A generalized ϕ-divergence for asymptotically multivariate normal models

Stefan Wegenkittl *

Abstract

Csiszár's [4] φ-divergence which was considered independently by Ali and Silvey [1] gives a goodness-of-fit statistic for multinomial distributed data. We define a generalized ϕ-divergence which unifies the φ-divergence approach with that of Rao [25] and derive weak convergence to a chi-square distribution under the assumpotion of asymptotically multivariate normal distributed data vectors. As an example we discuss the application to the frequency count in Markov chains and thereby give a goodness-of-fit test for observations from dependend processes with finite memory.
AMS classification: 62H10 Distribution of statistics, 62H15 Hypothesis testing, 62M02 Markov processes: hypothesis testing (Inference from stochastic processes), 62E20 Asymptotic distribution theory.

1 Introduction

Let $m \geq 2$ and denote by $D_m = \{\mathbf{P} = (P_1, \ldots, P_m) \in \mathbb{R}^m, P_i \geq 0, \sum_{i=1}^{m} P_i = 1\}$ the set of discrete probability distributions and by $D_m^\circ = \{\mathbf{P} \in D_m, P_i > 0\}$ the subset of non-degenerate probability distributions of D_m. Fix a vector $\mathbf{P} \in D_m^\circ$ and consider a sequence $(\hat{\mathbf{P}}^{(n)})_{n\in\mathbb{N}}$ of m-variate random vectors such that $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$ converges in distribution to a multivariate normal distribution with zero mean and covariance matrix Σ with rank $R(\Sigma)$ as $n \to \infty$. An important example is the class of multinomial models where $n\hat{\mathbf{P}}^{(n)}$ is distributed multinomial with parameters n and $\mathbf{P}$, $n\hat{\mathbf{P}}^{(n)} \sim \mathcal{MN}(n, \mathbf{P})$: here,

*Research supported by the Austrian Science Foundation (FWF), project no. P11143-MAT and FSP "Number-theoretic Algorithms and Their Applications"

$\Sigma = \Sigma_{\mathbf{P}}$, where $\Sigma_{\mathbf{P}}$ denotes the covariance matrix of a $\mathcal{MN}(1, \mathbf{P})$ distributed vector. Define the statistic

$$\mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) := n \sum_{i=1}^{m} \frac{(\hat{P}_i^{(n)} - P_i)^2}{P_i} = n \sum_{i=1}^{m} P_i \left(\frac{\hat{P}_i^{(n)}}{P_i} - 1 \right)^2 . \tag{1}$$

The famous Theorem of Pearson [24] states that under the condition $n\hat{\mathbf{P}}^{(n)} \sim \mathcal{MN}(n, \mathbf{P})$ (the multinomial case), $\mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ converges to a chi-square distribution with $m-1$ degrees of freedom in distribution as n approaches infinity, $\mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\mathrm{d}} \chi^2_{m-1}$. This statistic plays a central role in the field of hypothesis testing where $\hat{\mathbf{P}}^{(n)}$ is the relative frequency of the number of occurrences of each possible outcome of a discrete i.i.d. process.

Our aim in this paper is a generalization of Pearson's statistic (1) towards the general multivariate normal case $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$ which arises in the context of counter vectors arising from sampling Markov processes, and towards a more general class of loss functions of the likelihood-ratios $\hat{P}_i^{(n)}/P_i$.

The middle term in (1) can be interpreted as a quadratic form in a diagonal matrix $diag(P_1{}^{-1}, \ldots, P_m{}^{-1})$. In Section 2 we recall the theory of distributions of quadratic forms in generalized inverses which allows to extend the statistic $\mathcal{X}^2$ to suit the general case of asymptotically multivariate normal sequences $(\hat{\mathbf{P}}^{(n)})_{n \in \mathbb{N}}$. In addition, the right hand term in (1) can be interpreted as the expected value of the squared loss with respect to the likelihood ratios $\hat{P}_i^{(n)}/P_i$. This suggests replacing the loss function $\varphi(u) = (u-1)^2$ by a function from a class of convex functions and yields the so-called φ-divergence which we recall in Section 2, too.

