Serdica J. Computing **8** (2014), No 2, 97–110

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

SEEKING RELATIONSHIPS IN BIG DATA: A BAYESIAN PERSPECTIVE

Nozer D. Singpurwalla

Abstract. The real purpose of collecting big data is to identify causality in the hope that this will facilitate credible predictivity. But the search for causality can trap one into infinite regress, and thus one takes refuge in seeking associations between variables in data sets. Regrettably, the mere knowledge of associations does not enable predictivity. Associations need to be embedded within the framework of probability calculus to make coherent predictions. This is so because associations are a feature of probability models, and hence they do not exist outside the framework of a model. Measures of association, like correlation, regression, and mutual information merely refute a preconceived model. Estimated measures of associations do not lead to a probability model; a model is the product of pure thought. This paper discusses these and other fundamentals that are germane to seeking associations in particular, and machine learning in general.

0. Preamble: motivation and viewpoint. The impetus for writing this paper is an article in *Science* by Reshef et al. [14], and the strong reaction

ACM Computing Classification System (1998): H.1.2, H.2.4., G.3.

Key words: Association, Correlation, Dependence, Mutual Information, Prediction, Regression, Retrospective Data.

that it has spawned by Kinney and Atawal in *The Proceedings of the National Academy of Sciences* [7]. Given the high visibility that these outlets are endowed with, some discussion clarifying the foundational issues that underlie the said writings seems germane. Both articles pertain to the quantification of the strength of "association" between variables in large data sets. Whereas the term **association** has a precise mathematical meaning in the context of probability theory (cf. [4]), its use here is colloquial and alludes to dependence. The focus of both articles is a heuristic notion, called equitability. The former uses Pearson's *correlation* as its core; the latter Shannon's *mutual information*. Whereas the need for pursuing equitability needs to be made more convincing, at least to this author, a closer reading of these articles underscores the importance of fundamentals when discussing associations. The purpose of this paper is to articulate on the philosophical and the mathematical underpinnings of the notion of dependence; what does it mean to assess it, how best to assess it, and how best to exploit it? The hope is that doing so will make the debates and discussions about seeking relationships in data sets less volatile.

With the above in mind, it is best to present at the outset, the seven bullets given below; these outline the position/viewpoint of this author. This viewpoint is a consequence of a personalistic interpretation of probability in the sense of de Finetti [3] and Savage [15].

1. All probability models are subjectively specified, and they reflect one's disposition to uncertainty, as exemplified by one's attitude to a 2-sided bet.
2. Dependence and association are properties of a probability model, and since probability models are subjectively specified, dependence and associations are judgements.
3. Observed data can only refute a model; they can never endorse it for perpetuity.

   In practice one behaves as if the model at hand is the best one to use, until new evidence falsifies it. To quote George Box [2], "all models are wrong but some are useful". Thus any model is waiting to be falsified, and observed data is the main falsifier (cf. [13]).
4. Assertions 2 and 3 imply that since dependence and association do not exist outside the framework of a model, seeking for associations in the absence of any preconceived notion by merely looking at data is philosophically not tenable.

   This means that when one looks for correlations in data-sets, one has at the back of one's mind a linear relationship. As with regression, either a linear or nonlinear relationship lurks in the mind. This viewpoint car-

ries forward when one asseses mutual information, because to do so one needs to estimate a joint density with the histogram as a starting point, and the preconceived notion underlying a histogram is a bivariate uniform distribution.

5. Like dependence, independence is also a judgment; it implies an absence of learning or the failure of memory. Its mathematical construction entails a hierarchy of assumptions, and these can result in the form of an infinite regress.
6. Unlike correlation and regression which encapsulate specific forms of linear and non-linear relationships, the notion of mutual information is an omnibus measure which can only assert the presence or the absence of an association.
7. The mere act of seeking relationships in data sets is a limited exercise. In actuality what is needed is predictivity. But prediction needs to be probabilistic, and to do so one needs to embed all associations within the framework of probability calculus. This is discussed in Section 4.

   Finally, in the context of the topic of this paper on seeking associations, there is one other caveat; it is also discussed in Section 4. Specifically, when assessing dependence using regression based methods, it matters whether the values of the dependent variables are preselected, or they are retrospectively observed.

