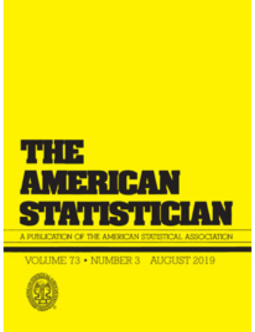

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 73 • NUMBER 3 AUGUST 2019

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Inducing Any Feasible Level of Correlation to Bivariate Data With Any Marginals

Hakan Demirtas

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. 3, 273–277: Teacher's Corner
https://doi.org/10.1080/00031305.2017.1379438

Taylor & Francis
Taylor & Francis Group

TEACHER'S CORNER

Inducing Any Feasible Level of Correlation to Bivariate Data With Any Marginals

Hakan Demirtas

Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, Chicago, IL

ABSTRACT

A simple sorting approach for inducing any desired Pearson or Spearman correlation to independent bivariate data, whose marginals can be of any distributional type and nature is described and illustrated through examples that span a broad range of situations. The proposed method has substantial potential in simulated settings that involve random number generation.

ARTICLE HISTORY
Received March 2017
Revised July 2017

KEYWORDS
Correlation; Random number generation; Simulation

1. Introduction and Motivation

Stochastic simulation is an indispensable part and major focus of scientific inquiry and has grown to be an intrinsic part of statistical practice. Describing a real notion by creating mirror images and imperfect proxies of the perceived underlying truth in a repeated manner allows researchers to study the performance of their statistical methods through simulated data replicates that mimic the real-data characteristics of interest in any given setting. Accuracy and precision measures that are obtained in a simulation study regarding the quantities under consideration signal if a procedure works properly, and may suggest corrective action to maximize the concordance between expectation and reality. Estimation and testing steps usually require verification via simulation to assess the reliability, validity, and plausibility of inferential techniques, to evaluate how well the implemented statistical models capture the specified true population values, and how reasonably these models respond to deviations from assumptions, among other things. Simulations have been commonly employed in a wide spectrum of disciplines across physical, medical, social, and managerial sciences. A central aspect of every simulation study is the quantification of the model components and parameters that jointly define a scientific process. When this quantification cannot be performed through deterministic tools, researchers often resort to random number generation (RNG) as a starting point in finding simulation-based solutions to their problems. Such situations usually involve variables of different kinds on a structural level; causal and correlational interdependencies are a function of a mixture of binary, ordinal, count, and continuous variables, which act simultaneously to characterize the mechanisms that collectively delineate a phenomenon. In nearly every RNG algorithm, an association structure among model variables needs to be specified. The Pearson and Spearman correlations are the most widely used associational quantities. The Pearson correlation measures the degree of linear association, whereas the Spearman correlation captures any monotonic relationship including nonlinear ones.

It is well known that the marginal distributions may impose bounds on correlations, leading to an interval that is narrower than the nominal $(-1, +1)$ range in most bivariate settings (Hoeffding 1940; Fréchet 1951). The concept of correlation bounds that differ from the broadest possible limits due to the marginal features applies to all types of random variables (discrete, continuous, and mixed). These correlation bounds need to be carefully considered and computed. In our experience, students as well as researchers often specify the correlations using the broadest possible limits (i.e., -1 to $+1$) in designing simulations, yet some correlation specifications are simply infeasible. Failure to define the association parameters within a plausible range can easily result in incorrect, suboptimal, imprecise, or inconsistent conclusions. Furthermore, it is a common tendency among simulation designers to try extreme situations to gauge the limits of their statistical procedures or models. Demirtas and Hedeker (2011) suggested a simple sorting approach to find the approximate bounds, taking the Wasserstein distance, which is a natural way of comparing the probability distributions of two variables (Dobrushin 1970), as a basis. The algorithm that appeared in Demirtas and Hedeker (2011) hinges upon the heuristic maximal and minimal covariation arguments within the Wasserstein distance framework, whose mathematically cogent proof was given by Rachev and Ruschendorf (1998, chap. 3). This algorithm is involved with generating random samples from the intended distributions independently using a large number of observations (e.g., $N = 100{,}000$). Subsequently, estimation of the lower/upper correlation bound is performed by sorting the two variables in the same/opposite directions, before computing the sample correlation that corresponds to maximal correlation and anti-correlation, respectively. Obviously, sorting is conducted with respect to the magnitude of numbers. Sorting in the same direction means sorting both variables from smallest to largest or from largest to smallest; sorting in the opposite direction means sorting one variable from smallest to largest, and the other one from largest to smallest. In what follows, sorting and reverse sorting stand for the same and opposite directions, respectively. In effect, this is equivalent to the inverse cdf (cumulative distribution function) method. If a RNG routine is available (it trivially is these days) to simulate

