ニューラルネットワークについて

前川眞一

19980328, 051102,0320, 20151013,20210624, 20210719,20dnc,21,27dnc,2021. 12. 22

1 教師信号ありの順伝搬型ニューラルネットワーク

Q 個の説明変数 $\boldsymbol{x}$ を第 1 層に、m 個の基準変数 $\boldsymbol{y}$ を予測する最終層を第 n 層とするネットワークを考える。各層のニューロンの数は $n_k, k = 1, 2, \ldots, n$ とする。（ただし、後述する常に 1 を出力するニューロンは数えない。）従って、$n_1 = Q, n_n = m$ である。また、個体 a の第 k 層の第 j 番目のニューロンへの入力を i^k_{aj}, 同じく出力を o^k_{aj} と書く。なお、本節では、上付きの添え字 $k, k-1, k+1, n$ 等はすべて添え字であり冪乗を示すものではないことに注意されたい。各個体を表す添え字は $a, a = 1, 2, \ldots, N$ である。各ニューロンはそれぞれ上位のニューロンに向かって同一の出力を発する。また、各ニューロンは下位のニューロンの発する出力を重み付けして加えた形でそのニューロンへの入力を形成する。すなわち、個体 a の第 $k-1$ 層の第 i 番目のニューロンの出力 o^{k-1}_{ai} から個体 a の第 k 層の第 j 番目のニューロンへの入力 i^k_{aj} を作る際の重みを w^k_{ij} とすれば、

$$i^k_{aj} = \sum_{i=1}^{n_{k-1}} o^{k-1}_{ai} w^k_{ij} + \theta^k_j \tag{1}$$

である。ただし、θ_k はバイアスと呼ばれる定数項（閾値）である。また、第 k 層の入力はシグモイド関数（たとえばロジスティク関数） ψ により、

$$o^k_{aj} = \psi(i^k_{aj}) = \frac{1}{1 + \exp(-i^k_{aj})} \tag{2}$$

という変換を受けた形で出力される。ただし、最終層 o^n_{aj} は y_{aj} を近似するためのものであるため、シグモイド関数による変換を施さない場合もあるが、その場合は、 $o^n_{aj} = i^n_{aj}$ である。この過程を、第 k 層の前後に関して図示したものが Figure 1 である。図中、第 $k-1$ 層のニューロンの数は 3、第 k 層のニューロンの数は 2、第 $k+1$ 層のニューロンの数は 2 であるが、定数項 θ は、常に 1 を出力するニューロンへの重みとして記述してある。また、個体を表す添字 a は省略してある。

これを全ての $i^k_{aj}, o^k_{aj}, a = 1, 2, \ldots, N$ に関してまとめて書くと、$\boldsymbol{O}^k$ を $N \times n_k$ の行列、$\boldsymbol{W}^k$

