A model for pricing data bundles based on minimax risks for estimation of a location parameter

Mokshay Madiman and Andrew Barron
Department of Statistics, Yale University,
24 Hillhouse Avenue, New Haven, CT 06511, USA
Email: mokshay.madiman@yale.edu, andrew.barron@yale.edu

Abram Kagan and Tinghui Yu
Department of Mathematics, University of Maryland,
College Park, MD 20742, USA
Email: amk@math.umd.edu, yuth@math.umd.edu

***Abstract*—Consider a situation involving many sources of finite-length data, with buyers potentially interested in purchasing data from any bundle (subset) of the sources. A principled way is presented to assign a price to each source, when the value of the data is measured in terms of how much information about an underlying location parameter can be extracted from it. Apart from the operational relevance to data pricing, these results also have relevance to sensor network theory.**

I. Introduction

The collection, storage, sale, communication, interpretation, and use of massive amounts of data is a distinguishing feature of today's world. This gives rise to many new problems, that relate to the modeling and optimization of performance in dealing with all of these data-related tasks. Indeed, whole new fields of scientific investigation have developed under the impetus of the dominance of data in modern culture– including communication and information theory, machine learning, and so on. Now, after several decades of progress on the fundamental conceptual issues of the mathematical modeling of data, increasingly attention is also being paid to the "soft" aspects of data. For example, in recent years, there have been efforts by economists, computer scientists and engineers to develop better mechanisms for the numerous kinds of transactions involving data that are possible (and necessary). In this note, we discuss a specific contribution to this line of work; in particular, we propose and analyze a model for the pricing of data when the value of the data derives from its ability to shed light on an underlying parameter.

First we discuss the rationale for considering this kind of model. The value of data may derive from many different considerations (for instance, their confidentiality) that has nothing to do with traditional statistical considerations. However, for many applications, the value of the data derives from its ability to shed light on some distributional characteristics underlying the data. For instance, in the context of democratic elections, the purchase of data (either by political campaigns or by newspapers and media outlets) from "pollsters" in order to gauge voter sentiment on various issues is now common in many countries. Clearly, what is of interest here is the underlying distribution of voter sentiment, and it would be natural to model the data as a finite sample drawn from this distribution. Furthermore, a typical poll would only examine one particular set of issues or explore one particular state or demographic; so in order to develop a more complete picture, the customer may wish to purchase data from several different polls. An important question for the data vendor that arises is how to price the data from different polls.

Any rational pricing mechanism for the data would take into account the value of the data for the customer (and therefore how much they would potentially be willing to pay for it). We will consider a simplified model for the quantification of value, and for exploring the pricing of data. Our model rests on the assumption that all distributional characteristics of the data are known, except for a location parameter, and this is what is of interest to the customer. In the polling example, one may think of this in terms of the common (albeit possibly naive) mapping of the political spectrum as a one-dimensional range spanning "right" and "left". The data from each poll may be thought of reflecting some underlying "leftistness" or "rightistness" parameter of the population at large. Having access to several polls relating to different issues can improve the customer's ability to estimate this parameter, and hence what the customer is willing to pay. The data vendor wishes to assign prices for the sale of data from each source, keeping in mind the value to the customer of arbitrary bundles of data. For a model of the type outlined above, we discuss a pricing mechanism and its properties.

This note is organized as follows. Section II describes our main results, which focus on developing a rational pricing model. Section III develops some preliminary facts about the Pitman estimator that are used to strong effect in our proofs. Sections IV–VI contain the proofs, which are rather elementary. Section VII discusses various related issues, including connections to sensor network theory.

II. Main Results

First we describe a model for quantifying the value of data, based on which we build a pricing mechanism.

Let $[N]$ denote the index set $\{1, 2, \ldots, N\}$. A precise description of the components of the model is as follows:

1) There are N sources, indexed by the set $[N]$, and these sources are independent of each other (i.e., the data streams produced by the sources are independent of each other).
2) For each $i \in [N]$, source i produces the data stream $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,n_i})$ of length n_i from some known

joint distribution $F_i(\cdot - \theta)$ on $\mathbb{R}^{n_i}$. The distribution F_i is arbitrary, except for the reasonable assumption that its covariance matrix is finite.

3) The parameter θ is some unknown real number, whose determination is the goal of all customers.
4) For any subset $\mathbf{s} \subset [N]$, a customer may choose to purchase the "$\mathbf{s}$-bundle". This refers to the data bundle $\mathbf{Y_s} = (\mathbf{X}_i : i \in \mathbf{s})$.

