THE STATISTICAL PROPERTIES OF RANDOM BITSTREAMS AND THE SAMPLING DISTRIBUTION OF COSINE SIMILARITY

GRAHAM L. GILLER

ABSTRACT. We summarize the statistical properties of statistics computed from independent random bitstreams including the commonly discussed *support* and *cosine similarity*. We derive the moments of the asymptotically normal approximation to the sampling distribution of the cosine similarity of independent random bitstreams and compare those computed moments to those measured by Monte-Carlo simulation. We find agreement for bitstreams of *internet scale* in length (i.e. of order 10 000 bits), and also much smaller (10 bits), and demonstrate that the expected value of the cosine similarity of *independent* bitstreams may be very significantly distant from zero. To compensate for this bias we propose a new statistic, the *Support Adjusted Cosine Similarity* or SACS.

1. DEFINITION OF RANDOM BITSTREAMS

A random bitstream (N, p) is a vector $\boldsymbol{B}$ in $\mathbb{B}^N$ where each component of the vector is independently drawn[1] from $\mathrm{Bernoulli}(p)$. i.e.

$$b_i = \boldsymbol{B} \cdot \boldsymbol{e}_i \sim \mathrm{Bernoulli}(p) \; \forall \; i \in [1, N] \tag{1}$$

We will represent this as the notation $\boldsymbol{B} \sim \mathrm{Bitstream}(N, p)$. Two bitstreams of equal length, $\boldsymbol{A} \sim \mathrm{Bitstream}(N, p)$ and $\boldsymbol{B} \sim \mathrm{Bitstream}(N, q)$, are *independent* if $E(a_i b_j) = E(a_i)E(b_j) \; \forall \; i, j \in [1, N]$. From this it follows that

$$\boldsymbol{A} \wedge \boldsymbol{B} \sim \mathrm{Bitstream}(N, pq) \tag{2}$$

where $\boldsymbol{A} \wedge \boldsymbol{B}$ represents the Boolean *and* operation performed pairwise on bitstream elements and is identically equal to the vector of the pairwise products of the bitstream elements. i.e.

$$(\boldsymbol{A} \wedge \boldsymbol{B}) \cdot \boldsymbol{e}_i = a_i b_i \tag{3}$$

Trivially, $\boldsymbol{A} \wedge \boldsymbol{A} = \boldsymbol{A}$ for any bitstream, $\boldsymbol{A}$.

2. THE SUPPORT AND INNER PRODUCT OF BITSTREAMS

2.1. **Definition.** In market basket and collaborative filtering analysis it is common to introduce a term called the "support"[2][1] of the bitstream which represents the

Date: February 28, 2013. *Giller Investments Research Note*: 20121024/1.

[1] In the following we use the notation $x \sim \mathrm{D}$ to mean "x is distributed as D."

[2] The usage has some commonalities with, but is not identical to, the usage of the term in measure theory.

count of the bits that are non-zero divided by the length of the bitstream, N. We write

$$\text{supp } \boldsymbol{B} = \frac{1}{N}\sum_{i=1}^{N} b_i = \bar{b}. \tag{4}$$

2.2. **Distribution.** From Equation 1 we see that

$$\boldsymbol{B} \sim \text{Bitstream}(N, p) \Rightarrow N \text{ supp } \boldsymbol{B} \sim \text{Binomial}(N, p). \tag{5}$$

As a scaled Binomial distribution, we immediately have

$$\begin{aligned} E(\text{supp } \boldsymbol{B}) &= p \\ \text{and } \text{Var}\,(\text{supp } \boldsymbol{B}) &= \frac{p(1-p)}{N} \end{aligned} \tag{6}$$

and it immediately follows that the distribution of $\text{supp } \boldsymbol{B}$ has a well understood limiting Normal distribution defined solely by these moments.

2.3. **Relationship to the Inner Product.** From Equation 4, we see that the inner product of two bitstreams is proportional to the support of their logical product i.e.

$$\boldsymbol{A} \cdot \boldsymbol{B} = N \text{ supp } \boldsymbol{A} \wedge \boldsymbol{B} \sim \text{Binomial}(N, pq) \tag{7}$$

where $\boldsymbol{A} \sim \text{Bitstream}(N, p)$ and $\boldsymbol{B} \sim \text{Bitstream}(N, q)$.