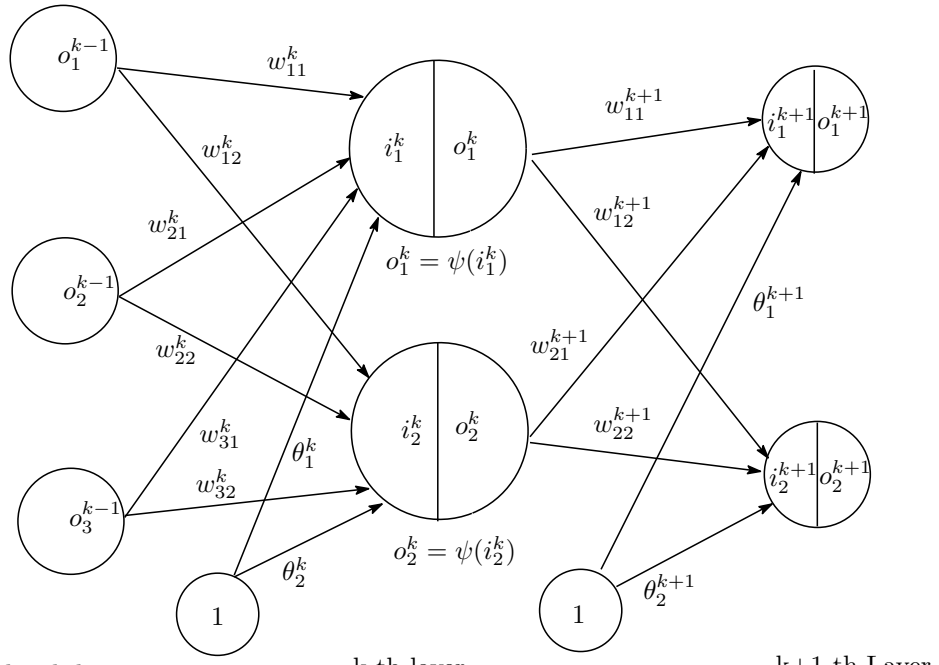

Figure1　ニューラルネットワークモデル （第 k 層前後の拡大図）

を $n_{k-1} \times n_k$ の行列、$\boldsymbol{\theta}^k$ を $1 \times n_k$ のベクトルとして、

$$\begin{aligned}
\boldsymbol{O}^k &= \boldsymbol{\Psi}(\boldsymbol{I}^k) && (3)\\
&= \boldsymbol{\Psi}(\boldsymbol{O}^{k-1}\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) && (4)\\
&= \boldsymbol{\Psi}(\boldsymbol{\Psi}(\boldsymbol{I}^{k-1})\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) && (5)\\
&= \boldsymbol{\Psi}(\boldsymbol{\Psi}(\boldsymbol{O}^{k-2}\boldsymbol{W}^{k-1} + \mathbf{1}\boldsymbol{\theta}^{k-1})\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) && (6)\\
&\quad \vdots \quad \vdots && (7)
\end{aligned}$$

となるが、これは

$$\boldsymbol{O}^k = \boldsymbol{\Psi}(\boldsymbol{O}^{k-1}\boldsymbol{W}^k + \mathbf{1}\boldsymbol{\theta}^k) = \boldsymbol{\Gamma}_k(\boldsymbol{O}^{k-1}) \tag{8}$$

と定義をすれば、最終層の出力が

$$\begin{aligned}
\boldsymbol{O}^n &= \boldsymbol{\Gamma}_n(\boldsymbol{O}^{n-1}) && (9)\\
&= \boldsymbol{\Gamma}_n(\boldsymbol{\Gamma}_{n-1}(\boldsymbol{O}^{n-2})) && (10)\\
&\quad \vdots \quad \vdots && (11)\\
&= \boldsymbol{\Gamma}_n \circ \boldsymbol{\Gamma}_{n-1} \circ \cdots \circ \boldsymbol{\Gamma}_2(\boldsymbol{O}^1) && (12)\\
&= \boldsymbol{\Gamma}(\boldsymbol{O}^1) \;=\; \boldsymbol{\Gamma}(\boldsymbol{X}) && (13)
\end{aligned}$$

と書けることを示している。ただし、

$$\boldsymbol{\Gamma}(.) = \boldsymbol{\Gamma}_n \circ \boldsymbol{\Gamma}_{n-1} \circ \cdots \circ \boldsymbol{\Gamma}_2(.) \tag{14}$$

である。従って、ニューラルネットワークのモデルは、基準変数 $\boldsymbol{Y}$ を説明変数 $\boldsymbol{X}$ の非線形関数 $\widehat{\boldsymbol{Y}} = \boldsymbol{O}^n = \boldsymbol{\Gamma}(\boldsymbol{X})$ で近似するという意味で

$$\boldsymbol{Y} = \boldsymbol{\Gamma}(\boldsymbol{X}) + \boldsymbol{\mathcal{E}} \qquad \text{or} \qquad \widehat{\boldsymbol{Y}} = \boldsymbol{\Gamma}(\boldsymbol{X}) \tag{15}$$