CONTACT Hakan Demirtas ✉ demirtas@uic.edu Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL 60612.

samples from the marginal distributions of interest, then one can generate random realizations, and sort them in these two different ways, as described above, to compute the approximate correlation limits. Demirtas and Hedeker (2011) gave several illustrations concerning the operational characteristics of this method. While it was not explicitly mentioned in their paper whose focus was the Pearson correlation, the method can be shown to work equally well for the Spearman correlation too.

In our teaching of graduate level courses involving computational statistics at the University of Illinois at Chicago in The Division of Epidemiology and Biostatistics, the design and execution of simulation studies is a critical part of the coursework. In such simulation studies, the specification of the correlation structure among variables is inevitably required. Students can easily employ the method in Demirtas and Hedeker (2011) when defining the range of correlations in their simulation designs. Here is a seemingly peripheral but fundamentally pivotal question posed by students: yes, we can find the approximate lower and upper correlation bounds by this technique. Can we also employ this method to induce any Pearson or Spearman correlation in any bivariate setting? In other words, sorting or reverse sorting of whole data gives the bounds, what if we do this partially? Can it be used to generate bivariate data with any desired correlation? In Section 2, we propose a potentially useful and practical sorting idea to address these questions that motivate the current work.

2. Algorithm and Illustration

Let the desired or specified correlation be δ_{spec}. The above algorithm that is designed for computing the correlation limits can be re-engineered to impose any feasible Pearson or Spearman correlation (δ_{spec}). The rationale behind this variation is extremely simple: If sorting 100% of data gives $\delta_{\max}$, sorting $100 * \delta_{\text{spec}}/\delta_{\max}$% of data yields a sample correlation δ_{spec}. Similarly, for negative correlations, if reverse sorting 100% of data gives $\delta_{\min}$, reverse sorting $100 * \delta_{\text{spec}}/\delta_{\min}$% of data yields a sample correlation δ_{spec}. Clearly, neither $\delta_{\text{spec}}/\delta_{\min}$ nor $\delta_{\text{spec}}/\delta_{\max}$ can exceed 1 for negative and positive δ_{spec}, respectively; that is, one cannot sort or reverse sort more than 100% of data. Here is a summary of findings presented in this work:

- The percentage of sorting is linearly related to the upper and lower bounds. If the upper bound is $\delta_{\max}$ and the desired correlation is $\delta_{\text{spec}} > 0$, the proportion of data that should be sorted is $\delta_{\text{spec}}/\delta_{\max}$. The same argument applies to negative correlation values and the lower bound: if $\delta_{\text{spec}} < 0$, the percentage of data that should be reverse sorted is $\delta_{\text{spec}}/\delta_{\min}$.
- The above logic is equally valid for the Pearson and Spearman correlations.
- This idea works for any type of bivariate data with any combination of binary, ordinal, count, and continuous marginals.
- One can induce any Pearson or Spearman correlation on independent bivariate data regardless of data type via the following algorithm:
  1. Generate random samples from the intended distributions independently using a large number of observations (e.g., $N = 100{,}000$).
  2. Compute the approximate upper ($\delta_{\max}$) and lower ($\delta_{\min}$) correlation bounds, by sorting and reverse sorting the whole data, respectively (Demirtas and Hedeker 2011).
  3. To induce a specified correlation value ($\delta_{\text{spec}} > 0$), sort $100 * \delta_{\text{spec}}/\delta_{\max}$% of the original, uncorrelated data in Step 1. If $\delta_{\text{spec}} < 0$, reverse sort $100 * \delta_{\text{spec}}/\delta_{\min}$% of data.