In Section 3 we join both aspects by defining a generalized ϕ-divergence depending on an $m \times m$ matrix $\overline{\Sigma}$ and a function ϕ: $[0, \infty)^2 \to (-\infty, \infty]$ and prove weak convergence to a chi-square distribution in the general multivariate normal case. We discuss a parameterization scheme which includes the aforementioned quadratic forms and φ-divergences as special cases in Section 4. Section 5 shows an application to the frequency count in Markov chains, the Appendix contains technical details. Our contribution which is part of the author's PhD thesis, is related to the work of [30, 18] concerning generalizations of divergence statistics, and to [16, 17, 3, 35] concerning tests for multinomial data and asymptotic distributions.

2 Quadratic Forms and φ-Divergences

Let the statistic

$$\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) := n \sum_{i,j=1}^{m} \overline{\Sigma}_{ij} \left(\hat{P}_i^{(n)} - P_i\right) \left(\hat{P}_j^{(n)} - P_j\right) \tag{2}$$

be the quadratic form in the $m \times m$ matrix $\overline{\Sigma} = (\overline{\Sigma}_{ij})$. It is well known (see e.g. [25, Theorem 9.2.2] and apply the Mapping Theorem in $\mathbb{R}^m$) that if $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\text{d}} \mathcal{N}(\emptyset, \Sigma)$ and if $\overline{\Sigma}$ is a generalized inverse of Σ (i.e. $\Sigma\overline{\Sigma}\Sigma = \Sigma$) then the statistic $\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)},\mathbf{P})$ is asymptotically distributed chi-square

$$\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) \xrightarrow{\text{d}} \chi^2_{R(\Sigma)} \tag{3}$$

with $R(\Sigma)$ degrees of freedom. The term weak inverse is also used for $\overline{\Sigma}$. Further reading on the theory of generalized inverses can be found in [6, 20]. The existence of a generalized inverse $\overline{\Sigma}$ of the covariance matrix Σ is guaranteed by the spectral decomposition $U'\Sigma U = diag(\lambda_1, \ldots, \lambda_m)$ with an orthogonal matrix U and m (not necessarily distinct) eigenvalues λ_i, $1 \le i \le m$, of Σ: putting $\lambda_i^- = \lambda_i^{-1}$, if $\lambda_i \neq 0$, and $\lambda_i^- = 0$ otherwise, and, finally, $D^- = diag(\lambda_1^-, \ldots, \lambda_m^-)$ we may define $\overline{\Sigma} = UD^-U'$, which trivially fulfills $\Sigma\overline{\Sigma}\Sigma = \Sigma$.

If $\Sigma = \Sigma_{\mathbf{P}}$ for some $\mathbf{P} = (P_1, \ldots, P_m) \in D_m^\circ$, we may choose a generalized inverse $\Sigma_{\mathbf{P}}^-$ of the form $\Sigma_{\mathbf{P}}^- = diag({P_1}^{-1}, \ldots, {P_m}^{-1})$, so that (2) equals the middle term in (1). The Theorem of Pearson thus follows from (3) by the Central Limit Theorem for a normalized sequence of multinomial random vectors.

Now we turn to the right hand term in (1). In 1963, Csiszár [4] defined a measure for the deviation of two probability densities $\hat{\mathbf{P}}$ and $\mathbf{P}$, which we will again assume to be discrete. His so-called φ-divergence was also introduced independently by Ali and Silvey [1] in 1966. Regarding a multinomial test setting, the original measure has to be scaled in order to obtain convergence in distribution to a chi-square distribution as n goes to infinity. Let $\varphi : [0, \infty) \to (-\infty, \infty]$ be a function with continuous second derivative on some non-empty interval $I_\delta = (1 - \delta, 1 + \delta) \subset [0, \infty)$, such that $\varphi(1) = \varphi'(1) = 0$ and $\varphi''(1) \neq 0$, and let φ be arbitrary outside of I_δ. Define the φ-divergence-statistic of $\hat{\mathbf{P}}^{(n)}$ and $\mathbf{P}$ by

$$I_\varphi(\hat{\mathbf{P}}^{(n)},\mathbf{P}) := \frac{2n}{\varphi''(1)} \sum_{i=1}^{m} P_i \, \varphi\left(\frac{\hat{P}_i^{(n)}}{P_i}\right). \tag{4}$$