と書くことができる。

なお、以下の記述では、各層の第 n_k 番目のニューロンに加えて、常に 1 を出力するニューロンがもう一つ存在し、その重みを $\boldsymbol{\theta}^k$ としている。すなわち、$\boldsymbol{\theta}^k$ は $\boldsymbol{W}^k$ 行列の最初の行、すなわち、第 0 行として取り扱われるが、これは $\boldsymbol{O}^k$ 行列を $N \times (1 + n_k)$ の行列と定義し、その第 0 列がすべて 1 であるとすることに相当する。

2　最小２乗法によるパラメタの推定に必要となる１次微分

データ $\boldsymbol{Y}$ とモデルの値 $\widehat{\boldsymbol{Y}}$ の差の２乗和

$$\mathrm{RSS} = \sum_{s=1}^{Q}\sum_{a=1}^{N}(y_{as} - \hat{y}_{as})^2 \tag{16}$$

を最小にするようにモデルのパラメタ

$$w_{ij}^k,\ k = 2, 3, \ldots, n,\ i = 0, 1, 2, \ldots, n_{k-1},\ j = 1, 2, \ldots, n_k$$

を求めるためには以下に示す $\hat{y}_{as}$ を各パラメタで偏微分したもの、すなわちヤコビアン行列が必要となる。ただし、w_{0j}^k は前節まで θ_j^k と記していたものであり、

$$\hat{y}_{as} = o_{as}^n \tag{17}$$

である。なお、以下の記述では、o と i に付く個体を表す添字 a は省略してある。

$$\frac{\partial\, o_s^n}{\partial\, w_{ij}^k} = \frac{\partial\, o_s^n}{\partial\, i_j^k} \times \frac{\partial\, i_j^k}{\partial\, w_{ij}^k} \tag{18}$$

$$\frac{\partial\, o_s^n}{\partial\, i_j^k} = \frac{\partial\, o_s^n}{\partial\, o_j^k} \times \frac{\partial\, o_j^k}{\partial\, i_j^k} \tag{19}$$

$$\frac{\partial\, o_s^n}{\partial\, o_j^k} = \sum_{q=1}^{n_{k+1}} \frac{\partial\, o_s^n}{\partial\, i_q^{k+1}} \times \frac{\partial\, i_q^{k+1}}{\partial\, o_j^k} \tag{20}$$

$$\frac{\partial\, o_j^k}{\partial\, i_j^k} = \begin{cases} o_j^k(1 - o_j^k) & \text{最終層も } \Psi \text{ で変換する場合} \\ 1 & \text{最終層は変換しない場合} \end{cases} \tag{21}$$

$$\frac{\partial\, i_q^{k+1}}{\partial\, o_j^k} = w_{jq}^{k+1} \tag{22}$$

$$\frac{\partial\, i_j^k}{\partial\, w_{ij}^k} = \begin{cases} o_i^{k-1} & \text{if } i \geq 1 \quad (w) \\ 1 & \text{if } i = 0 \quad (\theta) \end{cases} \tag{23}$$

なお、Eq.(20) に於いて i_0^{k+1} は定数 1 であり o_j^k に依存しないため、総和記号の和は $q = 1$ からとなる。従って、$\delta_j^{k,s} = \dfrac{\partial\, o_s^n}{\partial\, i_j^k}$ と定義すると、この値は一つ上の第 $k+1$ 層から最終層である

第 n 層までの値を使って、 $2 \leq k < n$ に関して

$$\delta_j^{k,s} = \frac{\partial\, o_s^n}{\partial\, i_j^k} \tag{24}$$

$$= \sum_{q=1}^{n_{k+1}} \frac{\partial\, o_s^n}{\partial\, i_q^{k+1}}\, w_{jq}^{k+1} \quad \times \frac{\partial\, o_j^k}{\partial\, i_j^k} \;=\; \sum_{q=1}^{n_{k+1}} \delta_q^{k+1,s}\, w_{jq}^{k+1} \quad \times \frac{\partial\, o_j^k}{\partial\, i_j^k} \tag{25}$$

$$= \sum_{q=1}^{n_{k+1}} \delta_q^{k+1,s}\, w_{jq}^{k+1} \quad \times \quad o_j^k(1-o_j^k) \tag{26}$$