2.4. **Covariance of the Inner Product with its Arguments.** Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be independent bitstreams of equal length as defined above. Then the inner products $\boldsymbol{A} \cdot \boldsymbol{B}$ and $\boldsymbol{A} \cdot \boldsymbol{A}$ are both Binomial random variables but they are not independent. Suppose for some $\boldsymbol{A}$, N supp $\boldsymbol{A} = n \sim$ Binomial(N, p) i.e. there are n non-zero bits where n is a Binomial variate. The elements of $\boldsymbol{B}$ at the unsupported bits are irrelevant as $x \wedge 0 = 0 \; \forall \; x \in \mathbb{B}$ and so $\boldsymbol{A} \cdot \boldsymbol{B} \leq n$.

We may reorder the *arbitrary* indices of *both* vectors so that $\boldsymbol{A} = (\mathbf{1}_n, \mathbf{0}_{N-n})$ and $\boldsymbol{B} = (\boldsymbol{B}_n, \boldsymbol{B}_{N-n})$ where the subscripted vectors represent the partition of a vector in the first n bits, $\boldsymbol{B}_n$, and the remaining $N - n$ bits, $\boldsymbol{B}_{N-n}$ and $\mathbf{1}$ and $\mathbf{0}$ are the unit and zero vectors respectively. Clearly, $\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{B}_n \cdot \mathbf{1}_n = n$ supp $\boldsymbol{B}_n$. Since the full length vectors are independent then any similar partition of them must also be independent and so n supp $\boldsymbol{B}_n \sim$ Binomial(n, q) and *independent* of $\boldsymbol{A}$ except for the dependence on n. Therefore

$$\begin{aligned} \boldsymbol{A} \cdot \boldsymbol{B} \mid \boldsymbol{A} \cdot \boldsymbol{A} &\sim \text{Binomial}(\boldsymbol{A} \cdot \boldsymbol{A}, q) \\ \text{where } \boldsymbol{A} \cdot \boldsymbol{A} &\sim \text{Binomial}(N, p). \end{aligned} \tag{8}$$

Thus $\boldsymbol{A} \cdot \boldsymbol{B}$ is seen to be drawn from a *nested* Binomial distribution with parameters (N, p, q).

The covariance of $\boldsymbol{A} \cdot \boldsymbol{B}$ and $\boldsymbol{A} \cdot \boldsymbol{A}$ is given by

$$\text{Cov}(\boldsymbol{A} \cdot \boldsymbol{B}, \boldsymbol{A} \cdot \boldsymbol{A}) = E(\boldsymbol{A} \cdot \boldsymbol{B} \times \boldsymbol{A} \cdot \boldsymbol{A}) - E(\boldsymbol{A} \cdot \boldsymbol{B})E(\boldsymbol{A} \cdot \boldsymbol{A}). \tag{9}$$

The latter two terms are known to be Npq and Np respectively. The cross term may be evaluated as

(10)

$$E(\boldsymbol{A}\cdot\boldsymbol{B}\times\boldsymbol{A}\cdot\boldsymbol{A}) = \sum_{\boldsymbol{A}\cdot\boldsymbol{A}=0}^{N} \boldsymbol{A}\cdot\boldsymbol{A}\ \Pr(\boldsymbol{A}\cdot\boldsymbol{A}|N,p) \sum_{\boldsymbol{A}\cdot\boldsymbol{B}=0}^{\boldsymbol{A}\cdot\boldsymbol{A}} \boldsymbol{A}\cdot\boldsymbol{B}\ \Pr(\boldsymbol{A}\cdot\boldsymbol{B}|\boldsymbol{A}\cdot\boldsymbol{A},q)$$

and (remarkably) the entire expression reduces[3] to

$$\operatorname{Cov}(\boldsymbol{A}\cdot\boldsymbol{B},\boldsymbol{A}\cdot\boldsymbol{A}) = Np(1-p)q. \tag{11}$$

Similarly

$$\operatorname{Cov}(\boldsymbol{A}\cdot\boldsymbol{B},\boldsymbol{B}\cdot\boldsymbol{B}) = Npq(1-q). \tag{12}$$

3. The Cosine Similarity of two Bitstreams

3.1. **Definition.** We define the *cosine similiarity* of two bitstreams of equal length as

$$\cos(\boldsymbol{A},\boldsymbol{B}) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\sqrt{\boldsymbol{A}\cdot\boldsymbol{A}\ \boldsymbol{B}\cdot\boldsymbol{B}}}, \tag{13}$$

following the usual definition for an inner product space. Some properties of this metric are discussed in [3]. Unlike the cosine in Euclidean inner product space, this statistic is bounded below at 0.