Any customer uses the data she has purchased, say the $\mathbf{s}$-bundle $\mathbf{Y_s}$, to construct an estimate $\tilde{\theta}_{\mathbf{s}}(\mathbf{Y_s})$ of the parameter θ. One may measure the goodness of an estimator by comparing to the "best possible estimator in the worst case", i.e., by comparing the risk (or mean square error) of the given estimator with the minimax risk. The minimax risk achievable by the $\mathbf{s}$-user is

$$R(\mathbf{s}) = \min_{\text{all estimators } \tilde{\theta}_{\mathbf{s}}} \max_{\theta} \mathbb{E}[(\tilde{\theta}_{\mathbf{s}}(\mathbf{Y_s}) - \theta)^2].$$

For location problems, assuming that there exists an estimator with finite risk, this risk is achievable (see Section III).

The greater the minimax risk associated with estimating θ from a particular set of data, the less valuable that data is to the customer. In particular, if the minimax risk is infinity, the data has no value, while if the minimax risk is 0 and θ can be exactly determined, the data may be considered to be infinitely valuable. It is thus plausible to consider the intrinsic value of the data bundle $\mathbf{Y_s}$ as being quantified by $\frac{1}{R(\mathbf{s})}$ (including a scale factor makes no difference to what follows).

The goal for the seller is to set a price for the data corresponding to each source. Let us say that the i-th source is assigned a price of P_i, so that the price associated with the data bundle $\mathbf{Y_s}$ is

$$P_{\mathbf{s}} = \sum_{i \in \mathbf{s}} P_i. \tag{1}$$

We will explore the possibility of assigning these prices in line with the following principles:

- As an incentive to a customer who decides to buy *all* of the data, the seller would like to price the "grand bundle" consisting of all the data at precisely its intrinsic value:

$$P_{[N]} = \frac{1}{R([N])}. \tag{2}$$

- In setting prices for the data, the seller may justifiably charge at least the intrinsic value of the data. Thus, for the data bundle $\mathbf{Y_s}$, the seller wishes to charge

$$P_{\mathbf{s}} \geq \frac{1}{R(\mathbf{s})}. \tag{3}$$

Our main result is an existence result.

Theorem 1: [Price Allocation] There exists a price allocation $(P_1, \ldots, P_N)$ that satisfies (1), (2) and (3).

This result is based on the following fundamental relationship between the minimax risks of estimating θ using different data bundles.

Let $\mathcal{C}$ be an arbitrary collection of subsets of $[N]$. A *fractional partition* using the collection $\mathcal{C}$ is a set $\beta = \{\beta(\mathbf{s}) : \mathbf{s} \in \mathcal{C}\}$ of non-negative real numbers satisfying

$$\sum_{\mathbf{s} \ni i, \mathbf{s} \in \mathcal{C}} \beta(\mathbf{s}) = 1 \tag{4}$$

for each i in $[N]$. We say a collection is *r-regular* if each index $i \in [N]$ appears in the same number r of sets in $\mathcal{C}$. For an r-regular collection, a canonical choice of fractional partition is $\beta(\mathbf{s}) = \frac{1}{r}$.

Theorem 2: Suppose $F_1, \ldots, F_N$ have finite covariance matrices. Then, for any fractional partition β using any collection $\mathcal{C}$,

$$\frac{1}{R([N])} \geq \sum_{\mathbf{s} \in \mathcal{C}} \frac{\beta(\mathbf{s})}{R(\mathbf{s})}.$$

Finally we show that in the special case when all the samples have the same size, i.e., $n_k = n$ for each $k \in [N]$, the inequality of Theorem 2 behaves gracefully as n becomes large in the sense that it cannot grow large compared with n.

Theorem 3: Suppose $F_1, \ldots, F_N$ have finite covariance matrices, and $n_k = n$ for each $k \in [N]$. Then, for any fractional partition β using any collection $\mathcal{C}$, the gap

$$\frac{1}{R([N])} - \sum_{\mathbf{s} \in \mathcal{C}} \frac{\beta(\mathbf{s})}{R(\mathbf{s})}$$

is $o(n)$ as $n \to \infty$.

Let us conclude this introduction with a critique of the simplistic model described above on economic grounds. A reviewer suggested that the value of information suggested in (2) is the "choke-up price", and the word "price" should be reserved for a representation of marginal utility. Also it is possible in some situations that one may not want to treat data as having no value even if the risk is infinite. At present, we are unable to provide a more nuanced set of results that would take these criticisms into account. However, inspite of these possible criticisms, we believe the results described highlight phenomena of theoretical interest, as well as the breadth of application of information theory and statistics.