1. Dependence: feature of joint distributions. The notions of causality, correlation, information, and regression, play a prominent role in statistics. Their precise definitions are cast in the language of probability. With the advent of big data and machine learning, these notions have gained additional prominence. Seeking patterns within variables, and relationships between variables, has now become a full time occupation for some. Whereas pinpointing causality is the holy grail which drives the collection of big data, many have taken heed of the dictum that "correlation is not causation". Indeed, causation is an elusive notion that has proven to be a challenge, not only to statisticians, but also to philosophers of science. Yet when one speaks of gaining knowledge from big data sets, one has at the back of one's mind the identification of a *genuine cause* (see [19]) which spawns the data. But since the search for a genuine cause can also trigger the problem of an infinite regress, one tends to take refuge in the next best thing, namely, an empirical assessment of measures of dependence, like correlation. Because all measures of dependence are properties of a joint probability model, it behooves one to ask: what is a probability model, and where does it come from? This is the topic of the next section.

2. The genesis of a probability model. At some reference time $\tau \geq 0$, consider an analyst ♀ who assesses his(her) uncertainty about an unknown quantity X, in the light of all the historical information $\mathscr{H}$ that ♀ has at τ, via probability. That is, ♀ needs to specify

$$P^{\tau}_{\text{♀}}(X \geq x; \mathscr{H}).$$

Since the dimension of $\mathscr{H}$ is large, conceptually infinite, ♀ seeks simplification by introducing a quantity θ (whose interpretation is given later), and invoking the law of total probability to write

$$P^{\tau}_{\text{♀}}(X \geq x; \mathscr{H}) = \int_{\theta} P^{\tau}_{\text{♀}}(X \geq x|\theta; \mathscr{H}) P^{\tau}_{\text{♀}}(\theta; \mathscr{H}) d\theta,$$

where $P^{\tau}_{\text{♀}}(\theta; \mathscr{H})$ encapsulates ♀'s uncertainty about θ in the light of $\mathscr{H}$, at time τ. It is called a *prior* for θ. Were ♀ to assume that given θ, the event $(X \geq x)$ is independent of $\mathscr{H}$, then $P^{\tau}_{\text{♀}}(X \geq x|\theta; \mathscr{H}) = P^{\tau}_{\text{♀}}(X \geq x|\theta)$, and now

$$P^{\tau}_{\text{♀}}(X \geq x; \mathscr{H}) = \int_{\theta} P^{\tau}_{\text{♀}}(X \geq x|\theta) P^{\tau}_{\text{♀}}(\theta; \mathscr{H}) d\theta.$$

The quantity $P^{\tau}_{\text{♀}}(X \geq x|\theta)$ is called a *probability model* for the event $(X \geq x)$, and θ is called the *parameter* of the probability model. The quantity $P^{\tau}_{\text{♀}}(X \geq x; \mathscr{H})$ is called the *predictive distribution* of X, and one endeavours to provide predictive distributions that are trustworthy. To do so, one's choice for a probability model and the prior need to be judicious and meaningful. The notion of independence is articulated above. Note that independence has been defined in the framework of probability. More often than not, independence is conditional; in the above case, $(X \geq x)$ is independent of $\mathscr{H}$, conditional on θ (i.e, were θ to be known).

The parameter θ can be a scalar or a vector whose dimension needs to be much smaller than that of $\mathscr{H}$; otherwise the parameter does not serve a useful purpose. This is because the role of the parameter has been to compress the information about $(X \geq x)$ contained in $\mathscr{H}$. Indeed $\mathscr{H}$ can comprise both qualitative and quantitative information, like previously observed data on $(X \geq x)$.

Thus parameters in probability models can be seen as devices which compress the high dimensional $\mathscr{H}$ to a lower dimensional θ. de Finetti referred to θ merely as a Greek symbol; i.e. an abstract entity which need not have an observable reality. Its role is to impart independence between $\mathscr{H}$ and the event $(X \geq x)$, and also as a device which facilitates the prediction of observables, like

X. There are other interpretations of θ, but for now it suffices to say that a statistician's approach to *data compression* is through the introduction of Greek symbols called parameters.