3.2. **Relationship to the Support.** The properties of the support then allow us to write

$$\cos(\boldsymbol{A},\boldsymbol{B}) = \frac{\operatorname{supp}\boldsymbol{A}\wedge\boldsymbol{B}}{\sqrt{\operatorname{supp}\boldsymbol{A}\operatorname{supp}\boldsymbol{B}}}. \tag{14}$$

3.3. **A Tighter Upper Bound.** The authors of [3] point out that it also follows from the definition of support that

$$\cos(\boldsymbol{A},\boldsymbol{B}) \le \sqrt{\frac{\operatorname{supp}\boldsymbol{A}}{\operatorname{supp}\boldsymbol{B}}}, \tag{15}$$

when $\operatorname{supp}\boldsymbol{A} \le \operatorname{supp}\boldsymbol{B}$.

3.4. **A Normal Approximation to the Sampling Distribution of the Cosine Similarity.** We will approximate the sampling distribution of the cosine similarity of equal length independent bitstreams by using the well-known "Delta Method"[2] of taking the expectations of the first order Taylor expansion of a function of random variables about the means of the underlying variables. i.e.

$$Ef(\boldsymbol{r}) \simeq f(E\boldsymbol{r}) + (\boldsymbol{\nabla}f)\cdot\ E\delta\boldsymbol{r} \tag{16}$$

$$\text{and } \operatorname{Var} f \simeq (\boldsymbol{\nabla}f)\cdot\operatorname{Var}\boldsymbol{r}\cdot\boldsymbol{\nabla}f \tag{17}$$

[3]This was simplified using *Mathematica* v6. as the expression
`FullSimplify[Sum[AA Pr[AA,N,p] Sum[AB Pr[AB,AA,q],{AB,0,AA}],{AA,0,N}]]`
where `Pr[n_,N_,p_]:=PDF[BinomialDistribution[N,p],n]`.

Consider

$$f(\boldsymbol{r}) = \frac{x}{\sqrt{yz}}, \tag{18}$$

where $\boldsymbol{r} = (x, y, z)'$ and x, y, and z are correlated asymptotically normal random variables with a known mean vector and covariance matrix. In this case we have

$$\frac{\partial f}{\partial x} = \frac{f}{x},\ \frac{\partial f}{\partial y} = -\frac{f}{2y} \text{ and } \frac{\partial f}{\partial z} = -\frac{f}{2z} \tag{19}$$

and so

$$E\cos(\boldsymbol{A}, \boldsymbol{B}) \simeq \sqrt{pq} \tag{20}$$

$$\operatorname{Var}\cos(\boldsymbol{A}, \boldsymbol{B}) \simeq \frac{4 - 3p - 3q + 2pq}{4N}, \tag{21}$$

where we have substituted in the previously developed values of the inner product expecations, variances and covariances using $x = \boldsymbol{A}\cdot\boldsymbol{B}$, $y = \boldsymbol{A}\cdot\boldsymbol{A}$ and $z = \boldsymbol{B}\cdot\boldsymbol{B}$. Therefore, we model $\cos(\boldsymbol{A}, \boldsymbol{B})$ as asymptotically distributed as $\operatorname{Normal}(\mu, \sigma^2)$ where μ is given by Equation 20 and σ^2 is given by Equation 21.

The expected value of the cosine similarity for fully independent bitstreams may be quite large, for example if $p = q = 0.5$ then it is 0.5 but for internet scale bitstreams of large lengths, e.g. $N = 10\,000$, the sampling error about this number is quite small in comparison (0.006 in this case). Note that as $p, q \to 1$ then $\sigma^2 \to 0$ whereas $p, q \to 0$ then $\sigma^2 \to 1/N$.

