GENERALIZED VARIANCE

Ashis Sengupta

Indian Statistical Institute, Applied Statistics University, Kolkata, India, ashis@isical.ac.in

Abstract. Generalized variance (GV), proposed by Wilks, is an one-dimensional measure of multidimensional scatter. It plays an important role in both theoretical and applied research. GV has been extended to Standardized GV (SGV) to enable comparison of scatters in differing dimensions. Interesting results on the distribution, estimation, and testing of GVs and SGVs have been emerging. Also, these measures have been finding novel applications in a variety of applied problems, ranging from such age-old areas as agriculture and sample surveys to the modern developments in signal processing and Bioinformatics.

Keywords and Phrases. Bayesian analysis, D-optimality, multidimensional scatter, Pincherle's H-function, $|S|$-chart, signal processing, standardized generalized variance, Time series analysis.

Blind Entry.

AMS Subject Classification.

Introduction

The generalized variance (GV) of a p-dimensional random vector variable X is defined as the determinant of its variance-covariance matrix. GV was introduced by Wilks [33, 34, 36] as a scalar measure of overall multidimensional scatter. We will denote the population and sample GVs by $|\Sigma|$ and $|S|$ respectively. The standardized

•Q1 GV (SGV) [24, 25•] of X is the positive pth root of GV. GV has several interesting interpretations. For an optimal estimator of a multidimensional parameter, GV is the reciprocal of the determinant of the information matrix. GV is the only criterion for which the function-induced ordering of information matrices is invariant under reparameterization. The determinant function is invariant under reparameterization in the D-group, that is, under such transformations as with determinant of the coefficient matrix being $+1$ or -1. The modal value of the pdf of a multivariate normal distribution is inversely proportional to the square root of the GV. Under normality, in a linear model set up, the optimal confidence ellipsoid of an estimable parametric function has (the smallest) volume that is inversely proportional to the square root of the GV of the optimal estimator.

While GV is used as a measure of multidimensional scatter, SGV can serve the same purpose as well as be used for comparing scatter in different dimensions. Further, for applying the theory of information functions, Pukelsheim [20] points out that SGV is to be preferred.

Here, we present the developments on GV and SGV in both the areas of theory and applications, mostly since the review by Kocherlakota and Kocherlakota [16]. While some interesting theoretical results on GV and SGV have emerged in the aspects of distributional derivations and statistical inference, to date little seems to be available beyond the underlying multivariate normal distribution or its ramifications. On the other hand, there has been a rich variety in novel applications of GV and SGV. This ranges from their use as theoretical measures for comparison of estimators and construction of test statistics to their use as applied scalar measures of overall multidimensional scatter spanning such an age-old area as sample surveys to a newly emerging area as bioinformatics.

Advances In Theory

An elementary yet expository interpretation of GV, including related geometry, has been given by Johnson and Wichern [15, pp. 129–145]. The exact distributions of $|S|$ for an underlying real and complex (S being the Hermitian sample covariance matrix then) multivariate Gaussian distribution are available from Mathai [18] and Goodman [12] respectively. The exact distribution of GV when a sample is drawn

from a mixture of two multivariate normal populations has been expressed in terms of Meijer's G-functions and related residues, as for the nonmixture case, by Castañeda and Nagar [5]. The asymptotic distribution of ln $|S|$ here also is normal.

Since the exact distribution of the sample GV, though available, is quite complicated, good approximations are of interest and usefulness. In the context of generalized Wilk's Λ statistic, a product of certain independent Beta random variables, near-exact as well as asymptotic distributions, have been suggested through generalized near-integer Gamma distributions [7]. For obtaining tail probabilities, saddle-point method has been enhanced [31]. These approaches may be explored explicitly for GV, a product of certain independent Gamma random variables.