The family of φ-divergences includes several well-known measures (in particular (1)) for the deviation of two probability distributions, some of which are:

1) the class φ_α introduced by Liese and Vajda [13],

$$\varphi_\alpha(u) = \begin{cases} u - 1 - \ln u & \alpha = 0 \\ \frac{\alpha u + 1 - \alpha - u^\alpha}{\alpha(1-\alpha)} & \alpha \in \mathbb{R} \setminus \{0, 1\} \\ 1 - u + u \ln u & \alpha = 1 \end{cases}$$

This class is also known under the name power-divergence in [28, 3], where the family is indexed by a parameter λ which equals $\alpha - 1$. For every member of the class $\varphi_\alpha(u)$, $\alpha \in \mathbb{R}$, we have $\varphi(1) = \varphi'(1) = 0$, and $\varphi''(1) = 1$. The following instances of φ_α play a major role in estimation and decision theory. For $\alpha = 2$ we get $\varphi_2(u) = \frac{1}{2}(u-1)^2$ which corresponds to the Pearson $\mathcal{X}^2$. Choosing $\alpha = -1$ on the other hand yields Neyman's [21] modified chi-square statistic $NM^2 = n \sum\limits_{i=1}^{m} \frac{(\hat{P}_i^{(n)} - P_i)^2}{\hat{P}_i^{(n)}}$, which is equivalent to exchanging $\hat{\mathbf{P}}^{(n)}$ and $\mathbf{P}$ in $\mathcal{X}^2$. For $\alpha = 1$ we get the I-Divergence of Kullback-Leibler [10], $G^2 = 2n \sum\limits_{i=1}^{m} \hat{P}_i^{(n)} \ln(\frac{\hat{P}_i^{(n)}}{P_i})$, which is also called log-likelihood ratio statistic. In the equiprobable case $\mathbf{P} = (\frac{1}{m}, \ldots, \frac{1}{m})$, $G^2/2n$ is equal to $(\ln(m) - \ln(2) H(\hat{\mathbf{P}}^{(n)}))$, where $H(\hat{\mathbf{P}}^{(n)}) = -\sum\limits_{i=1}^{m} \hat{P}_i^{(n)} \log_2(\hat{P}_i^{(n)})$ is the so-called sample entropy. The case $\alpha = 0$ gives the modified log-likelihood ratio statistic considered by Kullback [9, 8], $GM^2 = 2n \sum\limits_{i=1}^{m} P_i \ln(\frac{P_i}{\hat{P}_i^{(n)}})$. Finally, setting $\alpha = 1/2$ yields $\varphi_{1/2}(u) = 2(\sqrt{u} - 1)^2$ and hence the square of the Hellinger-Distance [14], $F^2 = 4n \sum\limits_{i=1}^{m} (\sqrt{\hat{P}_i^{(n)}} - \sqrt{P_i})^2$, see also [11, Chapter 4, p. 46].

2) the class h_α defined by Matusita [14, 15] for $\alpha \in (0,1)$ and by Vajda [31] for $\alpha \in (1, \infty)$, see also Boekee [2],

$$h_\alpha(u) = \begin{cases} |u^\alpha - 1|^{1/\alpha} & \alpha \in (0, 1] \\ |u - 1|^\alpha & \alpha \in (1, \infty) \end{cases}$$

For $\alpha = 2$ and $\alpha = 1/2$, h_α corresponds to a multiple of φ_α. The standard theory for asymptotic distributions of φ-divergences works only in these cases, although every $\alpha \in (0, 1]$ allows to define a distance of probability distributions in terms of the corresponding φ-divergence.

3) the class ϕ_κ as defined in [7],

$$\phi_\kappa(u) = \frac{|u-1|^\kappa}{2(u+1)^{\kappa-1}}, \ \kappa \in [1, \infty)$$

For each measure in this class, the corresponding φ-divergence allows the definition of a distance of probability distributions. Asymptotic theory, however, is applicable only in the case $\kappa = 2$ which was introduced by Vincze [32] and also investigated in Le Cam's book [11, Chapter 4, p. 47]. For $\kappa = 2$ we get

