simple 3-by-3 example.

Despite the different convergence patterns that are observed, when our successive renormalization is repeated with different starting values - a surprising phenomenon surfaces. The process seems always to converge, and moreover the convergence is very rapid. One is led naturally to ask whether this process will always converge and if so under what conditions. These questions lay the foundation for the work in this paper.

3. Preliminaries

We establish the notation that we will use by revisiting a normalization/standardization method that is traditional for multivariate data. If the main goal of a normalization of a rectangular array is achieving zero row and column averages, then a natural approach is to “mean polish” the row (i.e., subtract the row mean from every entry of the rectangular

array), followed by a column “mean polish”. This cycle of successive row and column polishes is repeated until the resulting rectangular array has zero row and column averages. The following theorem proves that this procedure attains a double mean standardized rectangular array in one iteration where an iteration is defined as constituting one row mean polish followed by one column mean polish.

Lemma 3.1

Given an initial matrix $\mathbf{X}^{(0)}$, an iterative procedure to cycle through repetitions of a row mean polish followed by a column mean polish until convergence terminates in one step.

Proof—Let $\mathbf{X}^{(0)}$ be an $n \times k$ matrix and define the following:

$$\mathbf{X}^{(0)} = [X_{ij}^{(0)}]$$

$$\bar{X}_i^{(0)} = \frac{1}{k} \sum_{j=1}^k X_{ij}^{(0)}$$

Now the first part of the iteration, termed as a “row mean polish” subtracts from each element its respective row mean:

$$\mathbf{X}^{(1)} = [X_{ij}^{(1)}] = X_{ij}^{(0)} - \bar{X}_i^{(0)}$$

The second step of the iteration, termed a “column mean polish” subtracts from each element of the current matrix its respective column mean:

$$\mathbf{X}^{(2)} = [X_{ij}^{(2)}] = X_{ij}^{(1)} - \bar{X}_{.j}^{(1)},$$

where

$$\bar{X}_{.j}^{(1)} = \frac{1}{n} \sum_{i=1}^n X_{ij}^{(1)}$$

After the second step of the iteration it is clear that the columns sum to zero; the previous operation enforces this. In order to prove that the iterative procedure terminates at the second part of the iteration it is sufficient to show that the rows of the current iterate sum to zero. Now note that

$$\begin{aligned} \mathbf{X}^{(2)} &= [X_{ij}^{(2)}] \\ &= [X_{ij}^{(1)}] - \bar{X}_{.j}^{(1)} \\ &= (X_{ij}^{(0)} - \bar{X}_i^{(0)}) - \left(\frac{1}{n} \sum_{r=1}^n X_{rj}^{(1)} \right) \\ &= (X_{ij}^{(0)} - \bar{X}_i^{(0)}) - \left(\frac{1}{n} \sum_{r=1}^n (X_{rj}^{(0)} - \bar{X}_r^{(0)}) \right) \end{aligned}$$

The remaining is to show that the row sum of this matrix $\mathbf{X}^{(2)}$ expressed as the elements of $\mathbf{X}^{(0)}$ sum to zero. So,

$$\begin{aligned}
\sum_{j=1}^k X_{ij}^{(2)} &= \sum_{j=1}^k \left(X_{ij}^{(0)} - \bar{X}_i^{(0)} \right) - \sum_{j=1}^k \left(\frac{1}{n} \sum_{r=1}^n \left(X_{rj}^{(0)} - \bar{X}_r^{(0)} \right) \right) \\
&= \left(k\bar{X}_i^{(0)} - k\bar{X}_i^{(0)} \right) - \frac{1}{n} \sum_{r=1}^n \sum_{j=1}^k \left(X_{rj}^{(0)} - \bar{X}_r^{(0)} \right) \\
&= \left(k\bar{X}_i^{(0)} - k\bar{X}_i^{(0)} \right) - \frac{1}{n} \sum_{r=1}^n \left(k\bar{X}_r^{(0)} - k\bar{X}_r^{(0)} \right) \\
&= 0 - 0 \\
&= 0
\end{aligned}$$

Note that the above double standardization is implicit in a 2-way ANOVA, and though not explicitly stated it can be deduced from the work of Scheffé [9]. It is nevertheless presented here first in order to introduce notation, second as it is not available in this form above in the ANOVA framework, but third for the intuition it gives since it is a natural precursor to the subject of work in the remainder of this paper.

3.1. Example 1(cont.): A 3-by-3 example with only mean polishing

We proceed to illustrate the previous theorem on the motivating example given after the introduction and draw contrasts between the two approaches. As expected the successive normalization algorithm terminates in one iteration. The initial matrix, the final solution, and the column and row standard deviations of the final matrix are given below:

$$Y^{(0)} = \begin{bmatrix} 0.1182 & 0.7069 & 0.4145 \\ 0.9884 & 0.9995 & 0.4648 \\ 0.5400 & 0.2878 & 0.7640 \end{bmatrix} \quad (4)$$

$$Y^{(column-polished)} = \begin{bmatrix} -0.4307 & 0.0422 & -0.1333 \\ 0.4396 & 0.3347 & -0.0829 \\ -0.0089 & -0.3769 & 0.2162 \end{bmatrix} \quad (5)$$

$$Y^{(row-polished)} = Y^{(final)} = \begin{bmatrix} -0.2568 & 0.2161 & 0.0407 \\ 0.2091 & 0.1043 & -0.3134 \\ 0.0477 & -0.3204 & 0.2727 \end{bmatrix} \quad (6)$$

$$Std(columns) = \begin{bmatrix} 0.1932 & 0.2311 & 0.2410 \end{bmatrix} \quad (7)$$

$$Std(rows) = \begin{bmatrix} 0.1952 \\ 0.2257 \\ 0.2445 \end{bmatrix} \quad (8)$$

We note unlike in the motivating example, and as expected, the row and column means are both 0; but the standard deviations of the rows and the columns are not identical, let alone identically 1. Since mean polishing has already been attained, and we additionally require that row and column standard deviations to be 1, it is rather tempting to row and column standard deviation polish the terminal matrix $Y^{(final)}$ above. We conclude this example by

observing the simple fact that doing so results in the loss of the zero row and column averages.

3.2. The 2-by-2 problem

We now examine the successive row and column mean and standard deviation polishing for a 2×2 matrix and hence illustrate that for the results in this paper to hold true, the minimum of row(k) and column dimension(n) of the matrix under consideration must be at least 3, that

is $\min(k, n) \geq 3$. Consider a general 2×2 matrix: $\mathbf{X}^{(0)} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. If $a < b$ and $c < d$, then

after one row normalization, $\mathbf{X}^{(1)} = \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}$; so $(S_j^{(1)})^2 = 0$. Therefore, allowing for both inequalities to be reversed, and, assuming that, for example, a, b, c , and d are *iid* with

continuous distribution(s), then $P((S_j^{(1)})^2 = 0) = 1/2$, in which case the procedure is no longer well-defined.

A moment's reflection shows that if \mathbf{X} is $n \times 2$, with n odd, then after each row normalization, each column has an odd number of entries, each entry being -1 or $+1$.

However, each row has exactly one -1 and one $+1$. Thus $(S_i^{(m)} - S_j^{(m+1)})^2 \rightarrow 0$ occurs only on a set of product Lebesgue measure 0. With $\min(k, n) \geq 3$, both tend to 1 a.e. as $m \nearrow \infty$. Henceforth, assume $\min(k, n) \geq 3$.

4. Theoretical Considerations

For a matrix \mathbf{X} as defined, take $\mathbf{x} \in \mathcal{R}^{n \times k}$. Let λ denote Lebesgue measure on $\mathcal{R}^{n \times k}$, and P be the probability on the coordinates that renders them *iid* $N(0, 1)$. That is, $\mathbf{X}_{ij}(x_{ij}) = \mathbf{x}_{ij}$. \mathbf{X}_{ij} are *iid* $N(0, 1)$. What is more, for any orthogonal linear transformation \mathcal{O} that preserves orthogonality, $\mathcal{O}\mathbf{X} \sim \mathbf{X}$, that is, $\mathcal{O}\mathbf{X}$ is distributed as \mathbf{X} Muirhead [8], Anderson [1].

Because λ and P are mutually absolutely continuous, if $\mathcal{C} = \{\text{algorithm for successive row and column normalization converges}\}$, then $P(\mathcal{C}) = 1$ iff $\lambda(\mathcal{R}^{n \times k} \setminus \mathcal{C}) = 0$, though only one direction is used.

For the remainder, assume that P governs \mathbf{X} . Positive results will be obtained for $3 \leq \min(n, k) \leq \max(n, k) < \infty$. Our arguments require notation, to which we turn attention now. We redefine the notation and symbols introduced in Lemma 3.1 as we now bring in row and column standard deviation polishing. Define $\mathbf{X} = \mathbf{X}^{(0)}$.

$$\begin{aligned} \bar{\mathbf{X}}_i^{(0)} &= \frac{1}{k} \sum_{j=1}^k \mathbf{X}_{ij}; \\ (s_i^{(0)})^2 &= \frac{1}{k} \sum_{j=1}^k (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i^{(0)})^2 \\ &= \left(\frac{1}{k} \sum_{j=1}^k \mathbf{X}_{ij}^2 \right) - (\bar{\mathbf{X}}_i^{(0)})^2 \end{aligned}$$

$\mathbf{X}^{(1)} = [\mathbf{X}_{ij}^{(1)}]$, where $\mathbf{X}_{ij}^{(1)} = (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i^{(0)}) / S_i^{(0)}$; a.s. $(S_i^{(0)})^2 > 0$ since $k \geq 3 > 1$.