Likelihood ratio tests (LRTs) for SGVs in one or several different-dimensional multivariate normal populations and the exact, in terms of Pincherle's H-function, and asymptotic distributions of some of these test statistics have been given by SenGupta [24]. A multivariate version of Hartley's F_{max} statistic that provides a shortcut test for homogeneity of variances in the univariate case may be based on SGVs. Such a test and associated distributional results have been given by SenGupta [25]. Union–intersection tests for SGVs are available from Reference 21. Tests for SGVs with equicorrelation structures and under order constraints have been considered through the technique of isotonic regression by SenGupta [23]. Bhandary [4] has derived LRTs for GVs having specified structures under some additive models relevant to signal processing. Specifically, the testing problem under white noise case can be formulated as follows. Let $X_1, \ldots, X_n$ be independent and identically distributed as $N_p(0, \Gamma + \sigma^2 I_p)$, where Γ is a nonnegative definite matrix of rank $q(< p)$. The problem then is to test $H_0 : |\Gamma + \sigma^2 I_p| = \sigma_0^2$ against $H_1 : |\Gamma + \sigma^2 I_p| \neq \sigma_0^2$. The LRT statistic can be written down explicitly. For the case of colored noise, the underlying distribution is taken as $N_p(0, \Gamma + \sum_1)$, where $\sum_1$ is some arbitrary positive definite matrix. Assuming that there is an independent estimate of $\sum_1$, the null and alternative hypotheses are formulated similar to those in the case of white noise above and the LRT has been derived.

Results of Reference 24 on LRTs for a single specified SGV and for the equality of several SGVs have been extended by Behara and Giri [3] to the complex case. More specifically, they formulate the problem as follows. For $i = 1, 2$, let $\xi_1^{(i)}, \ldots, \xi_{N_i}^{(i)} (N_i >$

p_i) be a sample from a p_i-dimensional complex Gaussian population with mean γ_i and Hermitian positive definite covariance matrix $\sum_i$. Let $\overline{\xi}^{(i)} = \sum_{j=1}^{N_i} \xi_j^{(i)}/N_i$ and $d_i^2 = [\sum_{j=1}^{N_i}(\xi_j^{(i)} - \overline{\xi}^{(i)})(\xi_j^{(i)} - \overline{\xi}^{(i)})^*/N_i]^{1/p_i}$. The one- and two- population testing problems for the complex setup then respectively become those of testing the hypotheses (i) $H_{01} : |\sum_1|^{1/p_1} = \sigma_0^2$ (given) and (ii) $H_{02} : |\sum_1|^{1/p_1} = |\sum_2|^{1/p_2}$ against not equality alternatives. The LRTs take similar forms as for the real cases. Their critical regions are given respectively by (i) $\omega_1 : d_1^2/\sigma_0^2 > C_0$ or $< C_1$ and (ii) $\omega_2 : F = d_1^2/d_2^2 < D_0$. The exact distributions of these test statistics can be represented, as for the real cases, in terms of products of independent χ^2 and F random variables.

The admissibility and Bahadur optimality of the LRT for a GV for an underlying multivariate normal population have been established by SenGupta and Pal [28]. However, detailed studies on the properties, including those on unbiasedness and monotonicity, of the above LRTs for GVs are to be explored. In yet another direction, ranking and selection procedures based on SGVs, as already available for GVs, remain to be developed.

Nonparametric multivariate notions of "Scatter" and "More Scattered" based on statistical depth functions are being studied by several researchers (see e.g., Ref. 39), which may be viewed as counterparts of GV, primarily proposed for the parametric situation.

Advances In Applications

GV has been found quite useful in statistical inference for multidimensional parameters. In the theory of estimation, GV plays an important role as a measure of efficiency of an estimator. Also estimators having minimum GV are functions of the
•Q2 sufficient statistic see, for example, Reference• 40, Sec. 5a.3. In the theory of testing of hypotheses, Isaacson [14] has used the criterion of maximizing the total curvature of the power hypersurface to suggest a locally best (optimal) test for a multidimensional simple hypothesis. This criterion is related to the determinant of the power Hessian matrix [29].