$\phi_\kappa(1) = \phi'_\kappa(1) = 0$ and $\phi''_\kappa(1) = 1/2$.
4) the class f_p introduced by Österreicher and Vajda [23]: let $\mathbb{R}_+ = (0,\infty)$ and put

$$f_p(u) = \begin{cases} u\ln(u) - (1+u)\ln(1+u) + (1+u)\ln(2) & p = 1 \\ \frac{1}{1-1/p}\left[(1+u^p)^{1/p} - 2^{(1/p)-1}(1+u)\right] & p \in \mathbb{R}_+ \setminus \{1\} \\ |1-u|/2 & p = \infty \end{cases}$$

Similar to φ_α in the case $\alpha = 1$, f_1 can be represented by Shannon's entropy measure. Even more, this representation is not limited to the equiprobable case since for every $\mathbf{P} \in D_m^\circ$ we have $I_{f_1}^\circ(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \ln(2)(2H(\frac{\hat{\mathbf{P}}^{(n)}+\mathbf{P}}{2}) - [H(\hat{\mathbf{P}}^{(n)}) + H(\mathbf{P})])$. In the case $p = 1/2$, f_p yields the Hellinger divergence. The case $p = 2$ has a peculiar appeal from the geometric point of view, see [22]. For every $p \in (0,\infty]$, the corresponding φ-divergence allows to define a distance. The asymptotic theory works for $p \in (0,\infty)$, where $f_p(1) = f_p'(1) = 0$ and $f_p''(1) = p2^{1/p-2}$.

In the multinomial setting $n\hat{\mathbf{P}}^{(n)} \sim \mathcal{MN}(n,\mathbf{P})$, $\mathbf{P} \in D_m^\circ$, all the mentioned measures for which the asymptotic theory is applicable are stochastically equivalent in the limit since $I_\varphi(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ converges to a chi-square distribution with $m-1$ degrees of freedom, $I_\varphi(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\mathrm{d}} \chi^2_{m-1}$. This can be shown by representing φ as a Taylor series and thereby achieving a reduction to the Pearson case, see also [19]. Thus, all generalizations so far apply to the multinomial case. We refer the reader to [30] for an extensive treatment within the framework of $(\underline{h}, \underline{\phi})$-divergences, which extend I_φ to an even larger class of divergence measures.

3 A generalized ϕ-divergence

In an attempt to unify (2) and (4) in a single statistic, we define

$$I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) := n \sum_{i,j=1}^{m} c_{ij}\; P_i P_j\; \phi\left(\frac{\hat{P}_i^{(n)}}{P_i}, \frac{\hat{P}_j^{(n)}}{P_j}\right). \tag{5}$$

Here, $c_{ij} = c_{ij}(\mathbf{P}, \overline{\Sigma}, \phi) \in \mathbb{R}$ denote weights which depend on $\mathbf{P}$, $\overline{\Sigma}$, and ϕ. Furthermore, $\phi : [0,\infty)^2 \to (-\infty,\infty]$ is a real valued function which takes over the role played by φ in (4). Let us assume that ϕ has continuous partial derivatives up to order 2 on an open square I_δ^2 with $I_\delta = (1-\delta, 1+\delta)$ and $0 < \delta < 1$. For $(\tau_1, \tau_2) \in \{x,y\}^2$ we denote these derivatives by

$$\phi_{\tau_1}(a,b) = \frac{\partial\phi(x,y)}{\partial\tau_1}_{|(x,y)=(a,b)}, \text{ and } \phi_{\tau_1\tau_2}(a,b) = \frac{\partial^2\phi(x,y)}{\partial\tau_1\partial\tau_2}_{|(x,y)=(a,b)},$$

and abbreviate $\phi_{\tau_1}(1,1)$ by ϕ_{τ_1}, and $\phi_{\tau_1\tau_2}(1,1)$ by $\phi_{\tau_1\tau_2}$. In analogy to the conditions on φ in I_φ we further assume that $\phi(1,1)=\phi_x=\phi_y=0$. Note that no restrictions are imposed on ϕ outside the open square I_δ^2.