By analogy, set

$$\begin{aligned} \mathbf{X}^{(2)} &= [\mathbf{X}_{ij}^{(2)}], \text{ where } \mathbf{X}_{ij}^{(2)} = (\mathbf{X}_{ij}^{(1)} - \overline{\mathbf{X}}_{.j}) / S_j^{(1)}; \\ \overline{\mathbf{X}}_{.j}^{(1)} &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{ij}^{(1)}; \\ (S_j^{(1)})^2 &= \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_{ij}^{(1)} - \overline{\mathbf{X}}_{.j}^{(1)})^2. \end{aligned}$$

Arguments sketched in what follows entail that a.s. $(S_j^{(1)})^2 > 0$ since $n \geq 3$.

For m odd, $\mathbf{X}_{ij}^{(m)} = (\mathbf{X}_{ij}^{(m-1)} - \overline{\mathbf{X}}_{i.}^{(m-1)}) / S_i^{(m-1)}$, with $\overline{\mathbf{X}}_{i.}^{(m-1)}$ and $(S_i^{(m-1)})^2$ defined by analogy to previous definitions.

For m even, $\mathbf{X}_{ij}^{(m)} = (\mathbf{X}_{ij}^{(m-1)} - \overline{\mathbf{X}}_{.j}^{(m-1)}) / S_j^{(m-1)}$, again with $\overline{\mathbf{X}}_{.j}^{(m-1)}$ and $S_j^{(m-1)}$ defined by analogy to earlier definitions.

Back to the general problem

We first note that because the process we study is a coordinate process, there is no difference between regular conditional probabilities and regular conditional distributions (see Durrett [4], Section 4.1.c, p33 and pp. 229–331 for more details). They can be computed as densities with respect to Lebesgue measure on a finite Cartesian product $\times Sph(q)$, where $q = k$ refers to after row normalization and where $q = n$ if subsequent to column normalization. In a slight abuse of notation, for any positive integer q we define $Sph(q) = \{\mathbf{x} \in \mathcal{R}^q: \|\mathbf{x}\|^2 = q\}$, where the norm is the usual Euclidean norm.

Let $\{r_{ij}: i = 1, \dots, n; j = 1, \dots, k\}$ be a set of $n \cdot k$ rational numbers, not all 0. Obviously,

$$P(\cup_{r_1, \dots, r_{nk}} \sum_{i,j} r_{ij} \mathbf{X}_{ij} = 0) = 0.$$

An inductive argument involving conditional densities shows that

$$P(\cup_{m=1}^{\infty} \cup_{r_1, \dots, r_{nk}} \sum_{i,j} r_{ij} \mathbf{X}_{ij}^{(m)} = 0) = 0. \quad (9)$$

Consequently

$$P((\cap_{m=1}^{\infty} \cap_{i=1}^n (S_i^{(2m)})^2 > 0) \cap (\cap_{m=1}^{\infty} \cap_{j=1}^k (S_j^{(2m-1)})^2 > 0)) = 1.$$

Further, a.s. $\mathbf{X}^{(m)}$ is defined and finite for every m .

What we know about the t distribution Efron [5] and geometric arguments entail that $\mathbf{X}^{(1)}$ can be viewed as having probability distribution on $\mathcal{R}^{n \times k}$ that is the n -fold product of independent uniform distributions on $Sph(k) \times \dots \times Sph(k)$. For the sake of clarity we note again that conditional probabilities are always taken to be “regular” and concentrated on the relevant product of spheres. Readers will note that in the cited arguments for

$\sum_{i,j} r_{ij} \mathbf{X}_{ij}^{(m)}$, $(S_i^{(2m)})^2$, and $(S_j^{(2m-1)})^2$, relevant conditional probabilities and densities are used. Of course, after the m^{th} row standardization $\mathbf{X}^{(2m-1)} = [\mathbf{X}_{ij}^{(2m-1)} : i=1, \dots, n; j=1, \dots, k]$, and analogously for column standardization and $\mathbf{X}^{(2m)}$.

As an aside, write $g_1(\mathbf{X}) = \mathbf{X}^{(1)}$ on $\{\min(S_i^{(0)})^2 > 0\}$. Elsewhere, let $g_1(\mathbf{X}) = 0$. Necessarily g_1 depends on \mathbf{X} only through its direction in $\mathcal{R}^{n \times k}$. Equivalently, g_1 is a function of \mathbf{X} that is homogeneous of degree 0. Moreover, $g_1(\mathbf{X}) \sim g_1(\mathcal{O}\mathbf{X})$ for all orthogonal linear \mathcal{O} on $\mathcal{R}^{n \times k}$. $\mathbf{X}^{(1)}$ has independent rows, each uniformly distributed on $Sph(k)$.

We turn now to study $\mathbf{X}_{ij}^{(2m-1)}$ as m increases without bound. Note first that for $m = 1, 2, \dots$ $\mathbf{X}^{(2m-1)}$ has joint distribution that is unchanged if two columns of \mathbf{X} are transposed, therefore if two columns of $\mathbf{X}^{(2m-1)}$ are transposed. Since every permutation is a product of transpositions, $\mathbf{X}^{(2m-1)}$ is column exchangeable. Each is also row exchangeable.

Write π for a permutation of the integers $\{1, \dots, k\}$; let be the finite σ -field of all subsets of $\{\pi\}$. The marginal probability induced on $\{\pi\}$ from the joint distribution of $(\mathbf{X}, \{\pi\})$ is discrete and uniform, assigning probability $1/k!$ to each π .

Write $\mathcal{G}_{2m-1}^{(i)}$ to be the σ -field

$$\mathcal{F}([\mathbf{X}_{ij}^{(q)} : j=1, \dots, k; q=2m-1, 2m+1, \dots]) \times \Pi.$$

THEOREM 3.2— $E\{(\mathbf{X}_{i\pi(1)}^{(1)})^2 | \mathcal{G}_{2m-1}^{(i)}\} = (\mathbf{X}_{i\pi(1)}^{(2m-1)})^2$ a.s. for $m = 1, 2, \dots$

Proof: Write $(\mathbf{X}_{i\pi(1)}^{(2m-1)})^2 = \sum_{l=1}^k (\mathbf{X}_{il}^{(2m-1)})^2 I_{\{\pi(1)=l\}}$, where I_A is the indicator function of the event A . Obviously, $(\mathbf{X}_{i\pi(1)}^{(2m-1)})^2$ is $\mathcal{G}_{2m-1}^{(i)}$ measurable; $\{\alpha_{q_1, q_2}\}$ is a set of real numbers, doubly indexed by $\{\alpha_{q_1, q_2}\} \in F$; F is a finite subset of $\{1, \dots, k\} \times \{1, 2, \dots\}$. Form $B = \{(\mathbf{X}_{iq_1}^{(2m-1+q_2)})^2 \leq \alpha_{q_1, q_2}; (q_1, q_2) \in F\}$.

Note that $\mathcal{G}_{2m-1}^{(i)}$ is generated by $\{B \times Q\}$, B of the cited form and $Q \in \Pi$. In particular, each $B \times Q \in \mathcal{G}_{2m-1}^{(i)}$. Proof of our claim is complete if we show that for $m = 2, 3, \dots$

$$\int_{B \times Q} (\mathbf{X}_{i\pi(1)}^{(2m-1)})^2 = \int_{B \times Q} (\mathbf{X}_{i\pi(1)}^{(1)})^2.$$

The left hand side of the display can be expressed

$$E\left\{\sum_{l=1}^k (\mathbf{X}_{il}^{(2m-1)})^2 I_{\{\pi(1)=l\}} I_{\{\pi \in Q\}} I_B\right\} = E\left\{I_B \sum_{l=1}^k (\mathbf{X}_{il}^{(2m-1)})^2 I_{\{\pi(1)=l\}} I_{\{\pi \in Q\}}\right\}.$$

Now, for any π , the expression inside the sum is $(k)(1/k) = 1$ if $\pi \in Q$ and 0 if not. That is, the expression constituting the sum is $k I_{[\pi(1)=1]} I_{[\pi \in Q]}$. Now the, expectation factors into $P(B)P(\pi(1) = 1 | \pi \in Q) P(\pi \in Q) = P(B)P(Q)$. Retracing steps shows clearly that $(X_{il}^{(2m-1)})^2$ in the computation just completed could be replaced by $(X_{il}^{(1)})^2$ with all equalities remaining true. The claim is now proven.

Convergence of $(X_{ij}^{(m)})^2$

The backwards martingale convergence theorem Doob [3] entails that $(X_{i\pi(1)}^{(2m-2)})^2$ converges a.s. as $m \rightarrow \infty$. So, for each fixed $j \in \{1, \dots, k\}$, $(X_{i\pi(1)}^{(2m-1)})^2 I_{[\pi(1)=j]}$ converges a.s. It follows that $[(X_{ij}^{(2m-1)})^2]$ converges a.s. as $m \rightarrow \infty$.