Some of the well known examples of probability models are the exponential, wherein $P^{\tau}_{\text{♀}}(X \geq x|\lambda) = \lambda \exp(-\lambda x)$, for $\lambda > 0$ and $x \geq 0$, and the Weibull, wherein
$P^{\tau}_{\text{♀}}(X \geq x|\lambda,\ \beta) = \exp(-\lambda x^{\beta}), \quad \lambda,\ \beta > 0$ and $x \geq 0$. In the first case $\theta = \lambda$, is a scalar, and in the second case $\theta = (\lambda,\ \beta)$ is a vector. The best known and the most commonly used example of a probability model is the normal (or the Gaussian) wherein for $\theta = (\mu,\ \sigma^2)$, with $-\infty < \mu < +\infty$, and $\sigma > 0$

$$P^{\tau}_{\text{♀}}(X \geq x|\mu, \sigma) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx,$$

for $-\infty < x < +\infty$.

In the context of large data sets interest is focussed on two or more unknown quantities and the relationships between them. For purposes of discussion consider two unknowns, say X and Y, and their predictive distribution $P^{\tau}_{\text{♀}}(X \geq x,\ Y \geq y; \mathscr{H})$. As before, $\mathscr{H}$ is the historical (or background) information about X and Y possessed by ♀ at time $\tau \geq 0$. Here again ♀ may choose to introduce a parameter θ, invoke the law of total probability, assume independence of the event $(X \geq x,\ Y \geq y)$ and $\mathscr{H}$, given θ, and write:

$$(2.0) \qquad P^{\tau}_{\text{♀}}(X \geq x, Y \geq y|\mathscr{H}) = \int_{\theta} P^{\tau}_{\text{♀}}(X \geq x, Y \geq y|\theta) P^{\tau}_{\text{♀}}(\theta; \mathscr{H}) d\theta,$$

The quantity $P^{\tau}_{\text{♀}}(X \geq x, Y \geq y|\theta)$ is the joint (bivariate) probability model for the compound event $(X \geq x, Y \geq y)$. Note that whereas the event $(X \geq x, Y \geq y)$ has been judged independent of $\mathscr{H}$ conditional on θ, nothing has yet been said about the dependence or independence of the events $(X \geq x)$ and $(Y \geq y)$. Clearly, whenever $(X \geq x)$ and $(Y \geq y)$ share a θ (or a sub-set of θ), and θ is unknown, they will be unconditionally (of θ) dependent. This form of dependence is called *dependence by mixture*. However, conditional on θ, the events $(X \geq x)$ and $(Y \geq y)$ could be dependent or independent depending on ♀'s judgment. For example, if ♀ judges the events $(X \geq x)$ and $(Y \geq y)$ conditionally (given θ), independent, then the bivariate probability model is

$$P^{\tau}_{\text{♀}}(X \geq x, Y \geq y|\theta) = P^{\tau}_{\text{♀}}(X \geq x|\theta) P^{\tau}_{\text{♀}}(Y \geq y|\theta),$$

where each term on the right is a univariate probability model. If the above judgment of conditional independence is not tenable, then ♀ is faced with the task

of specifying a probability model for X and Y which encapsulates dependence. An example is the bivariate exponential distribution of Gumbel, wherein for some $\theta = \sigma,\ \sigma \in [0,1]$, and $x, y \geq 0$,

$$P^{\tau}_{\mathop{\scriptstyle\mathrm{Q}}\limits_{\scriptstyle\mathrm{+}}}(X \geq x, Y \geq y|\theta = \sigma) = e^{-(x+y+\sigma x \cdot y)} \tag{2.1}$$

Here the marginals are $P^{\tau}_{\mathop{\scriptstyle\mathrm{Q}}\limits_{\scriptstyle\mathrm{+}}}(X \geq x|\sigma) = e^{(-x)}$, and $P^{\tau}_{\mathop{\scriptstyle\mathrm{Q}}\limits_{\scriptstyle\mathrm{+}}}(Y \geq y|\sigma) = e^{(-y)}$, implying that the dependency parameter σ has no role to play with respect to the marginals. Note that when σ is assumed known, and is zero, then the events $(X \geq x)$ and $(Y \geq y)$ are independent; with σ unspecified, they are dependent.