To derive the asymptotic distribution of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P})$ as $n\to\infty$, we represent ϕ for $(x,y)\in I_\delta^2$ by a Taylor series:

$$\begin{aligned}\phi(x,y) \quad &= \quad \phi(1,1)+\phi_x(x-1)+\phi_y(y-1)+ \\ &+\frac{1}{2}\left\{\phi_{xx}(a,b)(x-1)^2+2\phi_{xy}(a,b)(x-1)(y-1)+\phi_{yy}(a,b)(y-1)^2\right\},\end{aligned}$$

with $a=1+\Delta(x-1)$ and $b=1+\Delta(y-1)$ for suitable $\Delta\in(0,1)$. Let $K_\delta(\mathbf{P})=\{\hat{\mathbf{P}}\in D_m\ :\ (\frac{\hat{P}_1}{P_1},\ldots,\frac{\hat{P}_m}{P_m})\in I_\delta^m\}$ and $Q_i^{(n)}:=\frac{\hat{P}_i^{(n)}}{P_i}-1$. If $\hat{\mathbf{P}}^{(n)}\in K_\delta(\mathbf{P})$, the Taylor representation can be used for every $\phi(\frac{\hat{P}_i^{(n)}}{P_i},\frac{\hat{P}_j^{(n)}}{P_j})$, $(i,j)\in\{1,\ldots,m\}^2$ and we let

$$a_{ij}^{(n)}=1+\Delta_{ij}^{(n)}Q_i^{(n)},\text{ and } b_{ij}^{(n)}=1+\Delta_{ij}^{(n)}Q_j^{(n)} \tag{6}$$

for appropriate $\Delta_{ij}^{(n)}\in(0,1)$ and put $\epsilon_{xx}^{(n)}(i,j)=\phi_{xx}(a_{ij},b_{ij})-\phi_{xx}$, $\epsilon_{xy}^{(n)}(i,j)=\phi_{xy}(a_{ij},b_{ij})-\phi_{xy}$, and $\epsilon_{yy}^{(n)}(i,j)=\phi_{yy}(a_{ij},b_{ij})-\phi_{yy}$. If $\hat{\mathbf{P}}^{(n)}\notin K_\delta(\mathbf{P})$, let $\epsilon_{xx}^{(n)}(i,j)=\epsilon_{xy}^{(n)}(i,j)=\epsilon_{yy}^{(n)}(i,j)=0$. In the sequel we omit the upper indices (n) of $Q_i^{(n)}$, $a_{ij}^{(n)}$, and $b_{ij}^{(n)}$.

Now let T be the Taylor expansion of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P})$ up to the second order terms,

$$T(\hat{\mathbf{P}}^{(n)},\mathbf{P})=\frac{n}{2}\sum_{i,j=1}^m c_{ij}\ P_iP_j\left\{\phi_{xx}Q_i^2+2\phi_{xy}Q_iQ_j+\phi_{yy}Q_j^2\right\}.$$

For $\hat{\mathbf{P}}^{(n)}\in K_\delta(\mathbf{P})$, (5) can be represented by $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P})=T(\hat{\mathbf{P}}^{(n)},\mathbf{P})+R(\hat{\mathbf{P}}^{(n)},\mathbf{P})$, with a remainder term

$$R(\hat{\mathbf{P}}^{(n)},\mathbf{P})=\frac{n}{2}\sum_{i,j=1}^m c_{ij}\ P_iP_j\ R_{ij}(\hat{\mathbf{P}}^{(n)},\mathbf{P}),$$

where $R_{ij}(\hat{\mathbf{P}}^{(n)},\mathbf{P})=\epsilon_{xx}^{(n)}(i,j)Q_i^2+2\epsilon_{xy}^{(n)}(i,j)Q_iQ_j+\epsilon_{yy}^{(n)}(i,j)Q_j^2$. We further define the remainder term

$$U(\hat{\mathbf{P}}^{(n)},\mathbf{P}) \quad = \quad \begin{cases} 0 & \text{for}\quad \hat{\mathbf{P}}^{(n)}\in K_\delta(\mathbf{P}) \\ I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P})-T(\hat{\mathbf{P}}^{(n)},\mathbf{P}) & \text{for}\quad \hat{\mathbf{P}}^{(n)}\notin K_\delta(\mathbf{P})\end{cases},$$

and get the following representation of (5):

$$I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) \quad = \quad T(\hat{\mathbf{P}}^{(n)},\mathbf{P})+R(\hat{\mathbf{P}}^{(n)},\mathbf{P})+U(\hat{\mathbf{P}}^{(n)},\mathbf{P}). \tag{7}$$