If previous arguments are perturbed so that π denotes a permutation of $\{1, \dots, n\}$, with $(X_{i\pi(1)}^{(1)})^2$ replaced by $(X_{\pi(1)j}^{(2)})^2$, $\mathcal{G}_{2m-1}^{(i)}$ by $\mathcal{G}_{2m}^{(i)}$, $(X_{i\pi(1)}^{(2m-1)})^2$ by $(X_{\pi(1)j}^{(2m)})^2$, and $\sum_{l=1}^k (X_{il}^{(2m-1)})^2$ by $\sum_{l=1}^n (X_{lj}^{(2m)})^2$, then one concludes that also $[(X_{ij}^{(2m)})^2]$ converges a.s. as $m \rightarrow \infty$. Without further argument it is unclear the a.s. limits along odd, respectively even, indices are the same; and it is crucial to what remains that this is in fact true.

Obviously, $\bigcap_{m=1}^{\infty} \mathcal{G}_{2m-1}^{(i)} = \bigcap_{m=1}^{\infty} \mathcal{G}_{2m}^{(i)}$, so in a certain sense measurability is the same. Obviously, too, randomization of index is by columns in the first case and by rows in the second. But now a path to the required conclusion presents itself. Given success in proving a.s. convergence along odd indices after randomizing columns and along even indices randomizing rows, and given a requirement of our approach is that these two limits be identical a.s., perhaps there is a path by simultaneously randomizing both columns and rows? Fortunately, that is the case. Thus, let π_1 be a permutation of $\{1, \dots, n\}$ and π_2 be a permutation of $\{1, \dots, k\}$. With obvious product formulation of governing probability mechanism and further obvious formulation of decreasing σ -fields, as an example of what can be proved,

$$E((X_{\pi_2(1)\pi_2(1)}^{(1)})^2 | \mathcal{G}_2) = (X_{\pi_1(1)\pi_2(1)}^{(2)})^2 \text{ a.s.}$$

From this arguments for the display there are several paths by which one concludes that a.s., simultaneously for all (i, j) , $[(X_{ij}^{(m)})^2]$ converges. Dominated convergence entails that the limit random matrix has expectation 1 in all coordinates. As a consequence of this convergence, a.s. and simultaneously for all (i, j) , $[(X_{ij}^{(2m+1)})^2] - [(X_{ij}^{(2m)})^2] \rightarrow 0$ as $m \rightarrow \infty$.

Convergence of $[X_{ij}^{(m)}]$

We turn now to key ideas in extending our argument that $[(X_{ij}^{(m)})^2]$ converges almost surely simultaneously to the same conclusion with the square removed. To limit notational complexity, we study first only odd indices as m grows without bound. Conclusions are identical for even indices, and by extension for indices not constrained to be odd or even.

A first necessary step is to show that for arbitrary j ,

$$P\{\overline{\lim}_m (S_j^{(2m+1)})^2 > 0\} = 1$$

To that end, let A be the event $[\overline{\lim}_m (S_j^{(2m-1)})^2 = 0]$. Obviously,

$A = \{\mathbf{x} : (S_j^{(2m-1)})^2 \rightarrow 0\} = \{\mathbf{x} : S_j^{(2m-1)} \rightarrow 0\}$ regardless of square roots taken. We show that $P\{A\} = 0$.

By way of contradiction, suppose that $P\{A\} > 0$. Write

$$(S_j^{(2m-1)})^2 = \frac{1}{n} \sum_{l=1}^n (\mathbf{X}_{lj}^{(2m-1)})^2 - (\overline{\mathbf{X}}_{\cdot j}^{(2m-1)})^2$$

We know that the first term tends to 1 a.s. on A . Therefore, also the second term tends to 1 a.s. on A . Since for $m > 1$, $\mathbf{X}_{lj}^{(2m-1)}$ is bounded a.s., $\overline{\lim}_m \max_l \mathbf{X}_{lj}^{(2m-1)}(\mathbf{x})$ is a finite-valued random variable $C = C(\mathbf{x})$. Simple considerations show that the only possibilities are that for all l , $C = +1$ or $C = -1$ a.s. Similar arguments show that $\underline{\lim}_m \min_l \mathbf{X}_{lj}^{(2m-1)}(\mathbf{x})$ is $+1$ or -1 . It is clear that there is no sequence $\{m_k\}$ of $\{m\}$ along which

$$-1 < \underline{\lim}_m \min_l \mathbf{X}_{lj}^{(2m_k-1)}(\mathbf{x}) \leq \overline{\lim}_m \max_l \mathbf{X}_{lj}^{(2m_k-1)}(\mathbf{x}) < 1$$

It follows that $\lim_m \mathbf{X}_{lj}^{(2m-1)}$ exists a.s. on A , and that the limit of the sequence is $+1$ or -1 on A .

Recall that $\mathbf{X} \sim -\mathbf{X}$, and this equality is inherited by all joint distributions of $\mathbf{X}^{(m)}$. However, when \mathbf{X} is replaced by its negative, any limit of $\mathbf{X}_{lj}^{(2m-1)}$ becomes its negative; a.s. convergence is a property of the probability distribution of \mathbf{X} ; and $\{\mathbf{x} : \mathbf{X}_{lj}^{(2m-1)} \text{ converges}\} = \{\mathbf{x} : -\mathbf{X}_{lj}^{(2m-1)} \text{ converges}\}$. Therefore, the only possibility is that $C = 0$, which entails that $(\overline{\mathbf{X}}_{\cdot j}^{(2m-1)})^2 \rightarrow 0$ and $(S_j^{(2m-1)})^2 \rightarrow 1$. The upshot is that $P\{A\} = 0$; $P\{\overline{\lim}_m (S_j^{(2m-1)})^2 > 0\} = 1$.

Again, let us fix j . Consider a sample path of $\{\mathbf{X}^{(2m_q-1)}\}$ along which $\lim_m (S_j^{(2m_q-1)})^2 = D > 0$. Clearly, $\{i : \overline{\lim}_{m_q} |\mathbf{X}_{ij}^{(2m_q-1)}| > 0\} \neq \emptyset$. Indeed, let $E = E(j) = \{i : \text{for some } \{m_q\} = \{m_q(i)\}, \overline{\lim}_{m_q(i)} |\mathbf{X}_{ij}^{(2m_q-1)}| > 0\}$. Row and column exchangeability of $\mathbf{X}^{(m)}$ entail that necessarily the cardinality of E is at least 2.

Let $i_0 \neq i_1 \in E$. Because $\min(n, k) \geq 3$, there is a further subsequence of $\{\{m_q(i)\}\}$ – for simplicity again write it $\{m_q\}$ – along which

$$\begin{aligned} & \lim_{m_q} |\mathbf{X}_{i_0 j}^{(2m_q-1)}| \text{ and } \lim_{m_q} |\mathbf{X}_{i_0 j}^{(2m_q)}| \text{ both exist;} \\ & \lim_{m_q} |\mathbf{X}_{i_1 j}^{(2m_q-1)}| \text{ and } \lim_{m_q} |\mathbf{X}_{i_1 j}^{(2m_q)}| \text{ both exist; and} \\ & \lim_{m_q} |\mathbf{X}_{i_0 j}^{(2m_q)} - \mathbf{X}_{i_1 j}^{(2m_q-1)}| \text{ exists and is positive.} \end{aligned}$$

The first two requirements can always be met off of the set of probability 0 implicit in eqn(9). That the third can be met as well is a consequence of the argument just concluded. In any case, if there were no such subsequence, then our proof would be complete because all $\mathbf{X}_{ij}^{(m)}$ for j fixed tend to the same number. But now, write

$$\begin{aligned} & (\mathbf{X}_{i_0 j}^{(2m_q)} - \mathbf{X}_{i_0 j}^{(2m_q-1)}) - (\mathbf{X}_{i_1 j}^{(2m_q)} - \mathbf{X}_{i_1 j}^{(2m_q-1)}) = \\ & (\mathbf{X}_{i_0 j}^{(2m_q-1)} - \mathbf{X}_{i_1 j}^{(2m_q-1)})(S_j^{(2m_q-1)} - 1)/S_j^{(2m_q-1)}. \end{aligned}$$

Since $(\mathbf{X}_{i_0 j}^{(2m_q)})^2 - (\mathbf{X}_{i_0 j}^{(2m_q-1)})^2 \rightarrow 0$ a.s., and likewise with i_0 replaced by i_1 , the first expression of the immediately previous display has limit 0. Thus, so too does the second expression. This is possible only if $S_j^{(2m_q-1)} \rightarrow 1$ (where we have taken the positive square root). Further,

$$\mathbf{X}_{i_0 j}^{(2m_q)} - \mathbf{X}_{i_0 j}^{(2m_q-1)} = \frac{\mathbf{X}_{i_0 j}^{(2m_q-1)}(S_j^{(2m_q-1)} - 1) + \overline{\mathbf{X}}_{\cdot j}^{(2m_q-1)}}{S_j^{(2m_q-1)}}$$

As a corollary to the above, one sees now that $\overline{\mathbf{X}}_{\cdot j}^{(2m_q-1)} \rightarrow 0$. Since the original $\{m_q\}$ could be taken to be an arbitrary subsequence of $\{m\}$, we conclude that:

- $S_j^{(2m_q-1)} \rightarrow 1$ a.s.;
- $\overline{\mathbf{X}}_{\cdot j}^{(2m_q-1)} \rightarrow 0$ a.s.; and
- $\mathbf{X}_{ij}^{(m)}$ converges a.s.