と再帰的に書くことが出来るが、これを誤差逆伝搬 error back propagation と呼ぶ。なお、最終層である第 n 層は

$$\delta_j^{n,s} = \frac{\partial\, o_s^n}{\partial\, i_j^n} = \left\{ \begin{array}{ll} \left\{ \begin{array}{ll} o_s^n(1-o_s^n) & \text{最終層も } \Psi \text{ で変換する場合} \\ 1 & \text{最終層は変換しない場合} \end{array} \right\} & \text{if} \quad j = s \\ 0 & \text{otherwise} \end{array} \right. \tag{27}$$

である。

まとめると

$$\frac{\partial\, o_s^n}{\partial\, w_{ij}^k} = \frac{\partial\, o_s^n}{\partial\, i_j^k} \times \frac{\partial\, i_j^k}{\partial\, w_{ij}^k} \;=\; \delta_j^{n,s} \quad \times \left\{ \begin{array}{lll} o_i^{k-1} & \text{if } i \geq 1 & (w) \\ 1 & \text{if } i = 0 & (\theta) \end{array} \right. \tag{28}$$

$$= \sum_{q=1}^{n_{k+1}} \delta_q^{k+1,s}\, w_{jq}^{k+1} \quad \times \quad o_j^k(1-o_j^k) \quad \times \left\{ \begin{array}{lll} o_i^{k-1} & \text{if } i \geq 1 & (w) \\ 1 & \text{if } i = 0 & (\theta) \end{array} \right. \tag{29}$$

となる。また、RSS の w_{ij}^k に関する微分 (gradient) は

$$\frac{\partial\, \mathrm{RSS}}{\partial\, w_{ij}^k} = -\,2 \sum_{s=1}^{m} \sum_{a=1}^{N} (y_{as} - o_{as}^n) \times \frac{\partial\, o_{as}^n}{\partial\, w_{ij}^k} \tag{30}$$

となる。

2.1 行列表記

いま、$\boldsymbol{O}^k,\ k = 1, 2, \ldots, n$ を $N \times n_k$ の第 k 層の出力を要素として持つ行列とする。ただし、 $\boldsymbol{O}^1 = \boldsymbol{X}$ であり、最終層の第 s 列を $\boldsymbol{o}_{(s)}^n$ という $N \times 1$ のベクトルで表す。また、$\boldsymbol{W}^k,\quad k = 2, 3, \ldots, n$ を第 $k-1$ 層の出力から第 k 層の入力を

$$\boldsymbol{I}^k = \left[\mathbf{1}_N, \quad \boldsymbol{O}^{k-1} \right] \boldsymbol{W}^k \tag{31}$$

という形で作るための $(n_{k-1}+1) \times n_k$ の重み行列とする。なお、$\boldsymbol{O}^{k-1}$ 行列の最初の列を $\mathbf{1}$ と定義せずに、$N \times n_{k-1}$ の行列 $\boldsymbol{O}^{k-1}$ の第 1 列の左側に $\mathbf{1}$ を付け加えたものに $\boldsymbol{W}^k$ を掛けて $\boldsymbol{I}^k$ を定義していることに注意されたい。したがって、$(n_{k-1}+1) \times n_k$ の行列 $\boldsymbol{W}^k$ を行単位に

$$\boldsymbol{W}^k = (\boldsymbol{w}_0^k, \boldsymbol{w}_1^k, \boldsymbol{w}_2^k, \ldots, \boldsymbol{w}_{n_{k-1}}^k)' \tag{32}$$

と定義すれば、第 0 行 $\boldsymbol{w}_0^{k'}$ は定数項 1 に掛かる係数、第 1 行 $\boldsymbol{w}_1^{k'}$ が $\boldsymbol{O}^{k-1}$ の第 1 列、 $\boldsymbol{o}_{(1)}^{k-1}$、に掛かる係数、第 2 行 $\boldsymbol{w}_2^{k'}$ が $\boldsymbol{o}_{(2)}^{k-1}$ に掛かる係数、という具合になる。また、$\boldsymbol{\Delta}_s^k$ を Eq.(26)