III. Preliminaries on the Pitman estimator

The Pitman estimator is the estimator with minimum mean square error among all equivariant estimators of location (see, eg., [1], for definitions and details). Girshick and Savage [2] proved that the Pitman estimator is minimax.

We first recall some basic properties of the Pitman estimator for a location parameter, which is the best equivariant estimator with respect to squared error loss. The Pitman estimator for θ based on observations $Z_1, \ldots, Z_n$ from $F(x_1 - \theta, \ldots, x_n - \theta)$ is defined by

$$t^{(n)} = \bar{Z} - E(\bar{Z}|\mathcal{G}), \tag{5}$$

where

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \quad (6)$$

is the sample mean,

$$\mathcal{G} = \sigma(Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}) \quad (7)$$

is the σ-algebra generated by the residuals, and E stands for the expectation taken at $\theta = 0$. This expression for the Pitman estimator is based on the fact that if $\hat{\theta}$ is any equivariant estimator with finite risk (using the strictly convex squared error loss), then the unique location equivariant estimator with minimum risk is given by

$$t = \hat{\theta} - E[\hat{\theta}|\mathcal{G}], \quad (8)$$

see, e.g., [1][Chapter 3.1]. Since the sample mean $\bar{Y}_{\mathbf{s}}$ is equivariant, the expression (5) follows.

The Pitman estimator is also unbiased, and in fact is almost surely equal to the uniformly minimum variance unbiased estimator when the latter exists, as shown by Romano [3]. Since the Pitman estimator is unbiased, its risk based on squared error loss is simply its variance.

As mentioned above, Girshick and Savage [2] showed the minimaxity of the Pitman estimator, namely, the fact that

$$\mathrm{var}(t^{(n)}) = R_n, \quad (9)$$

where R_n is the minimax risk for estimating θ from $Z_1, \ldots, Z_n$.

IV. Comparing minimax risks

Suppose $(X^{(k)})$ is the sample, of size n_k, from $F_k(\cdot - \theta)$. Let $t^{(\mathbf{s})}$ be the Pitman estimator of θ from the combined sample $(X^{(k)} : k \in \mathbf{s})$ of size $n_{\mathbf{s}} = \sum_{k \in \mathbf{s}} n_k$. In particular, each $t^{(\mathbf{s})}$ is equivariant. Thus, for any collection of $\mathcal{C}$ of subsets, and for any weights $w_{\mathbf{s}}$ with

$$w_{\mathbf{s}} \geq 0, \sum_{\mathbf{s} \in \mathcal{C}} w_{\mathbf{s}} = 1,$$

the estimator

$$\sum_{\mathbf{s} \in \mathcal{C}} w_{\mathbf{s}} t^{(\mathbf{s})}$$

is equivariant. Since $t^{([N])}$ is the MRE estimator from all of the data, we have

$$\mathrm{var}(t^{([N])}) \leq \mathrm{var}\bigg(\sum_{\mathbf{s} \in \mathcal{C}} w_{\mathbf{s}} t^{(\mathbf{s})}\bigg).$$

To bound the variance of this sum, we need the variance drop lemma of Madiman and Barron [4].

Lemma 1: [4] For any $\mathbf{s} \subset [n]$, let $X_{\mathbf{s}}$ denote the . Suppose $\{\psi_{\mathbf{s}} : \mathbf{s} \in \mathcal{C}\}$ is a collection of functions $\psi_{\mathbf{s}} : \mathbb{R}^{|\mathbf{s}|} \to \mathbb{R}$, with $E\psi_{\mathbf{s}}^2(X_{\mathbf{s}}) < \infty$ and $E\psi_{\mathbf{s}}(X_{\mathbf{s}}) = 0$ for each $\mathbf{s} \in \mathcal{C}$. Let

$$U(X_1, \ldots, X_n) = \sum_{\mathbf{s} \in \mathcal{C}} \psi_{\mathbf{s}}(X_{\mathbf{s}}), \quad (10)$$

Then

$$EU^2 \leq \sum_{\mathbf{s} \in \mathcal{C}} \frac{1}{\beta(\mathbf{s})} E\big(\psi_{\mathbf{s}}(X_{\mathbf{s}})\big)^2, \quad (11)$$

for any fractional partition $\{\beta(\mathbf{s}) : \mathbf{s} \in \mathcal{C}\}$.