Now replace arguments for (i) and (ii) on columns by analogous arguments on rows. Deduce that every infinite subsequence of positive integers has a subsequence along which our desired conclusion obtains.

5. Properties of successive normalization

We now comment on theoretical properties of successive normalization. In particular, we elaborate on the generality of the result by showing that the Gaussian assumption is not necessary and serves only as a convenient choice of measure. We also discuss convergence in Lebesgue measure and the domains of attraction of successive normalization.

5.1. Choice of probability measure

Write λ for Lebesgue measure on $\mathcal{R}^{n \times k}$. Thus, $\mathbf{x} \in \mathcal{R}^{n \times k}$ is an $n \times k$ rectangular array of real numbers. Let P be a measure on the Borel sets of \mathcal{B} of $\mathcal{R}^{n \times k}$ that is mutually absolutely

continuous with respect of λ . By this is meant that if $B \subset \mathcal{R}^{nk}$ is Borel measurable, then $\lambda(B) = 0$ iff $P(B) = 0$. One obvious example of such a P is the measure on \mathcal{B} that renders its nk coordinates *iid* Gaussian. If x_{rc} is the (r, c) coordinate of $\mathbf{X} \in \mathcal{R}^{nk}$, and $P^{(r,c)}$ is the marginal measure of P on its (r, c) real coordinate, then $P^{(r,c)}((-\infty, x_0]) = \Phi(x_0)$, where Φ is the standard normal cumulative distribution function. This is what we mean by P henceforth. The *iid* specification entails that P is an nk -fold product of identical probabilities, and because P is now defined uniquely for all product rectangles, necessarily it is defined uniquely for all $B \in \mathcal{B}$. It is obvious that Φ being mutually absolutely continuous with respect to one-dimensional Lebesgue measure means that the product measure P is mutually absolutely continuous with respect to the nk -fold Borel product measure λ .

Now, let f_1, f_2, \dots be a sequence of \mathcal{B} -measurable functions, $\mathcal{R}^{nk} \rightarrow \mathcal{R}^{nk}$. Then $\{f_m \text{ converges}\} = \{\overline{\lim} f_m = \underline{\lim} f_m \text{ simultaneously for all } nk \text{ coordinates}\}$. When $f_m(\mathbf{x})$ converges, then $\lim f_m(\mathbf{x}) = \overline{\lim} f_m(x) I_{\{f_m \text{ converges}\}}$, where for any set C , I_C is its indicator. Because each f_m is \mathcal{B} -measurable, $\lim f_m(\mathbf{x})$ is also \mathcal{B} -measurable. If we are given that $\{f_m\}$ converges P -almost everywhere, then $\{f_m \text{ converges}\}$ is a Borel set of P -measure 1. Necessarily its complement has P -measure 0, therefore λ measure 0. It follows that $\{f_m\}$ converges except for a Borel subset of \mathcal{R}^{nk} of Lebesgue measure 0.

In the present paper, the (r, c) coordinate of $\{f_m\}$ is the set of successive normalizations of the initial real entry multiplied by the indicator of the subset of \mathcal{R}^{nk} that is $\{\text{successive normalizations possible}\}$. The latter is clearly a Borel subset of \mathcal{R}^{nk} of P -measure 1, so its complement is a Borel set of λ -measure 0. It is immaterial to convergence whether the first initialization is by row or by column. Readers should note that any study of asymptotic properties of $\{f_m\}$ under P may entail that $\{f_m(\mathbf{x})\}$ has properties such as row and column exchangeability, where the coordinate functions are taken to be random variables. They should note, too, that: (i) changing the original measure to one more conducive to computation is standard, and is what happens with “exponential tilting” as it applies to the study of large deviations of sums; and (ii) the interplay between measure and topology, as is utilized here, is a standard approach in probability theory that is applied seldom to statistical arguments; the lack thus far owes only to necessity.

5.2. Convergence in p^{th} mean for Lebesgue measure

Whenever standardization is possible, after one standardization the sum over all nk coordinates of squares of respective values is bounded by nk it follows from dominated convergence that as m grows without bound, each term converges not only P -almost everywhere but also in p^{th} mean for every $\infty > p \geq 1$ so long as P remains the applicable measure. Because λ is not a finite measure, convergence in p^{th} mean for underlying measure λ does not follow. It is impossible for such convergence to apply to Lebesgue measure on the full Euclidean space \mathcal{R}^{nk} . This is because there is a set of positive Lebesgue measure whose members are not fixed points for normalization by both rows and columns. No matter which normalization comes first in any infinite, alternating sequence, the normalization is invariant to fixed scale multiples of each $\mathbf{x} \in \mathcal{R}^{nk}$, in particular to fixed scale multiples of \mathbf{x} in the set of positive Lebesgue measure just described. Obviously, no further arguments are required, and convergence in p^{th} mean is immediate if Lebesgue measure λ is restricted to a bounded subset of \mathcal{R}^{nk} .

5.3. Domains of attraction

Reviewers of research presented here have wondered if we can describe simply what successive and alternating normalization does to rectangular arrays of data, beyond introductory comments about putting rows and columns on an equal footing and the analogy to computing correlation from covariance. We begin our reply here, though details await

further research and a subsequent paper. Please remember invariance to either row or column normalization (when possible), to scale multiples of $\mathbf{x} \in \mathcal{R}^{n \times k}$. In other words, results of normalization are constant along rays defined by these multiples, and without loss of generality we can assume that \mathbf{x} lies in an n -fold product of $Sph(k)$, where each of the n components is orthogonal to the linear one-dimensional subspace of \mathcal{R}^k consisting of its equiangular line; call it $Sph(k) \setminus \{\mathbf{1}\}$. Thus, without loss we assume that the object under study is $\mathbf{X}^{(m)}$ subject to a row normalization of the process of successive normalization. We study only subsets of the n -fold product space that was described and that is complementary to $\{\cup_{m=1}^{\infty} \cup_{r_1, \dots, r_k} \sum_{i,j} r_{i,j} \mathbf{X}^{(m)} = 0\}$, where $\{r_{ij}\}$ are rational. The set in braces just before the comma has P-measure 0; and we know that on its complement, $X^{(m)}$ can be defined for all $m = 1, 2, \dots$

Because normalization always involves subtraction of a mean and division by a standard deviation, and because each $\mathbf{X}^{(m)}$ is row and column exchangeable, the limiting process we study here when P applies seems on superficial glance to be analogous to “domains of attraction” as that notion applies to sequences of *iid* random variables. One obvious difference is that here limits are almost sure rather than in distribution. While a.s. limits of $\mathbf{X}^{(m)}$ are shown to have row and column means 0 and row and column standard deviations 1, $n \times k$ arrays of real numbers with this property are obviously the only fixed points of the alternating process studied here. The Hausdorff dimension of the set of fixed points is not difficult to compute, but we have been unable thus far to give rigorously supported conditions for the domain of attraction (in the sense described) of each fixed point. The simple case for which domains of attraction for limits in distribution were described was a major development in the history of probability (see Feller [6], Gnedenko and Kolmogorov [7], Zolotarev [10]). We report some intuitive results, and next a mathematical question that arose in our study of domains of attraction for which at present we have only a heuristic argument.

Is there a set $E \subset \mathcal{R}^{n \times k}$ for which $P(E) = 1$; $X^{(m)}(\mathbf{x})$ converges for $\mathbf{x} \in E$; and for each fixed i $\lim \mathbf{X}_{ij}(\mathbf{x}) \neq \lim \mathbf{X}_{ij'}(\mathbf{x})$ for all $j \neq j'$ if $\mathbf{x} \in E$? Clearly one could ask the equivalent question for each fixed j , and a corresponding subset $E' \subset \mathcal{R}^{n \times k}$. We conjecture the existence of $E \cap E'$. However, arguments available thus far do not confirm existence rigorously. Therefore, suggestions regarding domains of attraction as $\mathbf{X}^{(m)}$ grows without bound should be taken as only heuristic for now.