The best known and the most discussed example of a bivariate probability model is the bivariate normal, with $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Here

$$P^{\tau}_{\mathop{\scriptstyle\mathrm{Q}}\limits_{\scriptstyle\mathrm{+}}}(X \geq x, Y \geq y|\theta) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y|\theta)dxdy,$$

where

$$f(x,y|\theta) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right\}. \tag{2.2}$$

3. Measures of association. As mentioned, dependence is a feature of a joint probability distribution, and for purposes of discussion we will center discussion on the bivariate normal, and the bivariate Gumbel distributions. For the former, its $f(x,y|\theta)$ is given by Equation (2.2). For the latter, an analogue of $f(x,y|\theta)$ is the probability density generated by Equation (2.1).

There are several attractive features of the model of Equation (2.2), two of which are, closure under marginalization and under conditionalization. That is, the marginal of X is also a normal with the parameters μ_1, and σ_1. Or,

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}, \qquad -\infty < x < +\infty,$$

and the conditional of Y, given $X = x$, is also a normal, with the parameters $\mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(x-\mu_1)$ and $\sigma_2^2(1-\rho^2)$.

Indeed there are many families of bivariate distribution each possessing its own version of $f(x,y|\theta)$, so the discussion which follows is generic, and is in terms of $f(x,y|\theta)$.

The most commonly used measure of dependence is the Galton-Pearson coefficient of correlation $\rho(x, y)$. With $f(x, y|\theta)$ specified, $\rho(x, y)$ can always be mathematically obtained. Irrespective of what $f(x, y|\theta)$ is, $|\rho(X, Y)| \leq 1$, and this is its attractive feature. Furthermore, when X and Y are independent, $\rho(X, Y) = 0$. However, $\rho(X, Y) = 0$ does not imply that X and Y are necessarily independent, and this is one of its limitations. An exception is the bivariate normal, for which $\rho(X, Y) = \rho$ of Equation (2.2), and here $\rho = 0$, implies that X and Y are independent. Another limitation of $\rho(X, Y)$ is that it only encapsulates the extent of linear relationship between X and Y. This becomes transparent when one looks at the definition of $\rho(X, Y)$, namely, that $\rho(X, Y) \stackrel{\text{def}}{=} Cov(X, Y)/\sqrt{V(X)V(Y)}$, where $Cov(X, Y) = \mathscr{E}(X \cdot Y) - \mathscr{E}(X) \cdot \mathscr{E}(Y)$, and $\mathscr{E}$ denotes expectation. The variance of X, $V(X) = Cov(X, X)$. These properties boil down to the feature that the best known and commonly used measure of dependence (or association), namely, the correlation is limited in scope. Thus, alternates to correlation have been considered. Some of these are Kendall's Tau [6], Spearman's Rho [18], and the several non-parametric measures of dependence introduced by Lehmann [9], and further articulated by Barlow and Proschan [1]. We do not pursue here these alternatives.

After correlation, the next best known measure of a relationship is regression. Specifically, the *regression* of Y on X is $\mathscr{E}(Y|X = x, \theta)$; similarly, $\mathscr{E}(X|Y = y, \theta)$. The regression of Y on X can take several forms, such as linear wherein $\mathscr{E}(Y|X = x, \theta) = \alpha + \beta x$, with $\theta = (\alpha, \beta)$, quadratic wherein $\mathscr{E}(Y|X = x, \theta) = \alpha + \beta x + \gamma x^2$, with $\theta = (\alpha, \beta, \gamma)$, cubic, and so on. Like the correlation $\rho(X, Y)$, the regression can also be theoretically computed, once $f(x, y|\theta)$ is specified. Thus regression can encapsulate a variety of linear and non-linear relationships, and is one step up the ladder from correlation for describing relationships.

In the case of the bivariate normal distribution, the regression of Y on X takes the linear form

$$\mathscr{E}(Y|X = x, \theta) = \mu_2 + \rho \left(\frac{\sigma_2}{\sigma_1}\right)(x - \mu_1). \tag{3.1}$$

with $V(Y|X = x, \theta) = \sigma_2^2(1 - \rho^2)$. This means that the average value of Y increases (decreases) linearly in x, depending on whether ρ is greater or less than 0. Clearly, in the case of the bivariate normal, the regression provides little added insight about the relationship between X and Y beyond that which is provided by the correlation ρ. The one interesting feature here is that $V(Y|X = x)$ is independent of x. It is this property which makes the bivariate normal distribution attractive in the context of the standard Kalman Filter (cf. [12]).