By Lemma 1,

$$\mathrm{var}\bigg(\sum_{\mathbf{s} \in \mathcal{C}} w_{\mathbf{s}} t^{(\mathbf{s})}\bigg) \leq \sum_{\mathbf{s} \in \mathcal{C}} \frac{w_{\mathbf{s}}^2}{\beta(\mathbf{s})} \mathrm{var}(t^{(\mathbf{s})}).$$

Thus we have

$$\mathrm{var}(t^{([N])}) \leq \sum_{\mathbf{s} \in \mathcal{C}} \frac{w_{\mathbf{s}}^2}{\beta(\mathbf{s})} \mathrm{var}(t^{(\mathbf{s})}).$$

Picking $w_{\mathbf{s}}$ to optimize the upper bound leads to the desired result (Theorem 2), after noting that $\mathrm{var}(t^{(\mathbf{s})})$ is just the minimax risk $R(\mathbf{s})$.

V. The grand buyer need not be unfairly penalized

We now prove Theorem 1; in fact, let us show that Theorem 2 and Theorem 1 are equivalent by linear programming duality. Consider the linear program

$$\begin{array}{rl} \text{Maximize} & \sum_{\mathbf{s} \in \mathcal{C}} \frac{\beta(\mathbf{s})}{R(\mathbf{s})} \\ \text{subject to} & \beta(\mathbf{s}) \geq 0 \text{ for each } \mathbf{s} \subset [n] \\ \text{and} & \sum_{\mathbf{s} \in \mathcal{C}, \mathbf{s} \ni j} \beta(\mathbf{s}) = 1 \text{ for each } j \in [n]. \end{array}$$

The dual problem is easily obtained:

$$\begin{array}{rl} \text{Minimize} & \sum_{j \in [n]} P_j \\ \text{subject to} & \sum_{j \in \mathbf{s}} P_j \geq v(\mathbf{s}) \text{ for each } \mathbf{s} \subset [n]. \end{array}$$

If r^* and p^* denote the primal and dual optimal values, Theorem 2 tells us that $r^* \leq 1/R([n])$, while duality theory tells us that $p^* = r^*$. Thus

$$p^* \leq 1/R([n]),$$

but we already know the reverse inequality is true from the constraints of the dual problem. Hence there exists a point in the rate region such that the sum rate $P_1 + \ldots + P_n = 1/R([n])$.

Alternatively, once Theorem 2 is proved, Theorem 1 simply follows from the Bondareva-Shapley theorem in cooperative game theory (see [5]).

VI. Connection to Fisher information

Port and Stone [6], [7] showed that the Pitman estimator is asymptotically efficient under minimal regularity assumptions. Specifically, if F has a density function f that is differentiable, then

$$\lim_{n \to \infty} n\, \mathrm{var}(t^{(n)}) = \frac{1}{I(X)}. \quad (12)$$

Here $I(X)$ is the Fisher information on θ in $X + \theta$, where X is distributed according to F; note that

$$\begin{aligned} I(X) &= \int_{\mathbb{R}} \bigg[\frac{\partial}{\partial \theta} \log f(x - \theta)\bigg]^2 f(x - \theta) dx \\ &= \int_{\mathbb{R}} \bigg[\frac{\partial}{\partial x} \log f(x)\bigg]^2 f(x) dx \end{aligned}$$

does not depend on θ.

Now let us restrict attention to the setting of Theorem 3, where all the samples have the same size, i.e., $n_k = n$ for each $k \in [N]$. From the preceding paragraph, it is clear that the scaled minimax risk $nR(\mathbf{s})$ converges to $I^{-1}(\mathbf{Y_s})$. (One could write, for clarity, $n\,\mathrm{var}(t_{\mathbf{s}}^{(n)})$ instead of $nR(\mathbf{s})$ to highlight the fact that the sample size is implicit in the minimax risk as well.) Thus by multiplying Theorem 2 by n and letting $n \to \infty$, one obtains for any fractional partition β,

$$I(\mathbf{Y}_{[N]}) \geq \sum_{\mathbf{s} \in \mathcal{C}} \beta(\mathbf{s}) I(\mathbf{Y_s}).$$

In particular, when $N = 2$ and one considers the actual partition $\{\{1\}, \{2\}\}$, one obtains

$$I(\mathbf{Y}_{\{1,2\}}) \geq I(\mathbf{Y}_{\{1\}}) + I(\mathbf{Y}_{\{2\}}). \tag{13}$$

In fact, the inequality (13) is true not just for independent samples $\mathbf{Y}_{\{1\}}, \mathbf{Y}_{\{2\}}$, but for samples with arbitrary dependence. This is Carlen's [8] so-called "superadditivity of Fisher information"; see also Kagan and Landsman [9] for its statistical meaning. It is easy to see that for independent samples, one must have equality in (13).

Thus one must have the gap in Theorem 2 is $o(n)$, since only then would the gap divided by n go to 0 as $n \to \infty$. This proves Theorem 3.

