Journal of Hydrology 197 (1997) 286–292

Journal of Hydrology

Transformation and normalization of variates with specified distributions

Roman Krzysztofowicz

Department of Systems Engineering and Division of Statistics, University of Virginia, Charlottesville, VA 22903, USA

Received 9 October 1995; revised 12 July 1996; accepted 27 August 1996

Abstract

Given two continuous random variables X and Y, with specified strictly increasing cumulative distribution functions F and G, respectively, the one-to-one transform t which maps one variate into another, say $Y = t(X)$, has an analytic form, $t(X) = G^{-1}(F(X))$ or $t(X) = G^{-1}(1 - F(X))$, depending upon whether t is increasing or decreasing. This fact of probability theory is reviewed and compared with another method for finding t that was recently proposed. Applications to system identification, normalization of a variate, and normalization of a sample are briefly discussed. © 1997 Elsevier Science B.V.

1. Introduction

This paper is motivated by the work of Snyder et al. (1993) who considered the following problem. Let X and Y be two continuous variates with known probability density functions f and g, respectively, such that one variate results from a transformation of the other, say $Y = t(X)$, where t is an unknown function defined for all observations (x,y) of variates (X,Y). Under the assumption that function t is strictly increasing, Snyder et al. (1993) presented a method of identifying t. The method proceeds by finding a numerical solution to the nonlinear differential equation $g(t(x))\mathrm{d}t(x)/\mathrm{d}x = f(x)$ through which the densities f and g are related.

This note reviews an alternative method which yields an expression for t that is analytic and exact with respect to the specified distributions.

2. Analytic method

Let F and G denote the cumulative distribution functions corresponding to the known

PII S0022-1694(96)03276-3

densities f and g, respectively. Because for a specified density, the corresponding distribution is uniquely determined, the problem posed by Snyder et al. (1993) has the following equivalent form: given distributions F of X and G of Y, find the transform t so that $Y = t(X)$.

The distributions F and G are known in the sense that they may be specified in terms of parametric models with fixed parameter values, or specified nonparametrically in terms of probability plotting positions or free curves. Whether parametric or nonparametric, these distributions may be assumed, derived from physical or other laws, estimated from random samples, or assessed judgmentally by experts. In essence, the nature and source of these distributions are immaterial.

Suppose the known distributions F and G are strictly increasing, and the unknown transform t is one-to-one so that its inverse t^{-1} exists and specifies $X = t^{-1}(Y)$. Thus in terms of probability P,

$$F(x) = P(X \le x) = P\left(t^{-1}(Y) \le x\right). \tag{1}$$

When t is increasing,

$$P\left(t^{-1}(Y) \le x\right) = P(Y \le t(x)) = G(t(x)), \tag{2}$$

and by equating Eq. (1) with Eq. (2), one obtains $F(x) = G(t(x))$ from where

$$t(x) = G^{-1}(F(x)). \tag{3}$$

When t is decreasing,

$$P\left(t^{-1}(Y) \le x\right) = P(Y > t(x)) = 1 - G(t(x)), \tag{4}$$

and by equating Eq. (1) with Eq. (4), one obtains $F(x) = 1 - G(t(x))$ from where

$$t(x) = G^{-1}(1 - F(x)). \tag{5}$$

That is to say, the unknown transform t is uniquely specified by a composition of the inverse G^{-1} of the distribution of Y and the distribution F of X (when t is increasing) or the exceedance function $1 - F$ of X (when t is decreasing).

Eqs. (3) and (5) are analytic and exact. To evaluate $t(x)$, the only required tasks are evaluations of the distribution $F(x)$ and the inverse $G^{-1}(p)$ for $p = F(x)$ or $p = 1 - F(x)$. For many common probability models these are straightforward evaluations which, at most, require numerical integration.