With the bivariate exponential of Gumbel, the regression of Y on X, $\mathscr{E}(Y|X=x,\theta) = (1+\sigma+\sigma x)/(1+\sigma x)^2$, which for $\sigma > 0$ is a gracefully decreasing function of x, starting from $(1+\sigma)$ and tailing off to 0 as x goes to $+\infty$. Here the correlation $\rho(X,Y) < 0$, and depending on the values of the dependency parameter σ, it ranges from $-.4036$ to 0. Here, the regression provides more insight about the relationship between X and Y, than the correlation. A similar feature is also exhibited by other bivariate distributions like the bivariate exponential of Marshall and Olkin [11] and a second version of Gumbel's bivariate exponential distribution. The specifics about these distributions can be found in Singpurwalla [17, p. 89–93].

In the case of Marshall and Olkin's bivariate exponential distribution, when the correlation $\rho(X,Y) = \dfrac{1}{3}$, $\mathscr{E}(Y|X=x,\cdot) = 1 - \dfrac{3e^{-x}}{4}$, suggesting that the regression is an exponentially increasing function of x, starting from $\dfrac{1}{4}$ and tapering off at 1. In the case of the second version of Gumbel's bivariate exponential distribution, when $\rho(X,Y) = \dfrac{1}{4}$, the regression of Y on X, when $X = x$, is $\dfrac{3}{2} - e^{-x}$. This is an exponentially increasing function starting from $\dfrac{1}{2}$ and tapering off at $\dfrac{3}{2}$. When $\rho(X,Y) = -\dfrac{1}{4}$, the said regression is $\dfrac{1}{2} + e^{-x}$, which is an exponentially decreasing function starting at $\dfrac{3}{2}$, and tapering off at $\dfrac{1}{2}$; see Figure 3.1.

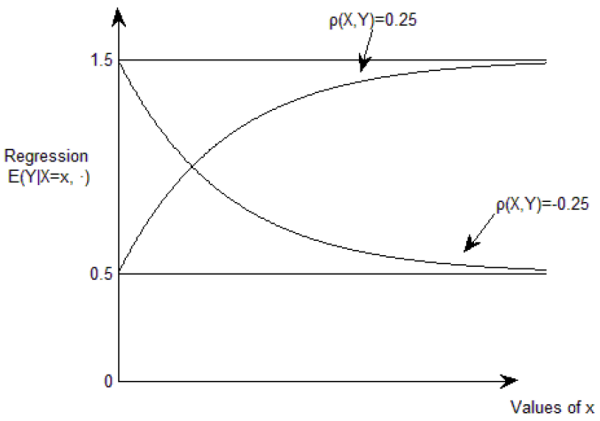

Fig. 3. 1. Regression for Gumbel's Distribution
for Positive and Negative Correlations.

Moving up the ladder for describing relationships, are the distance based measures, of which Shannon's *mutual information* is the best known. A more recent entry in this category is the work of Szekely and Rizzo [20]. Mutual information provides a measure of dependence without specifying its nature. It can be interpreted as the gain in information in going from independence to dependence. Since independence suggests an absence of learning, dependence implies knowledge. Consequently if the notion of information can be quantified, then the difference between information under dependence and information under independence, is the gain in knowledge. The quantification of information is due to the work of Shannon [16]. With $f(x, y|\theta)$ specified, and the marginals $f(x|\theta)$ and $f(y|\theta)$ known, Shannon's measure of information leads to the result that the gain in information (knowledge), say γ, is given by the expression

$$\gamma = \int_{x,y} \{f(x, y|\theta) log f(x, y|\theta) - f(x|\theta) f(y|\theta) log[f(x|\theta) f(y|\theta)]\} dxdy.$$

Thus γ is the amount of information conveyed to an individual who previously supposed X and Y to be independent, by the statement that the joint probability of X and Y is $f(x, y|\theta)$. It may of interest to note that γ provided Kullback and Leibler (1951) a motivation for defining their famous measure of discrepancy or divergence between two distributions.