Given that $(\mathbf{S}_i^{(m)})^2 \rightarrow 1$ on a subset of $\mathcal{R}^{n \times k}$ with complementary P-measure 0, therefore Lebesgue measure 0, almost surely ultimately (meaning for $m = m(\mathbf{x})$ large enough), row normalization does not perturb the (strict) sort of any row. By analogy, a.s. ultimately column normalization does not perturb the (strict) sort of any column. Thus, a.s. ultimately, successive and alternative row and column normalization do not perturb any row or column sort. This joint strict sort determines an open subset of $\mathcal{R}^{n \times k}$. If we restrict ourselves to rows, then we may speak more precisely of n -fold products of sorts of members of $Sph(k) \setminus \{\mathbf{1}\}$. For a fixed row there are $k!$ strict sorts of the squared entries. Given a fixed sort of the squares there are, say, $f(k)$ assignments of sign to the square roots of the k entries so that the sum of entries for that row is 0. For any assignments of signs to the necessarily non-trivial values of the k squares that renders the sum of entries for the row 0, there is always a scaling so that the sum of squares is any fixed value, therefore so that the variance of the set of numbers in that row is 1. One computes $f(3) = 2$; $f(4) = 4$; and so on. Computation of the exact value of $f(k)$ is slightly tricky and is not reported here. For all k , $f(k) \geq 2$. To summarize, for each fixed row there are a.s. ultimately $f(k)k!$ disjoint (open) invariant sets following row normalization, making for $[f(k)k!]^n$ a.s. ultimately (open) invariant sets simultaneously for all n rows. If we count a.s. ultimately invariant sets subsequent to column

normalization, then entirely analogous arguments result in a.s. ultimately $[f(n)n!]^k$ disjoint (open) sets simultaneously for all columns.

From computations, after a row normalization the surface area of the sphere in k -space orthogonal to the equiangular line - that corresponds to only one row of

$\mathbf{X}^{(m)}$ - has surface area $\approx \sqrt{\left(\frac{2}{\epsilon}\right)\left(\frac{2\pi e}{k-3}\right)} < 1$ for $k \geq 21$. The expression $\rightarrow 0$ as $k \nearrow \infty$. Even for $k = 4$, the quantity is only about 14.7 (larger than the actual value). Remember that there are at most $[f(k)k!]$ a.s. ultimately non-empty “invariant sets” for row normalization. Thus, one sees that for k large the quantity $\left(\frac{\text{surface area } Sph(k)}{\text{invariant sets of } Sph(k)}\right)^n$ is nearly 0.

6. Computational Results and Applications

We include three examples to highlight and illustrate some computational aspects of our iterative procedure. The first two examples are studies by simulation whereas the third example is an implementation on a real dataset.

For the simulation study, we consider a 3-by-3 matrix and a 10-by-10 matrix both with entries generated independently from a uniform distribution on $[0,1]$. For a given matrix, the algorithm computes/calculates the following 4 steps at each iteration:

- a. Mean polish the column
- b. Stand deviation polish the column
- c. Mean polish the row
- d. Stand deviation polish the row

These four steps, which constitute one iteration, are repeated until “convergence” - which we define as when the difference in some norm-squared (the quadratic/Frobenius norm in our case) between two consecutive iterations is less than some small prescribed value - which we take to be 0.00000001 or 10^{-8} .

6.1. Example 2: Simulation study on a 10-by-10 dimensional example

We proceed now to illustrate the convergence of the successive row and column mean-standard deviation polishing for the simple 10-by-10 dimensional example cited. The algorithm took 15 iterations to converge. The initial matrix, the final solution, and relative(and log relative) difference for the 15 iterations are:

$$X^0 = \begin{pmatrix} 0.8145 & 0.3551 & 0.7258 & 0.3736 & 0.0216 & 0.2486 & 0.0669 & 0.2178 & 0.6766 & 0.6026 \\ 0.7891 & 0.9970 & 0.3704 & 0.0875 & 0.9106 & 0.4516 & 0.9394 & 0.1821 & 0.9883 & 0.7505 \\ 0.8523 & 0.2242 & 0.8416 & 0.6401 & 0.8006 & 0.2277 & 0.0182 & 0.0418 & 0.7668 & 0.5835 \\ 0.5056 & 0.6525 & 0.7342 & 0.1806 & 0.7458 & 0.8044 & 0.6838 & 0.1069 & 0.3367 & 0.5518 \\ 0.6357 & 0.6050 & 0.5710 & 0.0451 & 0.8131 & 0.9861 & 0.7837 & 0.6164 & 0.6624 & 0.5836 \\ 0.9509 & 0.3872 & 0.1769 & 0.7232 & 0.3833 & 0.0300 & 0.5341 & 0.9397 & 0.2442 & 0.5118 \\ 0.4440 & 0.1422 & 0.9574 & 0.3474 & 0.6173 & 0.5357 & 0.8854 & 0.3545 & 0.2955 & 0.0826 \\ 0.0600 & 0.0251 & 0.2653 & 0.6606 & 0.5755 & 0.0871 & 0.8990 & 0.4106 & 0.6802 & 0.7196 \\ 0.8667 & 0.4211 & 0.9246 & 0.3839 & 0.5301 & 0.8021 & 0.6259 & 0.9843 & 0.5278 & 0.9962 \\ 0.6312 & 0.1841 & 0.2238 & 0.6273 & 0.2751 & 0.9891 & 0.1379 & 0.9456 & 0.4116 & 0.3545 \end{pmatrix} \quad (10)$$

$$X^{final} = \begin{pmatrix} 1.2075 & 0.2139 & 0.8939 & 0.2661 & -2.0026 & -0.5881 & -1.2477 & -0.4157 & 1.1023 & 0.5705 \\ -0.0736 & 1.7222 & -1.2202 & -1.0461 & 0.6465 & -0.8172 & 0.5144 & -1.1740 & 1.3022 & 0.1458 \\ 0.8858 & -0.8659 & 0.8816 & 0.7930 & 0.9515 & -0.9498 & -1.6621 & -1.1469 & 1.0831 & 0.0298 \\ -0.9296 & 1.5223 & 0.6537 & -0.7661 & 0.9476 & 0.9361 & 0.5467 & -1.3402 & -1.3775 & -0.1931 \\ -0.8358 & 0.8041 & -0.7288 & -2.0057 & 1.0328 & 1.4824 & 0.5929 & 0.0202 & 0.2768 & -0.6390 \\ 1.4926 & 0.1374 & -1.2120 & 1.1351 & -0.5035 & -1.2741 & 0.1766 & 1.3125 & -1.1642 & -0.1005 \\ -0.5156 & -0.7494 & 1.5647 & 0.2025 & 0.5610 & 0.2646 & 1.3840 & -0.0059 & -0.7521 & -1.9537 \\ -1.8680 & -1.1055 & -0.6428 & 0.9269 & 0.3515 & -0.8323 & 1.2448 & 0.1167 & 0.8827 & 0.9259 \\ 0.3596 & -1.0158 & 0.8070 & -0.5547 & -1.1339 & 0.1669 & -0.5895 & 1.0581 & -1.0805 & 1.9828 \\ 0.2771 & -0.6632 & -0.9973 & 1.0490 & -0.8509 & 1.6114 & -0.9601 & 1.5752 & -0.2727 & -0.7685 \end{pmatrix} \quad (11)$$

Successive Difference=	Iteration no.	difference	log(difference)
	1	84.1592	4.4327
	2	1.2860	0.2516
	3	0.1013	-2.2897
	4	0.0144	-4.2402
	5	0.0029	-5.8434
	6	0.0007	-7.2915
	7	0.0002	-8.6805
	8	0.0000	-10.0456
	9	0.0000	-11.4000
	10	0.0000	-12.7492
	11	0.0000	-14.0955
	12	0.0000	-15.4403
	13	0.0000	-16.7841
	14	0.0000	-18.1272
	15	0.0000	-19.4699

We note once more how the relative differences decrease linearly on the log scale(though empirically) and is once again suggestive of the rate of convergence. As both the figure (see Fig. 3) and the vector of relative differences indicate, there is a substantial jump at iteration 2; and then the curve behaves linearly.

The whole procedure takes about 0.37 seconds on a standard modern laptop computer and terminates after 15 iterations. It might appear that the increase in the number of iterations increases with increase in dimension. For instance, the number of iterations goes from 9 to 15 as we go from dimension 3 to 10. We should however bear in mind that when we go from dimension 3 to 10 the “tolerance level” is kept constant at 0.00000001. The number of elements that must be close to their respective limiting values, however, goes from 9 in the 3-dimensional case, to 100 in the 10-dimensional case. The rapidity of convergence was explored further, and the process above was repeated over 1000 simulations. The convergence proves to be stable in the sense that the mean and standard deviation of the number of steps till convergence over the 1000 simulations are 14.5230 and 2.0331 respectively. A histogram of the number of steps till convergence is given below (fig. 4).

A closer look at the vector of successive differences suggests that the “bulk of the convergence” is achieved during the first iteration. This seems reasonable since necessarily the first steps render the resulting columns, then, rows, members of their respective unit spheres (and even within cited subsets of them). Convergence then is only within these spheres. Our numerical results also indicate the the sequence of matrices $X^{(i)}$ changes most drastically during this first iteration. It suggests that mean polishing to a larger extent is responsible for the rapidity of convergence and is reminiscent of the result in Lemma 1 -

which states that if only row and column mean polishing are performed, then convergence is achieved immediately. We explore this issue further by looking at the distance between $X^{(1)}$ and $X^{(final)}$, and compare it to the distance between $X^{(0)}$ and $X^{(final)}$. The ratio of these two distance is defined below:

$$Ratio = \frac{dist(X^{(1)}, X^{(final)})}{dist(X^{(0)}, X^{(final)})}$$

For our 10-by-10 example, we simulated 1000 initial starting values and implemented our successive normalization procedure. The average value of the distance from the first iterate to the limit, as a proportion of the total distance to the limit from the starting value, is only 2.78%. One could interpret this heuristically as saying that on average the crucial first step does as much as 97.2% of the work towards convergence. We therefore confirm that the bulk of the convergence is indeed achieved in the first step (termed as a “one-step analysis” from now onwards). The distribution of the ratio defined above is graphed in the histogram below (fig. 5). We also note that none of the 1000 simulations yielded a ratio of over 10%.