In Bayesian inference, one approach (see e.g., Ref. 41, pp. 53–54) of constructing noninformative or vague priors for the scalar parameter θ is based on Jeffrey's prior. This prior results from the requirement of invariance of probabilistic inferences made

about the data and is proportional to $I(\theta)^{1/2}$, I being the Fisher's information for θ. Generalization of this approach to the multiparameter case yields priors proportional to the $|I(\theta)|^{1/2}$. However, caution must be taken with the choice of Jeffrey's prior when parameters are deemed dependent, see, for example, Reference 19, pp. 87–89, for further discussions. Recall now that $|I(\theta)|$ corresponds to the inverse of the GV associated with the optimal (frequentist) estimators of the parameters. In particular, with the multivariate normal distribution $N_p(\mu, \Sigma)$, Geisser and Cornfield [9] suggested the vague prior density $g(\Sigma)$ for Σ to be proportional to the inverse of the $(p+1)/2$-th power of the GV, that is, $g(\Sigma)\alpha|\Sigma|^{-(p+1)/2}$. Interestingly, in case μ and Σ are a *a priori* independent, this vague prior coincides with Jeffrey's invariant prior. While dealing with independent parameters in several Poisson populations, Leonard and Hsu [42, p.220] present Jeffrey's prior, that is, the prior inversely proportional to the positive square-root of such GV. They also illustrate its use by a real-life example from cross-classified data on performance evaluation of engineering apprentices. Applications of such priors are envisaged in many other areas including business and industry.

As noted earlier, while GV is used as a measure and for comparison of multidimensional scatters of equal dimensions, the same purposes are served by SGVs even for differing dimensions. The latter situation arises when one encounters missing data on the components of a vector variable arranged in several groups with data available on different number of components of the variable over the different groups. An example from speech recognition or talker identification problem is discussed in Reference 25.

For problems in sample surveys involving correlated characters, minimization of GV has been employed to achieve optimal allocation in a multistage survey [6] and in stratified sampling [2, 10].

Construction of optimal designs through the D-optimality criterion aims at maximizing the determinant of the moment matrix of a design measure. This is achieved by maximizing the determinant of the information matrix that is equivalent to the minimization of the SGV of the optimal (least-squares) estimator for the relevant parameter system of interest. Characterizations of the D-, equivalently, SGV-, optimality criterion are available on the basis of certain invariance property and on certain information functions, see , for example, Reference 20, Sec. 13.8. Even for discrete

design measures, optimal n-point discrete designs may be obtained by maximizing the relevant GV or SGV.

The use of GV as a measure of overall variability has been exemplified in agricultural science by Goodman [13] and in behavioral and life sciences by Sokal [30].

In the construction of optimal predictors, Garcia Ben and Yohai [38] demonstrate that in contrast to the Trace criterion, predictors obtained by minimizing the GV have the appealing property of coinciding with the canonical variables.

GV has been popular in several areas of applied multivariate analysis. In the techniques for reduction of dimensionality, GV has been used as a criterion for the construction of generalized canonical variables [22, 27] by various authors, for example, Kettenring [43], SenGupta [23], and Coelho [44].

In classification and discriminant analysis, Wilks [35] obtains the optimal linear discriminant function by determining the space of dimension $t < p$, where $|T_t|/|W_t|$, the ratio of the total to pooled within t-dimensional scatters (GVs), is maximized. To explore the structure of heterogeneous multivariate data, Friedman and Rubin [8] have advocated that partitioning in a predetermined number of groups that maximizes the criterion $|T_p|/|W_p|$. They argued that this criterion has the desirable property of being invariant under nonsingular linear transformations, has greater sensitivity to the local structure of several data sets, and is computationally faster than the Trace criterion. Clustering techniques based on GV have been proposed by SenGupta [21].

Behara and Giri [3] point out that GV and SGV are quite useful for assessing the variability of estimators of spectral density matrix of a multiple stationary Gaussian time series (see also Ref. 11) and for testing of hypotheses concerning the overall variabilities in terms of SGVs of multiple Gaussian time series of different dimensions.