3. Applications

Hydrologic and water resource applications which call for identification of transform t may be of three types (Snyder et al., 1993): (i) system identification, (ii) normalization of a variate, and (iii) normalization of a sample. We shall briefly illustrate how Eq. (3) could be used in the context of these types of problems.

3.1. System identification

Consider a deterministic system with a single input, X, and a single output, Y. The problem is to identify the system (transfer) function t, given the marginal distributions F and G of input and output. This problem arises, for instance, when there are observations of X and observations of Y, but there are no joint (simultaneous) observations of (X,Y) . Consequently, the system function t cannot be estimated via any classical input–output analysis. However, the marginal distributions F and G can be estimated and thus t can be identified via Eq. (3).

To illustrate, suppose that the daily precipitation had been recorded at a high elevation (variate Y) and then the rain-gauge was transferred to a low elevation at which daily precipitation has been recorded since (variate X). One may be interested in developing a model of dependence between X and Y. The general model takes the form of a joint distribution of (X,Y). However, in the absence of joint (simultaneous) observations of (X,Y), such a distribution is difficult to estimate. One may, therefore, consider, as the first approximation, a transform model $Y = t(X)$, conditional on the occurrence of precipitation at both sites and the hypothesis that the dependence between X and Y is dominated by a systematic effect of elevation. Because marginal distributions F and G can be estimated from the available records, t can be derived from Eq. (3).

Following the conclusion of a large empirical study by Selker and Haith (1990), daily precipitation is modelled in terms of the Weibull distribution. Thus the distribution of X is specified by

$$F(x)=1-\exp\left[-\left(\frac{x}{a}\right)^{b}\right],\quad x>0, \tag{6}$$

and the inverse of the distribution of Y is given by

$$G^{-1}(p)=\alpha[-\ln(1-p)]^{1/\beta},\quad 0<p<1, \tag{7}$$

where (a, b) are the scale and shape parameters of F, and (α, β) are the scale and shape parameters of G. After Eqs. (6) and (7) are inserted into Eq. (3) one finds

$$t(x)=\alpha\left(\frac{x}{a}\right)^{b/\beta},\quad x>0. \tag{8}$$

This is a power function whose shape is determined by the ratio of the shape parameters of the marginal distributions: t is concave if $b < \beta$, linear if $b = \beta$, or convex if $b > \beta$. Interestingly, this derived form of t matches the form of a transformation established empirically and used operationally to account for the elevation effect in quantitative precipitation forecasting by the Northwest River Forecast Center of the National Weather Service (Bissell, 1991). Thus, coincidentally, the probability theory verifies coherence of independent empirical investigations, those that supported the Weibull distribution for daily precipitation at a station and those that uncovered the power transform between daily precipitation amounts recorded by stations at different elevations.

3.2. Normalization of variate

The objective is to transform variate X, whose distribution F is known, to variate Y, whose distribution is Gaussian. This objective dates back to 1879 when Galton suggested the logarithmic transform (Stigler, 1986, p. 330). Numerous other transforms have been used since.

With Q denoting the standard normal distribution and Q^{-1} its inverse, the transform defined by Eq. (3) takes the form

$$Y = Q^{-1}(F(X)). \tag{9}$$

We call it the normal quantile transform (NQT). It is the most general transform and offers several advantages over others. It guarantees that variate Y is standard normal, regardless of the distribution of the original variate X. It is nonparametric and thus retains a fixed structure (and properties) regardless of the form of F. It is easy to implement since polynomial approximations of Q and Q^{-1} are readily available (Abramowitz and Stegun, 1972).

In nonparametric statistics, $s = Q^{-1}(p)$ is often referred to as the inverse normal score, or just the normal score, for a given p, $0 < p < 1$; and it appears in test statistics of Klotz (1962) and van der Waerden (1969), whose names are sometimes associated with the transform itself as, for instance, in Gibbons and Chakraborti (1992). In hydrology, NQT was used to generate samples from specified marginal distributions (Hosking and Wallis, 1988), to perform Bayesian revision of arbitrary prior distributions (Kelly and Krzysztofowicz, 1994a,b), and to obtain a bivariate distribution when the marginal distributions are specified (Krzysztofowicz et al., 1994).