On its own γ is a satisfactory measure of dependence which can be computed once $f(x, y|\theta)$ is specified. However, under certain circumstances, in particular those pertaining to the form of $f(x, y|\theta)$, it can be shown (cf. [10]), that γ is related to the Pearson correlation via the relationship

$$\rho(X, Y) = \sqrt{1 - e^{-2\gamma}}.$$

In general, irrespective of what $f(x, y|\theta)$ is, the quantity $\sqrt{1 - e^{-2\gamma}}$ lies between 0 and 1, and takes the value 0 when X and Y are independent. It takes the value 1 whenever X can be uniquely determined by Y and vice versa. Linfoot [10] refers to $\sqrt{1 - e^{-2\gamma}}$ as the *informational coefficient of correlation*, and besides the fact that its special case is Pearson's correlation, it has the virtue of invariance. In other words $\sqrt{1 - e^{-2\gamma}}$ does not change if X is replaced by $X' = \varphi_1(X)$ and Y replaced by $\varphi_2(Y)$, for any φ_1 and φ_2. The measure γ is also known as *mutual information*; its invariance was pointed out by Jeffreys [5].

3.1. Association measures provide insight about models. Bullet 7 of Section 0 makes the claim that the mere act of seeking relationships in data sets is of limited value. Limited, because a knowledge of associations can only provide

insight about the nature of a joint distribution that may be entertained. Without further analyses and development, associations on their own do not enable predictivity. Of the three measures of dependence discussed before, correlation is the easiest estimate. Its value indicates the extent to which the two variables in question bear a linear relationship to each other. On its own, correlation does not provide a mechanism for predicting the value of one variable, say Y, knowing the value of X. Mutual information is probably the most difficult measure to compute because it entails the estimation of joint and marginal densities. But having computed mutual information, all we know is that underlying variables are dependent or not. Even if mutual information computed from data supports the hypothesis of dependence, one is unable to predict Y knowing a value of X and vice versa. This suggests that both correlation and mutual information should be viewed as qualitative measures of dependence. Neither can help pinpoint a joint probability model; thus the main purpose served by these measures is to refute (or not) a contemplated model.

Matters become more attractive when it comes to regression. First, like correlation, (but unlike mutual information), regression is easy to estimate. Second, the functional form of regression can provide insight about the joint probability model to entertain. For example, a bivariate normal if the regression is linear, a bivariate Gumbel if it is exponential, and so on. Furthermore, the regression function can serve as a device – albeit naive – for prediction as well. It therefore appears that when seeking relationships in large data sets, it may be more fruitful to pursue the regression function as opposed to correlation or mutual information.

4. Inference and predictivity. For purposes of discussion, we focus attention on the bivariate case with the bivariate normal as the underlying model. In the absence of any observed data on X and Y, predictivity is achieved via Equation (2.0), once the model and the prior are specified. In the bivariate normal case, $\theta = (\mu_1, \mu_2, \rho, \sigma_1, \sigma^2)$, so that the simultaneous prediction of X and Y is:

$$P^{\tau}_{\text{♀}}(X \geq x, Y \geq y|\mathscr{H}) = \int_{\theta} f(x, y|\theta) P^{\tau}_{\text{♀}}(\theta; \mathscr{H}) d\theta, \tag{4.1}$$

where $f(x, y|\theta)$ is prescribed by Equation (2.2). There are two challenges to implementing Equation (4.1). One is a specification of the prior $P^{\tau}_{\text{♀}}(\theta; \mathscr{H})$; the second is computing, which entails integration in five dimensions. Given the conceptual character of this paper, both these "operational" matters are not discussed.

Data on the variables X and Y can arise under two scenarios, each calling for its own approach to predictivity. We label the two scenarios *retrospective*

and *designed*. Under the former scenario, one obtains n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ as the realizations of two random variables X and Y. This is the kind of data that arises in large data sets. Under the designed scenario, one fixes X at say x_i^* and observes the Y corresponding to x_i^* as y_i, $i = 1, \ldots, n$. Here y_i is the realization of a random variable, whereas x_i^* is not. To summarize (x_i, y_i), $i = 1, \ldots, n$, is a realization of a random variable (X, Y), whereas with (x_i^*, y_i), it is only y_i that is the realization of a random variable Y when X is set at x_i^*.