In the context of signal processing, Bhandary [4] has derived LRTs for several statistical hypotheses involving GVs under both white and colored noises.

For modelling in reliability analysis, Tallis and Light [32] compare GVs to study the efficiency of their generalized method of moments estimator relative to that of the m.l.e. for an underlying mixture distribution.

$|S|$-chart, which we may term as GV-chart also, and its ramifications motivated by GV have been enhanced in statistical quality control for monitoring process variability

of a product with multiple correlated quality characteristics. Yeh et al. [37] give several real-life applications of such charts.

In bioinformatics, the use of GV for the identification of differentially expressed genes from microarray data has been advocated by SenGupta [26].

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. John Wiley & Sons, New York.

[2] Aravantis, L.G. and Afonja, B. (1971). Use of the generalized variance and the gradient projection method in multivariate stratified sampling. *Biometrics*, **27**, 119–127.

[3] Behara, M. and Giri, N. (1983). Generalized variance statistic in testing of hypothesis in complex multivariate Gaussian distributions. *Arch. Math.*, **40**, 538–543.

[4] Bhandary, M. (1996). Test for generalized variance in signal processing. *Stat. Probab. Lett.*, **27**, 155–162.

[5] Castañeda, M.E. and Nagar, D.K. (2003). Distribution of generalized variance
•Q3 under mixture normal model. *J. Appl. Stat. Sci.*; to appear•.

[6] Chakravarti, I.M. (1954). On the problem of planning a multistage survey for multiple correlated characters. *Sankhyā Ser. A*, 211–216.

•Q4 [7] •Coelho, C.A. (2004). The generalized near-integer Gamma distribution: a basis for 'near-exact' approximations to the distribution of statistics which are the product of an odd number of independent Beta random variables. *J. Multivariate Anal.*, **89**, 191–218.

[8] Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data. *J. Am. Stat. Assoc.*, **62**, 1159–1178.

[9] Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. R. Stat. Soc. Ser. B*, **25**, 368–376.

[10] Ghosh, S.P. (1958). A note on stratified random sampling with multiple characters. *Cal. Stat. Assoc. Bull.*, **8**, 81–90.

[11] Goodman, N.R. (1963a). Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *Ann. Math. Stat.*, **34**, 152–177.

[12] Goodman, N.R. (1963b). The distribution of the determinant of a complex Wishart distributed matrix. *Ann. Math. Stat.*, **34**, 178–180.

[13] Goodman, M. (1968). A measure of 'overall variability' in populations. *Biometrics*, **24**, 189–192.

[14] Isaacson, S.L. (1951). On the theory of unbiased tests of simple statistical hypotheses specifying the values of two or more parameters. *Ann. Math. Stat.*, **22**, 217–234.

[15] Johnson, R.A. and Wichern, D.W. (1999). *Applied Multivariate Statistical Analysis*, 4th ed. Prentice Hall, N.J.

[16] Kocherlakota, S. and Kocherlakota, K. (1983). "Generalized Variance". In *Encyclopedia of Statistical Sciences*, Vol. 3, N.L. Johnson and S. Kotz, eds. John Wiley & Sons, New York, pp. 354–357.

[17] Leonard, T. and Hsu, J.S.J. (1999). *Bayesian Methods*. Cambridge University Press, Cambridge, Mass.

[18] Mathai, A.M. (1972). The exact distributions of three multivariate statistics associated with Wilk's concept of generalized variance. *Sankhyā Ser. A*, **34**, 161–170.

[19] Press, S.J. (2003). *Subjective and Objective Bayesian Statistics*, 2nd ed. John Wiley & Sons, New York.

[20] Pukelsheim, F. (1993). *Optimal Design of Experiments*. John Wiley & Sons, New York.

[21] SenGupta, A. (1982). Tests for simultaneously determining numbers of clusters and their shape with multivariate data. *Stat. Probab. Lett.*, **1**, 46–50.