3.5. **Numerical Simulations.** To verify the formulae of Section 3.4 we conducted a small number of Monte-Carlo experiments in *Mathematica*. We generate a random bitstream with the function

$$\texttt{RandomBitstream[N\_, p\_]} := \texttt{Table[If[Random[]} \leq \texttt{p, 1, 0], \{N\}]} \tag{22}$$

and generate the of cosine similarity of two independent bitstreams with

$$\begin{aligned} \texttt{Cosine[N\_, p\_, q\_]} := \texttt{Module[\{A, B\},A} &= \texttt{RandomBitstream[N, p]}; \\ \texttt{B} &= \texttt{RandomBitstream[N, q]}; \\ \texttt{(A.B)/Sqrt[(A.A)(B.B)]]}. & \end{aligned} \tag{23}$$

One thousand trials of the cosine similarity were then generated for bitstreams of length $10\,000$, 100 and 10 bits and various combinations of p and q. The sample mean and variance for these trials were then compared with the predicted values. This results are show in Table 1 on page 5. Without performing any precise statistical testing, it seems fair to conclude that the formulae established are reasonable estimates of the first two moments of the sampling distribution of the cosine similarity of independent random bitmaps.

3.6. **Support Adjusted Cosine Similarity.** To compensate for the bias in the cosine similarity as discussed in Section 3.4, we propose adjusting the formula of Equation 13 by subtracting the asymptotic expected value. i.e.

$$\operatorname{sacos}(\boldsymbol{A}, \boldsymbol{B}) = \cos(\boldsymbol{A}, \boldsymbol{B}) - \sqrt{\operatorname{supp} \boldsymbol{A} \operatorname{supp} \boldsymbol{B}} \tag{24}$$

$$= \frac{\operatorname{supp} \boldsymbol{A} \wedge \boldsymbol{B} - \operatorname{supp} \boldsymbol{A} \operatorname{supp} \boldsymbol{B}}{\sqrt{\operatorname{supp} \boldsymbol{A} \operatorname{supp} \boldsymbol{B}}} \tag{25}$$

where Equation 14 has been substituted into Equation 24 to give Equation 25. We note there is a parallel between the structure of this equation and that of the Pearson correlation coefficient.

REFERENCES

[1] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. *Proceedings of SIGMOD*, pages 265–276, 1997.

[2] G.W. Oehlert. A note on the delta method. *The American Statistician*, 46:27–29, 1992.

[3] Shiwei Zhu, Junjie Wu, Hui Xiong, and Guoping Xia. Scaling up top-k cosine similarity search. *Data & Knowledge Engineering*, 70:60–83, 2011.

TABLES

N	p	q	Exp. Mean	Obs. Mean	Exp. Var.	Obs. Var.
10 000	0.5	0.5	0.5	0.500091	0.000038	0.000039
10 000	0.5	0.25	0.353553	0.353659	0.000050	0.000049
10 000	0.5	0.125	0.25	0.249839	0.000056	0.000053
10 000	0.5	0.0625	0.176777	0.176880	0.000059	0.000058
100	0.5	0.5	0.5	0.499110	0.00375	0.003740
100	0.5	0.25	0.353553	0.347558	0.005	0.005328
100	0.5	0.125	0.25	0.248546	0.005625	0.005876
100	0.5	0.0625	0.176777	0.175621	0.005938	0.006224
10	0.5	0.5	0.5	0.485566	0.0375	0.041569
10	0.5	0.25	0.353553	0.330916	0.05	0.049284
10	0.5	0.125	0.25	0.287991	0.05625	0.055474
10	0.5	0.0625	0.176777	0.254209	0.059375	0.056185

TABLE 1. Results of numerical simulations of the cosine similarity between independent random bitmaps of length 10 000, 100 and 10 bits. For each parameter set, 1 000 trials were generated.

GILLER INVESTMENTS (NEW JERSEY), LLC, 121 RED HILL ROAD, HOLMDEL NJ 07733
E-mail address: graham@gillerinvestments.com
URL: http://blog.gillerinvestments.com