Remark: In the proof of Theorem 2, the only place where independence of sources was used was in order to apply the variance drop inequality, but it is essential there. If not for the independence requirement, an inequality such as that in Theorem 2 would have given an alternative proof of Carlen's superadditivity.

VII. Discussion

We now discuss other implications and interpretations of these results, as well as related work.

Theorem 2 is of intrinsic interest in the theory of Pitman estimation, and a version of it is discussed in [10] in this context. Indeed, Theorem 2 implies that the convergence of the sequence $n\,\mathrm{var}(t^{(n)})$ to its limit is, in fact, monotonic. To see this, assume that all the N sources each produce a sample of size 1 from the same distribution F, and consider the collection of leave-one-out subsets $\mathcal{C}_{N-1} = \{\mathbf{s} : |\mathbf{s}| = N-1\}$ with its canonical fractional partition $\beta(\mathbf{s}) = 1/(N-1)$; Theorem 2 now reads

$$\frac{1}{R([N])} \geq \frac{N}{N-1} \frac{1}{R([N-1])},$$

whence $NR([N]) \leq (N-1)R([N-1])$. Since $R([N])$ is just the variance of the Pitman estimator from N samples of size 1 from F, and hence from a single sample of size N from F, the result follows.

For completeness, let us first mention here some other works that have looked at cooperative games in the context of statistical questions. For instance, Bapat [11] discusses a linear regression game, while Grömping [12] treats the problem of allocating relative importance to predictor variables in linear regression. Madiman [5] reviews the relevance of cooperative games to several settings that arise in information theory, including the theory of robust hypothesis testing.

One may also interpret Theorems 1 and 2 as results about sensor networks. Indeed, one may consider a system of sensors embedded in a space of sources, such that any given sensor is associated to the subset of sources that it sees. In other words, the data bundle $\mathbf{Y_S}$ is what the sensor associated with the set $\mathbf{s}$ of sources sees. In this setting, $R(\mathbf{s})$ is the minimax risk incurred by the $\mathbf{s}$-sensor when trying to estimate θ locally, i.e., based only on data to which it has direct access. Then, for instance, Theorem 2 can be seen as relating the fundamental limits of such local estimation by sensors. Clearly understanding such relationships is an important first step to building a rigorous theory of fundamental limits of sensor networks in more real-life settings where one introduces global objectives as well as local communication and computation constraints. More details will be forthcoming in [13].

References

[1] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed., ser. Springer Texts in Statistics. New York: Springer-Verlag, 1998.

[2] M. A. Girshick and L. J. Savage, "Bayes and minimax estimates for quadratic loss functions," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950.* Berkeley and Los Angeles: University of California Press, 1951, pp. 53–73.

[3] J. Romano, "The Pitman estimator is UMVUE," *Unpublished*, 2001.

[4] M. Madiman and A. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Trans. Inform. Theory*, vol. 53, no. 7, pp. 2317–2329, July 2007.

[5] M. Madiman, "Cores of cooperative games in information theory," *EURASIP J. on Wireless Comm. and Networking*, no. 318704, 2008.

[6] S. C. Port and C. J. Stone, "Fisher information and the Pitman estimator of a location parameter," *Ann. Statist.*, vol. 2, pp. 225–247, 1974.

[7] C. J. Stone, "Asymptotic properties of estimators of a location parameter," *Ann. Statist.*, vol. 2, pp. 1127–1137, 1974.

[8] E. A. Carlen, "Superadditivity of Fisher's information and logarithmic Sobolev inequalities," *J. Funct. Anal.*, vol. 101, no. 1, pp. 194–211, 1991.

[9] A. Kagan and Z. Landsman, "Statistical meaning of Carlen's superadditivity of the Fisher information," *Statist. Probab. Lett.*, vol. 32, no. 2, pp. 175–179, 1997.

[10] A. Kagan, T. Yu, A. R. Barron, and M. Madiman, "Contributions to the theory of Pitman estimators," *Submitted*, 2008.

[11] R. B. Bapat, "On some classes of balanced games," in *Mathematical Programming and Game Theory for Decision Making*, S. K. Neogy, R. B. Bapat, A. K. Das, and T. Parthasarathy, Eds. World Scientific, 2008.

[12] U. Grömping, "Estimators of relative importance in linear regression based on variance decomposition," *Amer. Statist.*, vol. 61, no. 2, pp. 139–147, 2007.

[13] M. Madiman, A. R. Barron, A. M. Kagan, and T. Yu, "Fundamental limits for distributed estimation: the case of a location parameter," *Preprint*, 2009.