Operationally, let X_1 and X_2 be the two components in a bivariate setting with δ_{spec}. In the case of $\delta_{\text{spec}} > 0$, after generating univariate data of size N, independently for X_1 and X_2, one can randomly pick $N * 100\delta_{\text{spec}}/\delta_{\max}$% of rows, sort the numbers in the same direction (e.g., from smallest to largest) separately for each of the variables, leaving the data in unpicked rows intact. If $\delta_{\text{spec}} < 0$, the same procedure is applied in a way that the sorting operation is performed in different directions (e.g., from smallest to largest for X_1, from largest to smallest for X_2). The resulting sample correlation will be very close to δ_{spec}. From a procedural point of view, any subset of independent data (any combination of rows as long as the same set of rows are selected for sorting or reverse sorting) is fine as long as N is large enough. Even when N is not large, the discrepancies between the specified and empirical correlations should be negligibly small across simulation replicates on average, as demonstrated later in the article.

The illustrative examples come from combinations of five distributions: two continuous and three discrete distributions are used in what follows. A standard normal variable, an exponentially distributed variable with a rate parameter 1, a Bernoulli variable with a proportion parameter 0.8, a Poisson variable with a rate parameter 2, and an ordinal variable with probabilities p_i, where $p_i = (0.4, 0.3, 0.2, 0.1)$ for $i = 1, 2, 3, 4$. These distributions are denoted by Φ, $\exp(1)$, $\text{Bern}(0.8)$, $\text{Poi}(2)$, and Ord, in the above order. Twelve combinations that include continuous–discrete, continuous–continuous, discrete–discrete pairs (some are identically distributed, some are not) were considered, the results for the two of them are given in the text below for exposition, and those for the remaining ten combinations are shown in Table 1. Note that (a) four significant digits after decimal are reported throughout the article, (b) the results may not be exact due to rounding, (c) the reported quantities may slightly vary across different computing platforms given the stochastic nature of the procedure, and (d) if at least one of the variables has a symmetric distribution, the lower and upper bounds should be symmetric around zero in theory, but they may not precisely be so in practice (due to the inherent simulation error).

Let us focus on $\text{Poi}(2) - \text{Bern}(0.8)$ and $\Phi - \text{Ord}$ pairs for the purpose of illustration. Let P and S denote the Pearson and Spearman correlations, respectively. First, let $X_1 \sim \text{Poi}(2)$, $X_2 \sim \text{Bern}(0.8)$, and $N = 1{,}000{,}000$. It turns out that $(\delta_{\min}, \delta_{\max}) = (-0.7389, 0.5924)$ for P and $(-0.6800, 0.6521)$ for S. When the correlation type is P and $\delta_{\text{spec}} = 0.3$, $\delta_{\text{spec}}/\delta_{\max} = 0.5064$, meaning that 50.64% of data should be sorted. After this operation, the sample (empirical) correlation (δ_{emp}) is computed to be 0.3008. When we specify a negative (-0.5) δ_{spec}, it leads to a reverse sorting ratio of $\delta_{\text{spec}}/\delta_{\min} = 0.6767$ and $\delta_{\text{emp}} = -0.4996$. When the correlation type is S and $\delta_{\text{spec}} = 0.2$, the associated sorting ratio is $\delta_{\text{spec}}/\delta_{\max} = 0.3067$ and $\delta_{\text{emp}} = 0.2007$. Changing δ_{spec} to -0.6 results in a reverse sorting ratio of 0.8824 and $\delta_{\text{emp}} = -0.6000$. Second, let us examine the

Table 1. The large sample case: results for 10 distributional pairs across two different correlation types and two correlation levels. P, S, $\delta_{\min}$, $\delta_{\max}$, δ_{spec}, and δ_{emp} stand for the Pearson, Spearman, minimum, maximum, specified, and empirical correlation, respectively.