To illustrate the NQT, we turn to the example used by Snyder et al. (1993) in testing their method. The density of X is specified by an isosceles triangle on the interval [–4, 4]. The corresponding distribution takes the form:

$$F(x) = \begin{cases} \frac{1}{2} + \frac{x}{4} + \frac{x^2}{32} \text{ if } & -4 \le x \le 0, \\ \frac{1}{2} + \frac{x}{4} - \frac{x^2}{32} \text{ if } & 0 \le x \le 4. \end{cases} \tag{10}$$

Fig. 1 shows a plot of the transformation $y = Q^{-1}(F(x))$. This transformation is essentially exact (the errors resulting from the polynomial approximation of Q^{-1} are on the order of 10^{-4}), and it can be compared with the plot of the transformation obtained by Snyder et al. (1993, p. 102). The plots are identical in the center, but distinct for extreme values of x, where the method of Snyder et al. (1993) produced unsatisfactory curvatures with negative slopes. In addition to being exact, the transformation shown herein requires only a fraction of the computational effort. Also, when the polynomial approximating Q^{-1} is employed, the transformation has a closed-form expression.

3.3. Normalization of sample

Normalization of the sample is a method for comparing the empirical distribution of a given variate X with the distribution of a standard normal variate Y. This was the primary objective of Snyder et al. (1993) who also discuss various hydrologic inference problems (such as estimating confidence intervals) that can be conveniently accomplished via normalization.

Let $x_{(1)}$, ..., $x_{(n)}$ be the sample of X, ordered from smallest to largest observation. Let $p_{(1)}, ..., p_{(n)}$ be the cumulative probabilities, estimated by whichever method is preferable, such that $p_{(i)} = P(X \le x_{(i)})$; for example, $p_{(i)} = i/(n+1)$. Points $\{(x_{(i)}, p_{(i)}): i = 1, ..., n\}$ form the plotting positions of the empirical distribution of X.

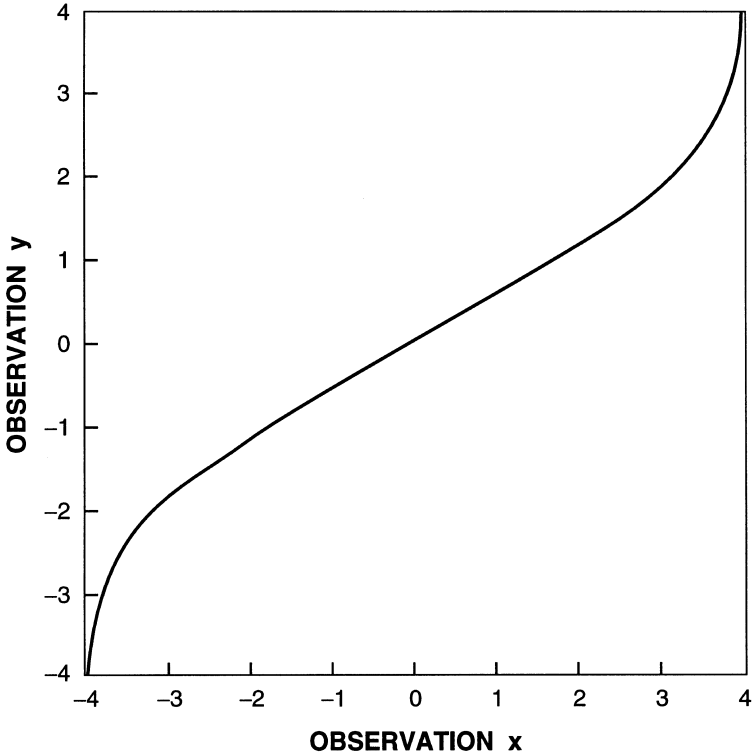

Fig. 1. Normal quantile transform of variate *X*, having distribution specified by Eq. (10).