で定義される $\delta_j^{k,s}$、すなわち $\delta_{aj}^{k,s} = \dfrac{\partial\, o_{as}^n}{\partial\, i_{aj}^n}$ をその a, j 要素として持つ $N \times n_k$ の行列と定義すると、Eq.(26) と Eq.(27) は

$$\boldsymbol{\Delta}_s^k = \begin{cases} \left\{ \begin{array}{ll} \boldsymbol{o}_{(s)}^n \odot (\mathbf{1}_N - \boldsymbol{o}_{(s)}^n)\, \boldsymbol{e}_s' & \text{最終層も } \Psi \text{ で変換する場合} \\ \mathbf{1}_N & \text{最終層は変換しない場合} \end{array} \right\} & \text{if} \quad k = n \\ \boldsymbol{O}^k \odot (\mathbf{1}_N \mathbf{1}_{n_k}' - \boldsymbol{O}^k) \odot (\boldsymbol{\Delta}_s^{k+1} \boldsymbol{W}^{k+1\dagger}) & \text{if} \quad k = n-1, \ldots, 2 \\ \quad k = 2, 3, \ldots, n-1, \quad s = 1, 2, \ldots, n_n & \end{cases} \tag{33}$$

と書くことが出来る。ただし、$\odot$ は行列のアダマール積を表し、 $\boldsymbol{e}_s$ はその第 s 行のみが 1 となっている $n_n \times 1$ のベクトルである。また、$\boldsymbol{W}^{k+1\dagger}$ は $\boldsymbol{W}^{k+1}$ 行列の第 0 行目を取り除いた $n_k \times n_{k+1}$ 行列を転置して得られる $n_{k+1} \times n_k$ の行列である。（この $\boldsymbol{W}$ の第 0 行目を取り除くことは Eq.(20) の和が $q = 1$ から始まっていることに対応している。）

したがって、Eq.(28) に示す $\dfrac{\partial\, o_{as}^n}{\partial\, w_{ij}^k}$ の $\boldsymbol{w}_i^k$ の要素 $w_{ij}^k,\ j = 1, 2, \ldots . n_k$ に関する部分をその aj 要素として持つ $N \times n_k$ のヤコビアン行列は

$$\frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \boldsymbol{w}_i^k} = \begin{cases} \left(\boldsymbol{o}_{(i)}^{k-1} \mathbf{1}_{n_k}'\right) \odot \boldsymbol{\Delta}_s^k & \text{if} \quad i \geq 1 \\ \boldsymbol{\Delta}_s^k & \text{if} \quad i = 0 \end{cases} \tag{34}$$

であり、第 k 層の重み行列 $\boldsymbol{W}^k$ のすべての要素に関するヤコビアン行列は Eq.(34) を $i = 0, 1, 2, \ldots, n_k$ と変えて横に並べた $N \times (n_{k-1}+1) n_k$ の行列

$$\frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k'}\right)} = \left[\frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \boldsymbol{w}_0^k},\ \frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \boldsymbol{w}_1^k},\ \frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \boldsymbol{w}_2^k},\ \ldots, \frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \boldsymbol{w}_{n_k}^k} \right] \tag{35}$$

$$= \left[\boldsymbol{\Delta}_s^k,\ \left(\boldsymbol{o}_{(1)}^{k-1} \mathbf{1}_{n_k}'\right) \odot \boldsymbol{\Delta}_s^k,\ \left(\boldsymbol{o}_{(2)}^{k-1} \mathbf{1}_{n_k}'\right) \odot \boldsymbol{\Delta}_s^k,\ \ldots,\ \left(\boldsymbol{o}_{(n_{k-1})}^{k-1} \mathbf{1}_{n_k}'\right) \odot \boldsymbol{\Delta}_s^k \right] \tag{36}$$

となる。ただし、$\mathrm{vec}\left(\boldsymbol{W}^{k'}\right)$ は $\boldsymbol{W}^k$ 行列の各行を横に並べて作られた横ベクトルの転置である。