X_1	X_2	Corr.Type	$\delta_{\min}$	$\delta_{\max}$	δ_{spec}	Sort.Ratio	δ_{emp}
Φ	Bern(0.8)	δ_P	−0.6999	0.7002	0.5	0.7141	0.5006
					−0.3	0.4289	−0.2993
		δ_S	−0.6928	0.6928	0.2	0.2887	0.1996
					−0.6	0.8649	−0.6002
Φ	Poi(2)	δ_P	−0.9596	0.9595	0.9	0.9380	0.9001
					−0.4	0.4169	−0.3983
		δ_S	−0.9753	0.9753	0.1	0.1025	0.0992
					−0.8	0.8202	−0.7996
Φ	Exp(1)	δ_P	−0.9032	0.9026	0.7	0.7755	0.7005
					−0.5	0.5535	−0.5025
		δ_S	−1.0000	1.0000	0.1	0.1000	0.1010
					−0.6	0.6000	−0.5992
Exp(1)	Bern(0.8)	δ_P	−0.8049	0.4463	0.4	0.8962	0.4001
					−0.3	0.3726	−0.2991
		δ_S	−0.6926	0.6926	0.5	0.7219	0.4995
					−0.2	0.2887	−0.2013
Exp(1)	Poi(2)	δ_P	−0.7824	0.9445	0.9	0.9529	0.8996
					−0.7	0.8936	−0.7002
		δ_S	−0.9753	0.9753	0.8	0.8203	0.7991
					−0.5	0.5127	−0.5000
Exp(1)	Ord	δ_P	−0.7104	0.8975	0.6	0.6685	0.5980
					−0.5	0.7028	−0.5021
		δ_S	−0.9488	0.9488	0.3	0.3162	0.2994
					−0.8	0.8433	−0.7994
Bern(0.8)	Ord	δ_P	−0.7499	0.4997	0.1	0.2001	0.1009
					−0.4	0.5329	−0.3972
		δ_S	−0.6853	0.5493	0.2	0.3641	0.2009
					−0.3	0.4379	−0.2999
Poi(2)	Ord	δ_P	−0.8311	0.9395	0.7	0.7451	0.7014
					−0.8	0.9612	−0.7999
		δ_S	−0.9071	0.9612	0.6	0.6242	0.6002
					−0.9	0.9919	−0.9000
Poi(2)	Poi(2)	δ_P	−0.8872	0.9994	0.5	0.5003	0.5005
					−0.4	0.4509	−0.4001
		δ_S	−0.9469	0.9995	0.9	0.9004	0.8997
					−0.7	0.7394	−0.7009
Ord	Ord	δ_P	−0.8009	0.9996	0.4	0.4002	0.3987
					-0.2	0.2500	−0.1982
		δ_S	−0.8735	0.9994	0.6	0.6004	0.5995
					−0.8	0.9156	−0.8000

pair of Φ − Ord. By following the same steps, $(\delta_{\min}, \delta_{\max}) = (-0.9095, 0.9093)$ for P and $(-0.9488, 0.9488)$ for S. When the correlation type is P and $\delta_{\text{spec}} = 0.8$, $\delta_{\text{spec}}/\delta_{\max} = 0.8797$, meaning that 87.97% data should be sorted. After this operation, the sample (empirical) correlation (δ_{emp}) is computed to be 0.8001. When $\delta_{\text{spec}} = -0.7$, then the reverse sorting ratio is $\delta_{\text{spec}}/\delta_{\min} = 0.7697$ and $\delta_{\text{emp}} = -0.6992$. When the correlation type is S and $\delta_{\text{spec}} = 0.6$, the sorting ratio is $\delta_{\text{spec}}/\delta_{\max} = 0.6324$ and $\delta_{\text{emp}} = 0.6003$. Changing δ_{spec} to -0.4 leads to a reverse sorting ratio of 0.4216 and $\delta_{\text{emp}} = -0.4007$.