Yet another illuminating perspective of our successive normalization technique is obtained when we track the number of sign changes in the individual entries of the matrices from one iteration to the next. Please remember that this is related to the “invariant sets” that were described in subsection 5.3. Naturally, one would expect the vast majority of sign changes to occur in the first step as the bulk of the convergence is achieved during this first step. We record the number of sign changes at each iteration, as a proportion of the total number of sign changes until convergence, over 1000 simulations, in our 10-by-10 case. The results are illustrated in the table below¹ (see Table 1). An empirical study of the occurrence of the sign changes reveals interesting heuristics. We note that on average 95% of sign changes occur during the first step and an additional 3% in the next step. The table also demonstrates that as much as 99% of sign changes occur during the first three iterations. When we examine infrequent cases where there is a sign change well after the first few iterations, we observe that the corresponding limiting value is close to zero, thus indicating that a sign change well into the successive normalization technique (i.e., a change from positive to negative or vice versa) amounts to a very small change in the actual magnitude of the corresponding value of the matrix.

We conclude this example by investigating more thoroughly whether the dimensions of the matrices have an impact on either the rapidity of convergence and/or on the one-step analysis. The following table gives the mean and standard deviations of the number of iterations needed for convergence for various values of the dimension of the matrix - denoted by p and n when keeping the total number of cells in the matrix constant 2. Once more our successive normalization procedure is applied to 1000 uniform random starting values. Results of this exercise are given in the table below (see Table 2).

We find that when n and p are close convergence appears to be faster, keeping everything else constant. Interestingly enough, a one-step analysis performed for the different scenarios above tends to suggest that the onestep ratio, defined as the distance from the first iterate to the limit as a proportion of the total distance to the limit from the starting value, seems largely unaffected by the row or column dimension of the problem.

¹Since the number of iterations to convergence depends on the starting point, the length of the vector of the number of sign changes will vary accordingly. We summarize this vector by averaging over all the 1000 simulations the relative frequency of the number of sign changes for the first nine iterations. The first nine iterations were chosen as each of the 1000 simulations required at least 9 iterations to converge.

²or approximately constant

6.2. Example 3: Simulation study on a 5-by-5 dimensional example

We now proceed to further investigate the successive normalization procedure when one begins with column mean-standard deviation polishing followed by row mean-standard deviation polishing or vice versa on a simple 5-by-5 dimensional example. The theory developed in the previous sections proves convergence of the successive normalization procedure whether the first normalization that is performed on the matrix is row polishing or column polishing.

The algorithm took 30 iterations to converge when one begins with column mean-standard deviation polishing, and when one begins with row mean-standard deviation polishing it took 26 iterations to converge. The initial matrix, the final solutions, log relative differences and their respective plots for both approaches are given below (see Fig. 6).

$$X^0 = \begin{pmatrix} 0.6565 & 0.2866 & 0.7095 & 0.4409 & 0.8645 \\ 0.3099 & 0.3548 & 0.9052 & 0.8758 & 0.0210 \\ 0.3316 & 0.5358 & 0.8658 & 0.8650 & 0.0768 \\ 0.1882 & 0.9908 & 0.1192 & 0.3552 & 0.3767 \\ 0.1007 & 0.0282 & 0.9553 & 0.6311 & 0.1492 \end{pmatrix} \quad (13)$$

$$X_{\text{starting with column polishing}}^{\text{final}} = \begin{pmatrix} 1.6360 & -0.4320 & -0.7863 & -1.0548 & 0.6371 \\ 0.1093 & -1.2446 & 0.9477 & 1.2170 & -1.0295 \\ -0.6979 & 1.1193 & -0.1716 & 1.1399 & -1.3897 \\ 0.2748 & 1.2421 & -1.3091 & -1.0112 & 0.8034 \\ -1.3223 & -0.6848 & 1.3192 & -0.2907 & 0.9786 \end{pmatrix} \quad (14)$$

$$X_{\text{starting with row polishing}}^{\text{final}} = \begin{pmatrix} 1.4956 & -0.4243 & -0.7386 & -1.1620 & 0.8293 \\ 0.3816 & -0.9267 & 0.5915 & 1.3267 & -1.3731 \\ -1.2158 & 1.1775 & 0.1052 & 0.9966 & -1.0634 \\ 0.3478 & 1.2181 & -1.4096 & -0.9138 & 0.7573 \\ -1.0092 & -1.0446 & 1.4514 & -0.2475 & 0.8499 \end{pmatrix} \quad (15)$$

$$\text{Successive Difference} = \begin{pmatrix} \text{Iteration no.} & \text{difference} & \text{log(difference)} \\ \text{---} & \text{starting with} & \text{starting with} \\ \text{---} & \text{column polishing} & \text{row polishing} \\ 1 & 2.9646 & 3.0255 \\ 2 & -0.5858 & 0.2539 \\ 3 & -1.5082 & -0.8731 \\ 4 & -1.8814 & -1.4650 \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ 24 & -15.0375 & -17.0730 \\ 25 & -15.7028 & -17.8229 \\ 26 & -16.3679 & -18.5728 \\ 27 & -17.0331 & --- \\ 28 & -17.6983 & --- \\ 29 & -18.3635 & --- \\ 30 & -19.0287 & --- \end{pmatrix} \quad (16)$$

As expected the final solutions are different. The simulations were repeated with different initial values, and we note that the convergence patterns (as illustrated in Fig. 6) are similar whether the procedure starts with column polishing or row polishing, though actual the number of iterations required to converge can vary.

6.3. Example 4: Gene expression data set: 20426-by-63 dimensional example

We now illustrate the convergence of the successive row and column mean-standard deviation polishing for a real life gene expression example -a 20426-by-63 dimensional example. This dataset arose originally from a study of human in-stent restenosis by Ashley et al. [2]. The algorithm took considerably longer in terms of time and computer resources but converged in eight iterations. The initial matrix and the final are too large to display, but the relative(and log relative) difference for the eight iterations are given subsequently:

$$\text{Successive Difference}=1.0e+005* \begin{pmatrix} \text{Iteration no.} & \text{difference} & \text{log(difference)} \\ 1 & 1.0465 & 11.5583 \\ 2 & 0.0008 & 4.4030 \\ 3 & 0.0000 & -0.2333 \\ 4 & 0.0000 & -4.7582 \\ 5 & 0.0000 & -9.2495 \\ 6 & 0.0000 & -13.7130 \\ 7 & 0.0000 & -18.1526 \\ 8 & 0.0000 & -22.5717 \end{pmatrix} \quad (17)$$

Note once more how the relative differences decreases linearly on the log scale(though empirically) and is once again suggestive of the rate of convergence. As both the figure (see Fig. 7) and the vector of relative differences indicates, there is a jump between iteration 1 and 2 and then the curve behaves linearly.

Additionally the whole procedure takes about 853.2 seconds or approximately 14.22 minutes on a desktop computer³ vs 0.4 seconds for the 10-by-10 example. However, the algorithm terminates after only eight iterations. In this example the number of iterations does NOT change with the increase in dimensionality. It may make sense to investigate this behavior more thoroughly, empirically, using simulation for rectangular but not square matrices. It seems that the ratio of the two dimensions or the minimum of the two dimensions may play a role. We should also bear in mind that the tolerance level, which is the sum of the individual differences squared, has been kept constant at 0.00000001.

7. Conclusion

In this section we attempt to lend perspective to our results and to point the way for future developments. Readers please note that for rectangular $n \times k$ arrays of real numbers with $\min(k, n) \geq 3$, the technique beginning with rows (alternatively columns) and successively subtracting row (column) means and dividing resulting differences, respectively, by row (column) standard deviations converges for a subset of Euclidean $\mathcal{R}^{n \times k}$ whose complement has Lebesgue measure 0. The limit is row and column exchangeable given the Gaussian probability mechanism that applies in our theoretical arguments. We do not offer other information on the nature of the exact set on which successive iterates converge. A single "iteration" of the process we study has four steps, two each respectively for rows and columns. Note that on the set for which the algorithm converges, convergence seems remarkably rapid, exponential or even faster, perhaps because after a half an iteration, the

³2GHz Core 2 Duo processor and 2GB of RAM

rows (alternately columns) lie as n (respectively k) points on the surface of the product of relevant unit spheres. Further iterations adjust only directions, not lengths.

Viewing the squares of the entries as the terms of a backwards martingale entails maximal inequalities for them, and therefore implicitly contains information on “rates of convergence” of the squares; but these easy results appear far from the best one might establish. Our arguments for (almost everywhere) convergence of the original signed entries do not have information regarding rates of convergence. One argues easily that if successive iterates converge, and no limiting entry is 0, then after finitely many steps, (the number depending on the original values and the limiting values) signs are unchanged. In our examples of small dimension, evidence of this can be made explicit. In particular we observe empirically that the vast majority of sign changes that are observed do indeed take place in the first few iterations. Any sign changes that are observed well after the first few iterations correspond to sign changes around entries with limiting values close to zero. We also have no information on optimality in any sense of the iterated transformations we study. One reason for our thinking that our topic is inherently difficult is that we were unable to view successive iterates as “contractions” in any sense familiar to us.