また、

$$\mathrm{RSS} = \sum_{s=1}^{m} (\boldsymbol{y}_{(s)} - \boldsymbol{o}_{(s)}^n)' (\boldsymbol{y}_{(s)} - \boldsymbol{o}_{(s)}^n) \tag{37}$$

$$= (\mathrm{vec}\,(\boldsymbol{Y}) - \mathrm{vec}\,(\boldsymbol{O}^n))' (\mathrm{vec}\,(\boldsymbol{Y}) - \mathrm{vec}\,(\boldsymbol{O}^n)) \tag{38}$$

であるから、全ての n_n 個の変数に関する $(N \times n_n) \times (n_{k-1}+1) n_k$ のヤコビアン行列は

$$\frac{\partial\, \mathrm{vec}\,(\boldsymbol{O}^n)}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k'}\right)} = \left[\left(\frac{\partial\, \boldsymbol{o}_{(1)}^n}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k'}\right)} \right)',\ \left(\frac{\partial\, \boldsymbol{o}_{(2)}^n}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k'}\right)} \right)',\ \ldots,\ \left(\frac{\partial\, \boldsymbol{o}_{(n_n)}^n}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k'}\right)} \right)' \right]' \tag{39}$$

となり、これらを層の添字 k を変えて横に並べたものが全ての重み行列に関するヤコビアン行列となる。

なお、$\dfrac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k'}\right)}$ 行列の列を commutation matrix $\boldsymbol{K}(n_k, n_{k-1}+1)$ を用いて並べ替えると

$$\frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k}\right)} = \frac{\partial\, \boldsymbol{o}_{(s)}^n}{\partial\, \mathrm{vec}\left(\boldsymbol{W}^{k'}\right)}\, \boldsymbol{K}(n_k, n_{k-1}+1) \tag{40}$$

が得られる。ただし、$\boldsymbol{e}_{a,b}$ をその第 b 要素の値のみが 1 でそれ以外の要素は 0 である a 次元の縦ベクトルとして

$$\boldsymbol{K}(r,c) = \sum_{i=1}^{r}\sum_{j=1}^{c}(\boldsymbol{e}_{r,i}\boldsymbol{e}'_{c,j}) \otimes (\boldsymbol{e}_{c,j}\boldsymbol{e}'_{r,i}) \tag{41}$$

と書ける。(commutation matrix に関しては Harville, 1997, や wikipedia を参照のこと。) したがって、$\dfrac{\partial \operatorname{vec}(\boldsymbol{O}^n)}{\partial \operatorname{vec}\left(\boldsymbol{W}^k\right)}$ は上記の $\dfrac{\partial\, \boldsymbol{o}^n_{(s)}}{\partial \operatorname{vec}\left(\boldsymbol{W}^k\right)}$ を縦に並べて得ることが出来る。

最後に、RSS の $\boldsymbol{W}^k$ に関する微分 (gradient) は以下の $(n_{k-1}+1)n_k \times 1$ のベクトル

$$\frac{\partial \operatorname{RSS}}{\partial \operatorname{vec}\left(\boldsymbol{W}^{k'}\right)} = -\,2\,\left[\frac{\partial \operatorname{vec}(\boldsymbol{O}^n)}{\partial \operatorname{vec}\left(\boldsymbol{W}^{k'}\right)}\right]' (\operatorname{vec}(\boldsymbol{Y}) - \operatorname{vec}(\boldsymbol{O}^n)) \tag{42}$$

となる。また、

$$\frac{\partial \operatorname{RSS}}{\partial \operatorname{vec}\left(\boldsymbol{W}^{k}\right)} = \boldsymbol{K}(n_k, n_{k-1}+1)\frac{\partial \operatorname{RSS}}{\partial \operatorname{vec}\left(\boldsymbol{W}^{k'}\right)} \tag{43}$$

である。

References

Mayekawa, Shin-ichi. (1995) An efficient algorithm for feed-forward neural network regression analysis. *Behaviormetrika*. Vol.22, No. 1, 67-90

Commutation matrix. (n.d.). In *Wikipedia* Retrieved from `https://en.wikipedia.org/w/index.php?title=Commutation_matrix&oldid=1027627556`

Harville,D.A. (1997) *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York, Inc.