[22] SenGupta, A. (1983). "Generalized Canonical Variables". In *Encyclopedia of Statistical Sciences*, Vol. 3, N. Johnson and S. Kotz, eds. John Wiley & Sons, New York, pp. 123–126.

[23] SenGupta, A. (1986). "On Tests Under Order Restrictions in Reduction of Dimensionality". In *Advances in Order Restricted Statistical Inference*, Vol. 37, R. Dykstra et al.•, ed. Lecture notes in Statistics, Springer-Verlag, New York, pp. 249–256.

•Q5

[24] SenGupta, A. (1987a). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *J. Multivariate Anal.*, **23**, 209–219 [For further details see SenGupta (1981), Tech Rep. 50, Dept of Statistics, Stanford University].

[25] SenGupta, A. (1987b). Generalizations of Bartlett's and Hartley's tests of homogeneity using "Overall Variability". *Commun. Stat.—Theory Methods*, **16**, 987–996.

[26] SenGupta, A. (2003). *Statistical analysis of differentially expressed genes in micro-array data*. Lecture Notes on *Statistical Methods in Bioinformatics*, Workshop on Bioinformatics, International Biometric Society (IR), Banaras Hindu University, India, pp. 35–42.

[27] SenGupta, A. (2004??). "Generalized Canonical Variables". In *Encyclopedia of Statistical Sciences*, N. Balakrishnan, N. Johnson and XXXXX, eds. John Wiley & Sons, New York•.

•Q6

[28] SenGupta, A. and Pal, C. (2003). On the admissibility and Bahadur optimality of the LRT for a generalized variance; submitted• for publication.

•Q7

[29] SenGupta, A. and Vermeire, L. (1986). Locally optimal tests for multiparameter hypotheses. *J. Am. Stat. Assoc.*, **81**, 819–825.

[30] Sokal, R.R. (1965). Statistical methods in systematics. *Biol. Rev. Cambridge Philos. Soc.*, **40**, 337–391.

[31] Srivastava, M.S. and Yau, W.K. (1989). Saddlepoint methods for obtaining tail probability of Wilk's likelihood ratio test. *J. Multivariate Anal.*, **31**, 117–126.

[32] Tallis, G.M. and Light, R. (1968). The use of fractional moments for estimating the parameters of a mixed exponential distribution. *Technometrics*, **10**, 161–175.

[33] Wilks, S.S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471–494.

[34] Wilks, S.S. (1960). "Multidimensional Statistical Scatter". In *Contributions to Probability and Statistics*, I. Olkin et al., ed. Stanford University Press, Stanford, Calif., pp. 486–503.

[35] Wilks, S.S. (1962). *Mathematical Statistics*. John Wiley & Sons, New York.

[36] Wilks, S.S. (1967). "Multidimensional Statistical Scatter". In *Collected Papers: Contributions to Mathematical Statistics*, T.W. Anderson, ed. John Wiley & Sons, New York, pp. 597–614.

[37] Yeh, A.B., Lin, D.K.J., Zhou, H., and Venkataramani, C. (2003). A multivariate exponentially weighted moving average control chart for monitoring process variability. *J. Appl. Stat.*, **30**, 507–536.

[38] Yohai, V.J. and Garcia Ben, M.S. (1980). Canonical variables as optimal predictors *Ann. Stat.*, **8**, 865–869.

[39] Zuo, Y. and Serfling, R. (2000). Nonparametric multivariate notions of 'Scatter' and 'More Scattered' based on statistical depth functions. *J. Multivariate Anal.*, **75**, 62–78.

Queries in Article ess6053

Q1. Please clarify if this citation should be 24 or 25.

Q2. References 40 to 44 have not been provided in the reference list. Please provide the complete details.

Q3. Please clarify if this article has since been published. If so, please provide the complete details for this reference.

Q4. References 1, 7 and 17 have not been cited in the text. Please provide the place of citation for these references.

Q5. Please list out all the authors' names for References 23 and 34.

Q6. Please provide the page range for this reference.

Q7. Please provide the complete details for this reference.