If we take any original set of numbers, and multiply each number by the realized value of a positive random variable with arbitrarily heavy tails, then convergence is unchanged. Normalization entails that after half a single iteration the same points on the surface of the relevant unit spheres are attained, no matter the multiple. The message is that what matters to convergence are the distributions induced on the surfaces of spheres after each half iteration, and not otherwise common heaviness of the tails of the probability distributions of individual entries.

Acknowledgments

The authors gratefully acknowledge Bradley Efron for introducing them to the research question addressed in the paper. They also thank Johann Won and Thomas Quattermous for useful discussions and Thomas Quattermous for granting us access to and understanding of data we study and reference; Bonnie Chung and Cindy Kirby are also acknowledged for administrative assistance. Richard Olshen was supported in part by NIH MERIT Award R37EB02784, and Bala Rajaratnam was supported in part by NSF grant DMS 0505303 DMS 0906392 and NIH grant 5R01 EB001988-11REV

References

1. Anderson, TW. An introduction to multivariate statistical analysis. 3. Wiley; New Jersey: 2003.
2. Ashley EA, Ferrara R, King JY, Vailaya A, Kuchinsky A, He X, Byers B, Gerckens U, Oblin S, Tsalenko A, Soito A, Spin J, Tabibiazar R, Connolly AJ, Simpson JB, Grube E, Quattermous T. Network analysis of human in-stent restenosis, in. *Circulation*. 2006; 114(24):2644–2654. [PubMed: 17145989]
3. Doob JL. Regularity properties of certain functions of chance variables, in. *Transactions of the American Mathematical Society*. 1940; 47(2):455–486.
4. Durrett, R. *Probability: Theory and Examples*. 2. Duxbury Press; Belmont: 1995.
5. Efron B. Student’s t-test under symmetry conditions, in. *Journal of the American Statistical Association*. 1969; 64(328):1278–1302.
6. Feller, W. *An Introduction to Probability Theory and Its Applications*. 2. Vol. 2. Wiley; New York: 1966.
7. Gnedenko, BV.; Kolmogorov, AN. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley; Boston: 1954.
8. Muirhead, RJ. *Aspects of Multivariate Statistical Theory*. Wiley; New York: 1999.
9. Scheffé, H. *The Analysis of Variance*. Wiley; New York: 1999. (Reprint)

10. Zolotarev, VM. One-dimensional Stable Distributions. American Mathematical Society; Providence: 1986.

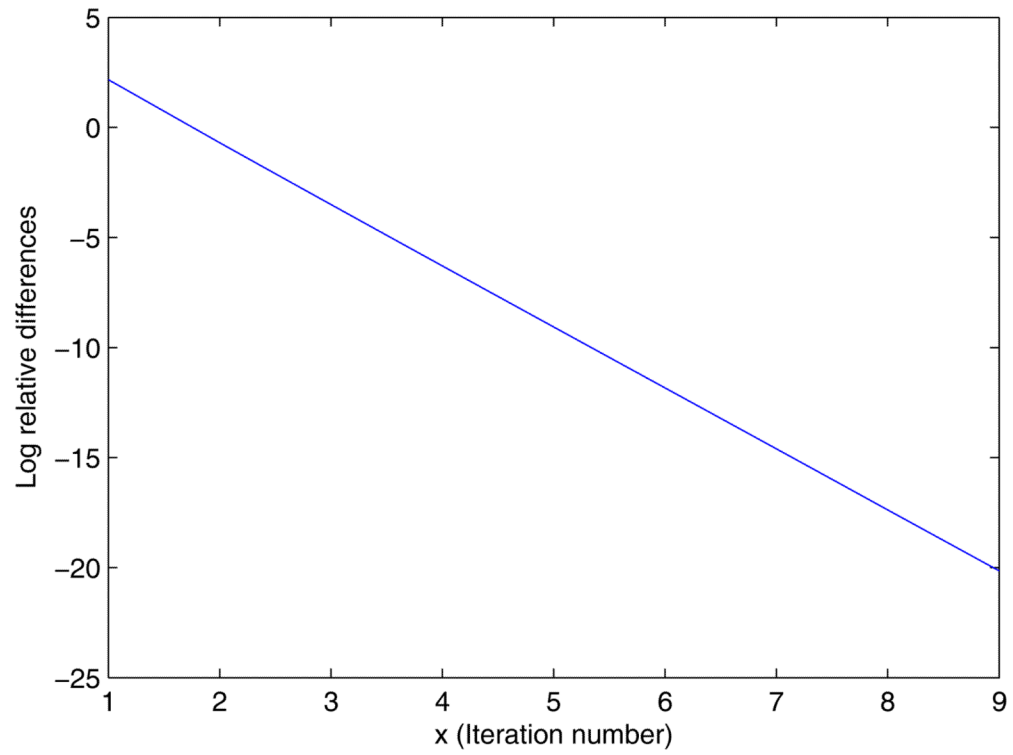


Fig 1.
Relative differences at each iteration on the log scale - 3-by-3 dimensional example