Each observation $x_{(i)}$ of X is next transformed into observation $y_{(i)} = Q^{-1}(p_{(i)})$ of the standard normal variate Y. Because $p_{(i)} = P(Y \leq y_{(i)})$, observations $x_{(i)}$ and $y_{(i)}$ form a pair of corresponding quantiles. The plot of $y_{(i)}$ versus $x_{(i)}$ for $i = 1, ..., n$ gives a discrete estimate of the transform t. This is, in fact, the quantile–quantile plot, or Q–Q plot, widely used in statistics after Wilk and Gnanadesikan (1968) as a simple and effective tool for visualizing transforms (Cleveland, 1993). Figs. 2 and 3 show an example.

The Q–Q plot is also advantageous in the process of identifying and estimating a parametric distribution of X. Because there are no à priori restrictions on the form of t, one can identify and estimate a parametric transform t that is optimal with respect to some criterion of goodness-of-fit to the points $\{(x_{(i)}, y_{(i)})$: $i = 1, ..., n\}$ in the Q–Q plot. Once such an optimal parametric transform t is found, the distribution of X is readily specified by $F(x) = Q(t(x))$.

4. Conclusion

In investigations of transforms between random variables, the constructs of choice are cumulative distribution functions. They ensure that the task of transform identification is straightforward, and the resultant transform has an analytic form and is exact with respect

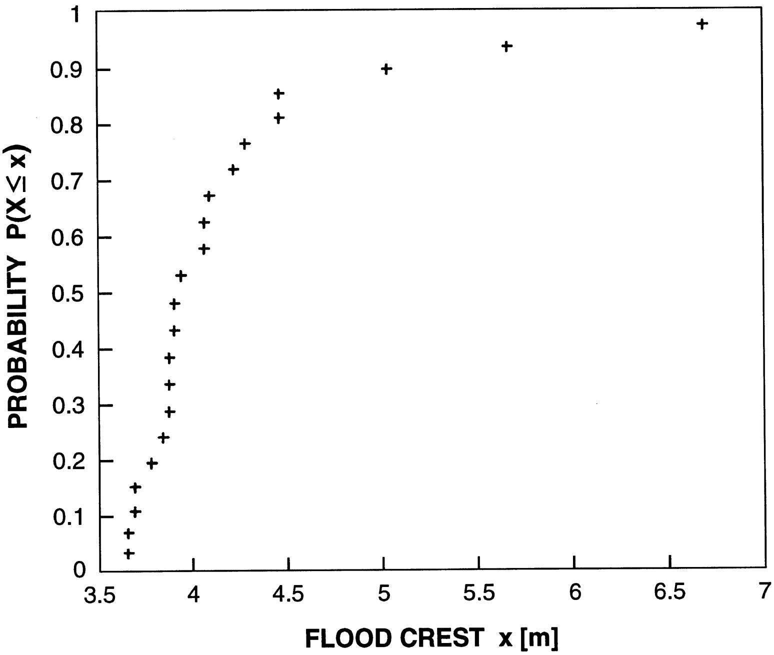

Fig. 2. Empirical distribution of the flood crest *X*, conditional on the occurrence of a flood (river stage exceeding 3.66 m), in Connellsville, Pennsylvania, based on river observations during the period 1943–1986.