Given the data $\underset{\sim}{d} : [(x_1, y_1), \ldots, (x_n, y_n)]$, the posterior distribution of θ is obtained, via Bayes' Law, as the proportionality relationship

$$P_{\text{♀}}^{\tau}(\theta; \underset{\sim}{d}) \propto P_{\text{♀}}^{\tau}(\theta; \mathscr{H}) \mathscr{L}(\theta; \underset{\sim}{d}),$$

where the likelihood $\mathscr{L}(\theta; \underset{\sim}{d})$, based on Equation (2.2), is:

$$\prod_{i=1}^{n} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_i - \mu_1}{\sigma_1} \right)^2 + \left(\frac{y_i - \mu_2}{\sigma_2} \right)^2 - \frac{2\rho(x_i - \mu_1)(y_i - \mu_2)}{\sigma_1\sigma_2} \right] \right\}.$$

With the above in place, the predictive distribution is:

$$P_{\text{♀}}^{\tau}(X \geq x, Y \geq y; \underset{\sim}{d}, \mathscr{H}) \propto \int_{-\infty}^{x} \int_{-\infty}^{y} \int_{\theta} f(x, y|\theta) P_{\text{♀}}^{\tau}(\theta; \underset{\sim}{d}) d\theta dy dx, \tag{4.2}$$

Continuing in this vein, that is, with (d) at hand, suppose that one wishes to predict Y conditional on observing $X = x_{n+1}$, say. That is, one wishes to assess $P_{\text{♀}}^{\tau}(Y > y | X = x_{n+1}; \underset{\sim}{d}, \mathscr{H})$; this is proportional to

$$\int_{-\infty}^{y} \int_{\theta} f(y|X = x_{n+1}, \theta) P_{\text{♀}}^{\tau}(\theta; \underset{\sim}{d}) d\theta dy,$$

where

$$f(y|X = x_{n+1}, \theta) = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \exp\{-\frac{1}{2}(\frac{y - \mu_2 - \rho\frac{\sigma_2}{\sigma_1}(x_{n+1} - \mu_1)}{\sigma_2\sqrt{(1-\rho^2)}})^2\}.$$

Note that in the above assessment, X has not been observed as x_{n+1}; rather, it is in the subjunctive mood and it means, were X to be observed as x_{n+1}.

4.1. The scenario leading to linear models. This sub-section pertains to predictivity when X is not random and its value is predetermined. Here $\underset{\sim}{d}^* = [(x_1^*, y_1), (x_2^*, y_2), \ldots, (x_n^*, y_n)]$, and one is interested in predictions about Y when X is set at x_{n+1}^*. To do so, one needs a probability model for Y, with $X = x^*$ acting as a parameter of the model. One possibility would be to consider a family of models called *linear models*, and studied under the (incorrectly) labeled term *regression analysis*. The scenario of having d^* at hand arises in the context of designed or (laboratory) controlled experiments.

We start with the question of what it is that motivates the development of linear models and why the term regression analysis? In other words, what is the genesis of linear models? Our answer to this question is suggested by two features. One is Equation (3.1) pertaining to $\mathscr{E}(Y|X = x, \theta)$, the regression of Y on X. The other is the elementary fact that any random variable Y can be written as the sum of its expectation (assuming that it is finite) and an error whose expectation is 0. That is,

$$Y = \mathscr{E}(Y) + \epsilon,$$

where ϵ is the error. Conditioning on $X = x$, we have

$$(Y|X = x) = \mathscr{E}(Y|X = x) + (\epsilon|X = x). \tag{4.3}$$

In the case of the bivariate normal with $\mu_1 = 0$ and $\sigma_1^2 = 1$, $\mathscr{E}(Y|X = x) = \mu_2 + \sigma_2\rho x$, and $V(Y|X = x) = \sigma_2^2\sqrt{1-\rho^2}$. Thus Equation (4.3) becomes, in the bivariate normal case,

$$(Y|X = x) = \mu_2 + \sigma_2\rho x + \epsilon,$$

if ϵ is assumed independent of $X = x$. Also $V(\epsilon) = V(Y|X = x) = \sigma_2^2\sqrt{1-\rho^2}$ when X is pre-selected and fixed at x^*, the above relationship gets written as:

$$Y(x^*) = \alpha + \beta x^* + \epsilon, \tag{4.4}$$

where $\alpha = \mu_\alpha$, $\beta = \sigma_\alpha\rho$, and ϵ has a normal distribution with mean 0, and variance $\sigma^2 = \sigma_2^2\sqrt{1-\rho^2}$.

This in turn means that the probability model for Y with x^* as a parameter is a univariate normal with mean $\alpha + \beta x^*$, and variance σ^2.

The relationship of Equation (4.4) can be generalized to a polynomial in x^*, and also to variables other than X giving us a family of linear models. This could be one way to describe the genesis of linear models and a use of the term regression in their context.