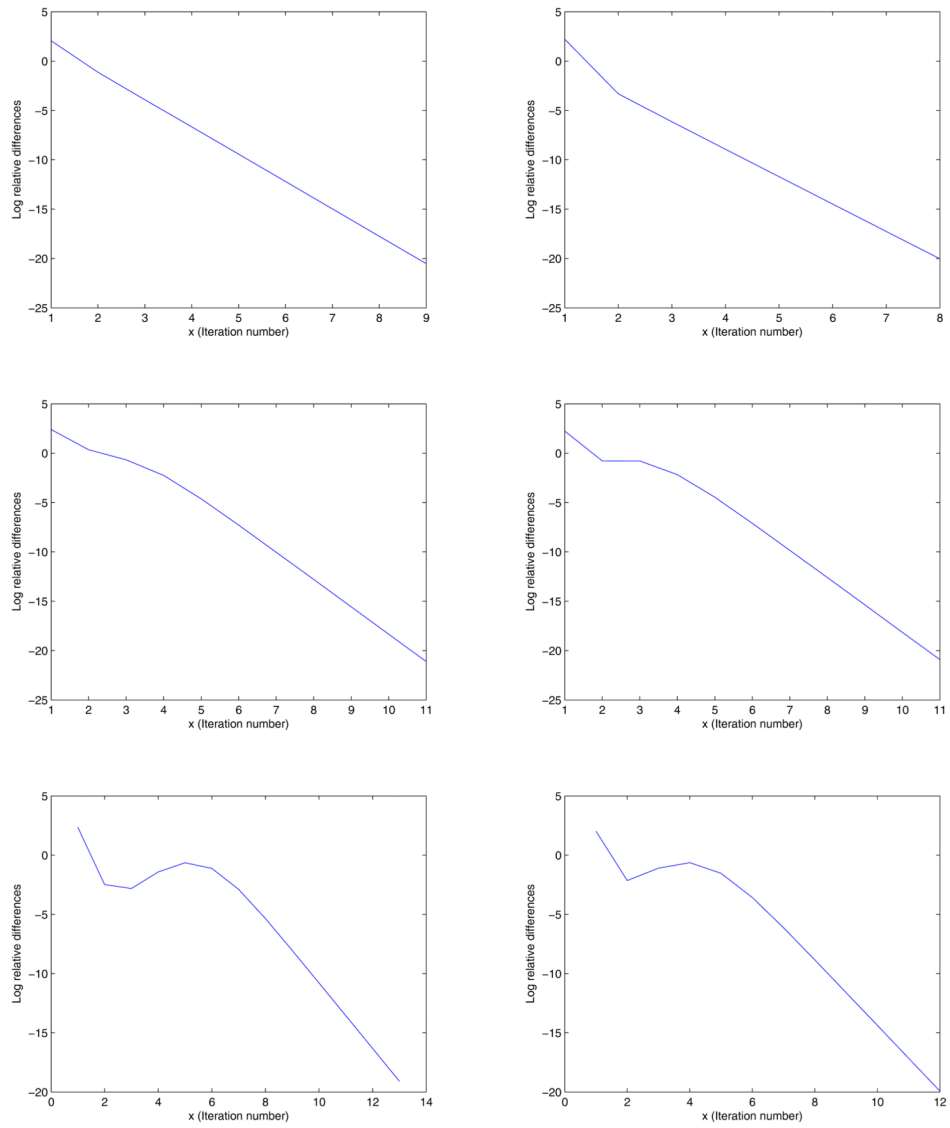


Fig 2. Convergence patterns for the 3-by-3 example

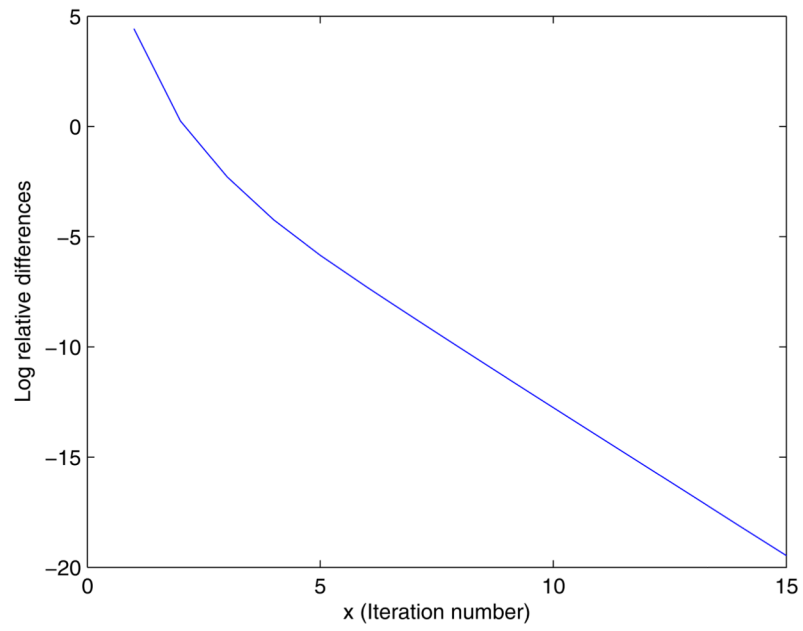


Fig 3.
Relative differences at each iteration on the log scale - 10-by-10 dimensional example

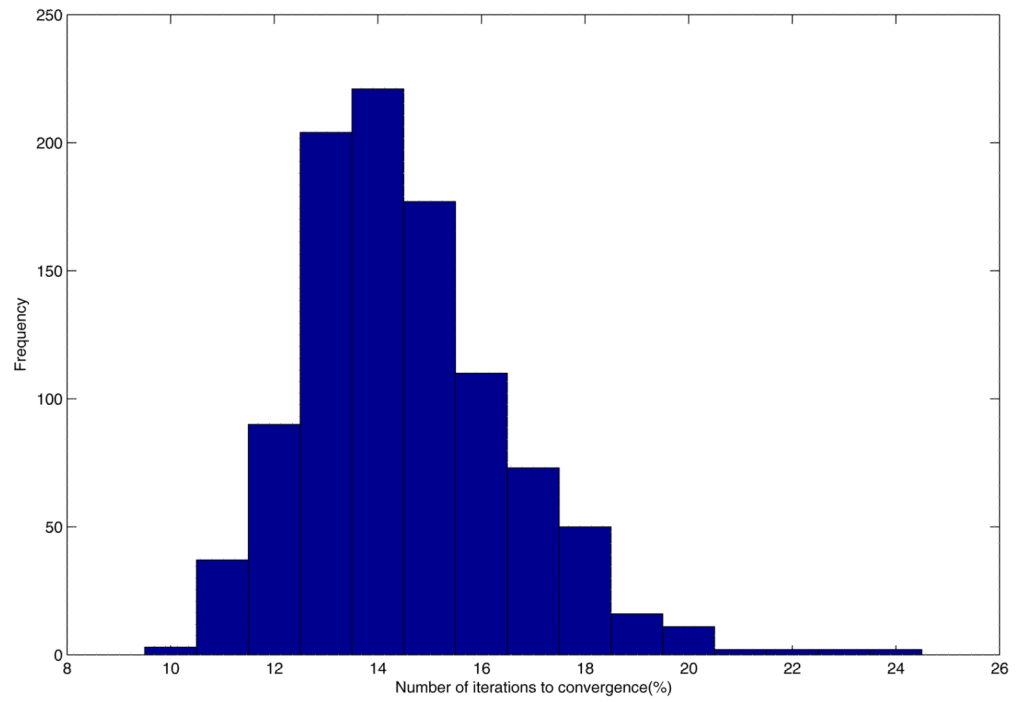


Fig 4.
Distribution of the number of steps to convergence

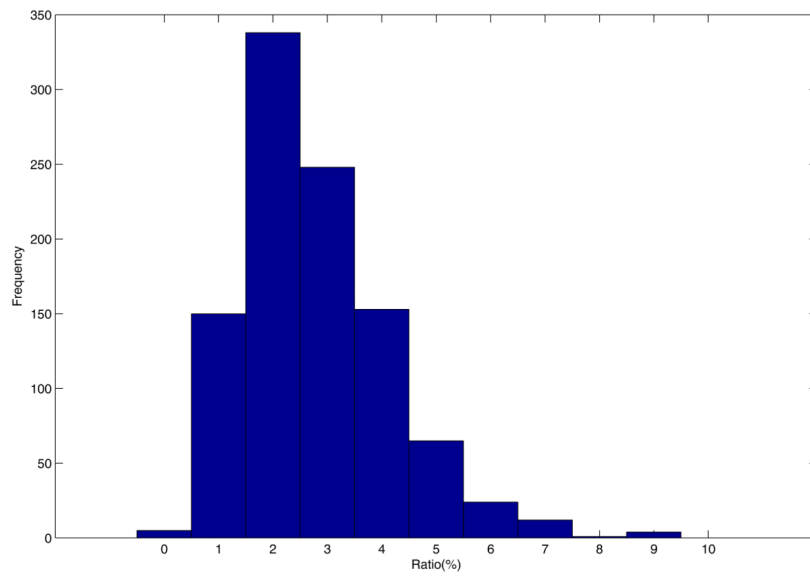


Fig 5. Distribution of distance to limit after 1-step as a proportion of distance to limit from initial-step

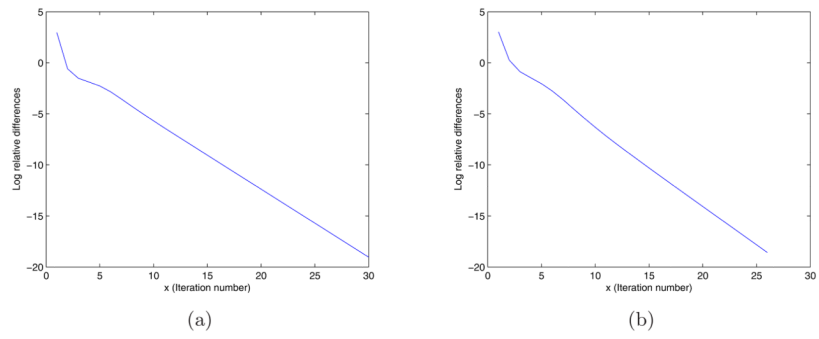


Fig 6. Relative differences at each iteration on the log scale for 5-by-5 dimensional example (a) starting with column polishing (b) starting with row polishing

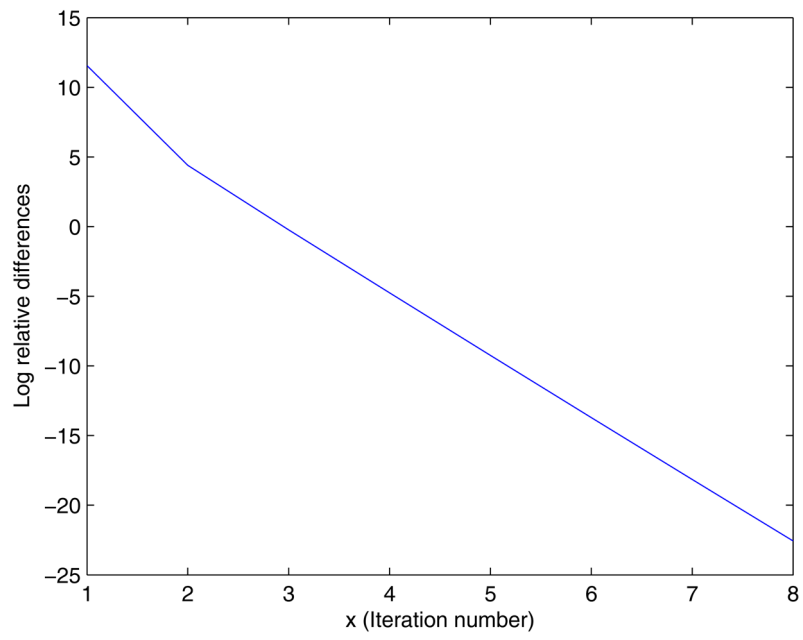


Fig 7.
Relative differences at each iteration on the log scale - 20426-by-63 dimensional example

Table 1

Distribution of the occurrence of sign changes

Iteration no.	relative frequency	relative cumulative frequency
1	94.97%	94.97%
2	3.15%	98.12%
3	1.03%	99.15%
4	0.40%	99.55%
5	0.20%	99.75%
6	0.12%	99.87%
7	0.05%	99.92%
8	0.03%	99.95%
9	0.02%	99.97%
10 and above	0.03%	100.00%
Total	100%	—

Table 2Rapidly of convergence and one-step analysis for various p and n combinations

p	3	4	5	10
n	33	25	20	10
mean(count)	34.0870	22.3500	18.0720	14.5790
std(count)	5.3870	3.3162	2.5477	2.1099
mean(ratio)	2.7026	2.5812	2.6237	2.7207
std(ratio)	1.8483	1.5455	1.4768	1.2699