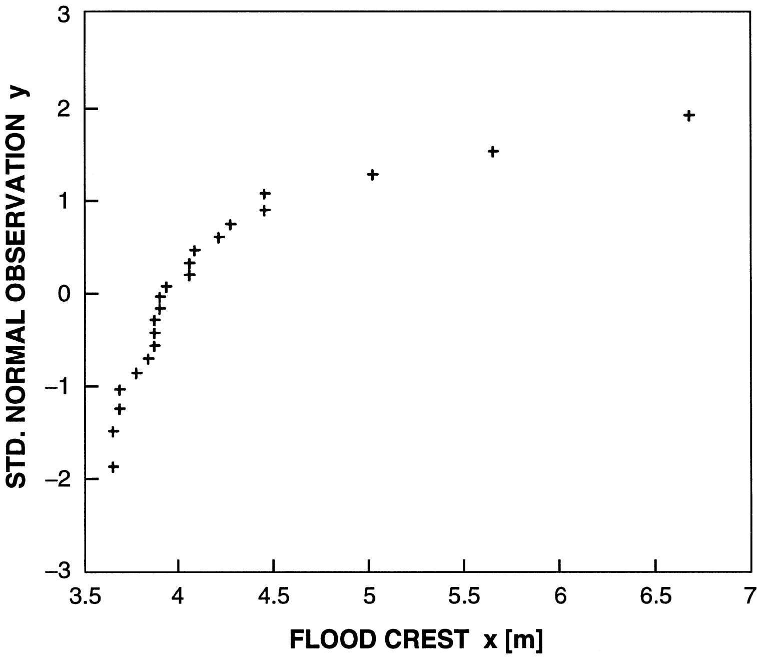

Fig. 3. Normalization of a sample of the flood crest *X* whose empirical distribution is shown in Fig. 2.

to the specified distributions. The most general method for normalizing a variate or a sample is the normal quantile transform (NQT), whose graphical representation takes the form of the quantile–quantile plot (Q–Q plot).

Acknowledgements

Research leading to this note was supported by the National Weather Service, under the project ‘‘Development of a Prototype Probabilistic Forecasting System’’, and under the Intergovernmental Personnel Act Agreement between the National Weather Service and the University of Virginia.

References

Abramowitz, M. and Stegun, I.A. (Editors), 1972. Handbook of Mathematical Functions. Dover Publications, New York.

Bissell, V.C., 1991. The ‘‘Z’’ transform: Normalizing daily precipitation for operational applications. Technical Note. Northwest River Forecast Center, National Weather Service, Portland, OR.

Cleveland, W.S., 1993. Visualizing Data. AT&T Bell Laboratories, Murray Hill, NJ.

Gibbons, J.D. and Chakraborti, S., 1992. Nonparametric Statistical Inference. Marcel Dekker, New York.

Hosking, J.R.M. and Wallis, J.R., 1988. The effect of intersite dependence on regional flood frequency analysis. Water Resour. Res., 24(4): 588–600.

Kelly, K.S. and Krzysztofowicz, R., 1994a. Probability distributions for flood warning systems. Water Resour. Res., 30(4): 1145–1152.

Kelly, K.S. and Krzysztofowicz, R., 1994b. Synergistic effect of dam and forecast on flood probabilities: a Bayesian analysis. In: L. Duckstein and E. Parent (Editors), Engineering Risk in Natural Resources Management. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 131–143.

Klotz, J., 1962. Nonparametric tests for scale. Ann. Math. Stat., 33(2): 498–512.

Krzysztofowicz, R., Kelly, K.S. and Long, D., 1994. Reliability of flood warning systems. J. Water Resour. Plann. Manage., 120(6): 906–926.

Selker, J.S. and Haith, D.A., 1990. Development and testing of single-parameter precipitation distributions. Water Resour. Res., 26(11): 2733–2740.

Snyder, W.M., Mills, W.C., Dillard, A.L. and Thomas, A.W., 1993. Normalization of a hydrologic sample probability density function by transform optimization. J. Hydrol., 149: 97–110.

Stigler, S.M., 1986. The History of Statistics. The Belknap Press of Harvard University Press, Cambridge, MA.

van der Waerden, B.L., 1969. Mathematical Statistics. Springer–Verlag, New York.

Wilk, M.B. and Gnanadesikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika, 55: 1–17.