With the above in place, and focussing on the simple linear model of Equation (4.4), the posterior distribution of $\theta = (\alpha, \beta, \sigma^2)$, with $\underset{\sim}{d}^*$ at hand is

$$P_{\text{♀}}^{\tau}(\theta; \underset{\sim}{d}^*) \propto P_{\text{♀}}^{\tau}(\theta; \mathscr{H}) \mathscr{L}(\theta; \underset{\sim}{d}^*),$$

where $P_{\text{♀}}^{\tau}(\theta; \mathscr{H})$ is the prior, and the likelihood

$$\mathscr{L}(\theta; \underset{\sim}{d}^*) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_i - \alpha - \beta x_i^*}{\sigma})^2}.$$

Finally, the predictive distribution of Y with X fixed at x_{n+1}^* is now obtained, at least in principle, as:

$$P_{\text{♀}}^{\tau}(Y \geq y; \underset{\sim}{d}^*, x_{n+1}^*, \mathscr{H}) \propto \int_{-\infty}^{y} \int_{\theta} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_i - \alpha - \beta x_{n+1}^*}{\sigma})^2} P_{\text{♀}}^{\tau}(\theta; \underset{\sim}{d}^*) d\theta dy, \tag{4.5}$$

Acknowledgement. Supported by a Grant from the City University of Hong Kong (Project No. 9380068). The several helpful comments of Michael Edesess and Robert Smythe are gratefully acknowledged.Thanks also go to Boyan Dimitrov who provided a platform which motivated the writing of this paper.

REFERENCES

[1] Barlow R. E., F. Proschan. Statistical Theory or Reliability and Life Testing: Probability Models. Holt, Rinehart & Winston, 1981.

[2] Box G. E. Science and statistics. *Journal of the American Statistical Association*, **71** (1976), No 356, 791–799.

[3] De Finetti B. Sur la condition déquivalence partielle. *Actualités scientifiques etindustrielles*, **739** (1938). Hermann et Cie, Paris, 5–18 (in French).

[4] Esary J. D., F. Proschan, D. W. Walkup et al. Association of random variables, with applications. *The Annals of Mathematical Statistics*, **38** (1967), No 5, 1466–1474.

[5] Jeffreys H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, **186** (1946), Series A, 453—461.

[6] Kendall M. G. A new measure of rank correlation. *Biometrika*, **30** (1938), No 1/2, 81–93.

[7] Kinney J. B., G. S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Science*s, **111** (2014), No 9, 3354–3359.

[8] Kullback S., L.R. Leibler. On information and suffiency. *The Annals of Mathematical Statistics*, **22** (1951), 79–86.

[9] Lehmann E. L. Some concepts of dependence. *The Annals of Mathematical Statistics*, **37** (1966), No 5, 1137–1153.

[10] Linfoot E. An informational measure of correlation. *Information and control*, **1** (1957), No 1, 85–89.

[11] Marshall A. W., I. Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, **62** (1967), No 317, 30–44.

[12] Meinhold R. J., N. D. Singpurwalla. Understanding the Kalman filter. *The American Statistician*, **37** (1983), No 2, 123–127.

[13] Popper K. The logic of scientific discovery. Routledge, 2014.

[14] Reshef D. N., Y. A. Reshef , H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti. Detecting novel associations in large data sets. *Science*, **334** (2011), No 6062, 1518–1524.

[15] Savage L. J. The foundations of statistics. Courier Dover Publications, 1972.

[16] Shannon C. A mathematical theory of communication. *Bell Sys. Tech. J.,* **27** (1948), 379–423.

[17] Singpurwalla N. D. Reliability and risk: a Bayesian perspective. John Wiley & Sons, 2006.

[18] Spearman C. The proof and measurement of association between two things. *The American journal of psychology*, **15** (1904), No 1, 72–101.

[19] Suppes P. A probabilistic theory of causation. North-Holland, Amsterdam, 1970.

[20] Székely G. J., M. L. Rizzo et al. Brownian distance covariance. *The annals of applied statistics*, **3** (2009), No 4, 1236–1265.

Nozer D. Singpurwalla
The City University of Hong Kong
Hong Kong, China
e-mail: nozer@gwu.edu

Received December 4, 2014
Final Accepted February 24, 2015