In Table 1, we report the results for ten other distributional pairs across two levels of the Pearson and Spearman correlations (one positive and one negative). In all 40 cases, the reported results lend strong support for how well the proposed algorithm works; the distances between expected (desired) and computed (empirical) quantities are minimal. In addition, as pointed out by a reviewer, in the spirit of evaluating the performance at distributional extremes, we take a look at a Poisson variable with a very small rate parameter (0.009) and a Bernoulli variable with a very small proportion parameter (0.001), $X_1 \sim \text{Poi}(0.009)$ and $X_2 \sim \text{Bern}(0.001)$. The algorithm delivers satisfactory results even when the parameters are chosen to be close to their boundaries. The correlation bounds in this case are $(\delta_{\min}, \delta_{\max}) = (-0.0031, 0.3512)$ for P and $(-0.003, 0.3347)$ for S. When the correlation type is P and $\delta_{\text{spec}} = 0.2$, $\delta_{\text{spec}}/\delta_{\max} = 0.5695$ and $\delta_{\text{emp}} = 0.1991$. When the correlation type becomes S and $\delta_{\text{spec}} = 0.1$, $\delta_{\text{spec}}/\delta_{\max} = 0.2988$ and $\delta_{\text{emp}} = 0.1003$.

For the assessment of the algorithmic performance in smaller samples ($N = 100$), we devised a simulation study with 1,000 replicates. We again use Poi(2) − Bern(0.8) and Φ − Ord pairs. The quantities reported in Table 2 are the same as the ones in Table 1 except that in the last column, the average empirical correlation across 1,000 simulation replicates ($\overline{\delta}_{\text{emp}}$) is shown. This is probably more realistic, because a typical simulation scheme includes many simulated samples of moderate sizes. The fundamental picture is the same, the proximity between the

Table 2. The moderate-size sample case: $\bar{\delta}_{\text{emp}}$ stands for the average empirical correlation across 1,000 simulation replicates when $N = 100$.

X_1	X_2	Corr.Type	$\delta_{\min}$	$\delta_{\max}$	δ_{Spec}	Sort.Ratio	$\bar{\delta}_{\text{emp}}$
Poi(2)	Bern(0.8)	δ_P	−0.7389	0.5924	0.3	0.5064	0.3042
					−0.5	0.6763	−0.4968
		δ_S	−0.6800	0.6521	0.2	0.3067	0.2016
					−0.6	0.8833	−0.5975
Φ	Ord	δ_P	−0.9095	0.9094	0.8	0.8797	0.7978
					−0.7	0.7696	−0.6996
		δ_S	−0.9488	0.9488	0.6	0.6324	0.5990
					−0.4	0.4216	−0.3967

specified and realized correlation levels is promising enough for us to take a point of advocacy for the proposed technique.

The reported results are reproducible and the R code (R Development Core Team 2018) used in this article is publicly accessible at *http://demirtas.people.uic.edu/computer_code_corr_paper.txt*

The only limitation of the proposed approach is that the specifications and the distributional assumptions are about the marginal distributions along with an appropriate correlational quantity. The algorithm does not necessarily preserve the properties of joint distribution since sorting (or reverse sorting) is performed independently and separately for each variable. However, the method's advantages far outweigh this concern. In practice, marginal behavior of the variables is often perceived as more consequential than positing a joint distribution that would be daunting to justify in the presence of radically different forms of marginals for modeling purposes. No complicated computational tools are needed, it works for all types of marginal distributions; it is very straightforward and useful. In most applications of RNG, the focus is on the marginal characteristics. The proposed algorithm does maintain the characteristics of marginal distribution mitigating the aforementioned concern. The association between variables is summarized by a correlation value without regard to the true underlying interdependence structure. While we realize that this is a simplified view, extra generality comes with this necessary evil and computational convenience given the operational attributes of the algorithm. The joint characteristics are typically distribution-specific and it is almost impossible to formulate a general umbrella that can handle a wide range of joint distributions that can be encountered in practice. Not fully accommodating the joint characteristics is the price we pay for broader applicability, flexibility, and far greater generality. This method literally covers the whole universe of bivariate distributions (a desired level of the Pearson or Spearman correlation is induced while the marginal features are preserved), and it does it in a way that anybody can implement without needing any statistical dexterity or sophisticated tools. Of note, some other techniques to induce correlations to independent data have appeared in the literature (e.g., Iman and Conover 1982) but they only encompass a subset of the general framework presented here (working only for continuous variables, allowing only one type of correlation, etc.).

Students' reaction has been very positive to this method considering its utility and generality. They expressed two concerns though: first, the required sorting ratio that comes out of this algorithm is a number that has at least a few digits after decimal (the choice of the number of significant digits is ultimately user-dependent, but we recommend at least four digits for accuracy). Suppose that $\delta_{\text{spec}}/\delta_{\max} = 0.7856$. If we need to generate data with say 100 rows, we are supposed to sort either 78 rows or 79 rows, which may cause some bias. Our recommendation was a simple algebraic operation in which there are 44% and 56% chances of sorting 78 and 79 rows, respectively. This will give the desired sorting ratio on average across simulation replicates. Alternatively, one could generate much bigger data and choose a subset of desired size by sampling rows without replacement (this is how we did it for $N = 100$). Second, we all know that most datasets in simulated and real life are multivariate. The ability to induce any feasible correlation in the bivariate case may not be sufficiently complex for some applications. There is no way to contradict this assertion, however, all research has to start somewhere, and the proposed method could serve as an important milestone for future work that will involve the multivariate extension. This can be accomplished via the use of linear algebraic tools or statistical distances. This augmentation has potential to grow instrumental in the RNG and simulation worlds. On a related note, existing multivariate data generation techniques (e.g., Amatya and Demirtas 2016, Demirtas 2006, Demirtas and Doganay 2012, Demirtas and Yavuz 2015, Demirtas, Hedeker, and Mermelstein 2012; Demirtas et al. 2017) require statistically specialized routines and mechanisms, whereas the multivariate version of the sorting approach set forth here is likely to be less esoteric and more comprehensible by a larger group of people.

Acknowledgment

The author is grateful to the former Editor, Nicole Lazar, to an anonymous referee and Prof. Robert C. Bailey for suggestions that considerably improved the article.

References

Amatya, A., and Demirtas, H. (2016), "Concurrent Generation of Multivariate Mixed Data with Variables of Dissimilar Types," *Journal of Statistical Computation and Simulation*, 86, 3595–3607. [276]

Demirtas, H. (2006), "A Method for Multivariate Ordinal Data Generation Given Marginal Distributions and Correlations," *Journal of Statistical Computation and Simulation*, 76, 1017–1025. [276]

Demirtas, H., Allozi, R., Hu, Y., Inan, G., and Ozbek, L. (2017), "Joint Generation of Binary, Ordinal, Count, and Normal Data with Specified Marginal and Association Structures in Monte-Carlo Simulations (pp. 3–15)," In *ICSA Book Series in Statistics, John Dean Chen and Ding-Geng (Din) Chen (Eds): Monte-Carlo Simulation-Based Statistical Modeling*. Singapore: Springer. [276]

Demirtas, H., and Doganay, B. (2012), "Simultaneous Generation of Binary and Normal Data with Specified Marginal and Association Structures," *Journal of Biopharmaceutical Statistics*, 22, 223–236. [276]

Demirtas, H., and Hedeker, D. (2011), "A Practical Way for Computing Approximate Lower and Upper Correlation Bounds," *The American Statistician*, 65, 104–109. [273,274]

Demirtas, H., Hedeker, D., and Mermelstein, J. M. (2012), "Simulation of Massive Public Health Data by Power Polynomials," *Statistics in Medicine*, 31, 3337–3346. [276]

Demirtas, H., and Yavuz, Y. (2015), "Concurrent Generation of Ordinal and Normal Data," *Journal of Biopharmaceutical Statistics*, 25, 635–650. [276]

Dobrushin, R. L. (1970), “Prescribing a System of Random Variables by Conditional Distributions,” *Theory of Probability and its Applications*, 15, 458–486. [273]

Fréchet, M. (1951), “Sur Les Tableaux de Corrélation Dont Les Marges Sont Donnés,” *Annales de l’Universite de Lyon Section A*, 14, 53–77. [273]

Hoeffding, W. (1940), in *Scale-Invariant Correlation Theory* eds. N. I. Fisher and P. K. Sen, The Collected Works of Wassily Hoeffding, pp. 57–107. New York: Springer-Verlag. [273]

Iman, R. L., and Conover, W. J. (1982), “A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables,” *Communications in Statistics: Simulation and Computation*, 11, 311–334. [276]

R Development Core Team (2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: Foundation for Statistical Computing. [276]

Rachev, S. T., and Ruschendorf, L. (1998), *Mass Transportation Problems, Volume I: Theory*, New York: Springer-Verlag. [273]