In the Appendix we show that under the assumption $\sqrt{n}\left(\hat{\mathbf{P}}^{(n)}-\mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$ both remainder terms $U(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ and $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ converge in probability to zero so that the asymptotic distribution of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ is equal to that of $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$. If furthermore $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ equals the quadratic form $\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ in a generalized inverse of Σ we actually know this asymptotic distribution which is a chi-square with the rank of Σ degrees of freedom. We thus will have to choose the values c_{ij} in such a way that $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$. Hence let $\Delta = \phi_{xy} + \frac{\phi_{xx}+\phi_{yy}}{2}$, and

$$T(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = n \sum_{i,j=1}^{m} d_{ij} \left(\hat{P}_i^{(n)} - P_i\right)\left(\hat{P}_j^{(n)} - P_j\right), \text{ where}$$

$$d_{ij} = \begin{cases} c_{ii}\Delta + \frac{1}{2P_i}\sum\limits_{k\neq i} P_k(\phi_{xx}c_{ik} + \phi_{yy}c_{ki}) & \text{for} \quad i=j \\ \phi_{xy}c_{ij} & \text{for} \quad i \neq j \end{cases}$$

To get the aforementioned equivalence we now solve $d_{ij} = \overline{\Sigma}_{ij}$ for all $(i,j) \in \{1,\ldots,m\}^2$ under the assumption that $\phi_{xy} \neq 0$ and that $\Delta \neq 0$. This, finally, yields

$$c_{ij}(\mathbf{P}, \Sigma, \phi) = \begin{cases} \frac{1}{\Delta}\left(\overline{\Sigma}_{ii} - \frac{1}{2P_i}\sum\limits_{k\neq i} P_k\left(\frac{\phi_{xx}}{\phi_{xy}}\overline{\Sigma}_{ik} + \frac{\phi_{yy}}{\phi_{xy}}\overline{\Sigma}_{ki}\right)\right) & \text{for} \quad i=j \\ \frac{\overline{\Sigma}_{ij}}{\phi_{xy}} & \text{for} \quad i\neq j \end{cases} \tag{8}$$

The conditions $\phi_{xy} \neq 0$ and $\phi_{xy} + \frac{\phi_{xx}+\phi_{yy}}{2} \neq 0$ are necessary in order to solve the equations for the coefficients c_{ij} in such a way that the quadratic form in the generalized inverse $\overline{\Sigma}$ can be implemented. If either of the condition is not satisfied, we get for some $\tilde{c}_i \in \mathbb{R}$

$$T(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \tilde{c}_0 \mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) + n\sum_{i=1}^{m} \tilde{c}_i \left(\hat{P}_i^{(n)} - P_i\right)^2.$$

Here, $\tilde{c}_0 = 0$ if $\phi_{xy} = 0$, and $\tilde{c}_0 = 1$ if $\phi_{xy} \neq 0$. The asymptotic distribution of $I_{\overline{\Sigma},\phi}$ ought not to be a central chi-square in this case which we will therefore not consider any further here.

We summarize our results in the following Theorem.

Theorem 3.1 (The Generalized ϕ-Divergence) *Assume, that the function $\phi : [0,\infty)^2 \to (-\infty,\infty]$ has continuous partial derivatives up to order 2 on an open square $I_\delta^2 \subset \mathbb{R}^2$ containing the point $(1,1)$ and that $\phi(1,1) = \phi_x = \phi_y = 0$, $\phi_{xy} \neq 0$, and $\Delta = \phi_{xy} + \frac{\phi_{xx}+\phi_{yy}}{2} \neq 0$. For a fixed $\mathbf{P} \in D_m^\circ$ assume a sequence of* *random vectors* $\hat{\mathbf{P}}^{(n)} \in D_m$, $n \in \mathbb{N}$, *which satisfies* $\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$, *with mean vector* $\emptyset$ *and covariance matrix* Σ *and let* $R(\Sigma)$ *denote the rank of* Σ, *and* $\overline{\Sigma}$ *a generalized inverse of* Σ. *Finally, let* c_{ij} *be defined as in (8) and let the generalized* ϕ*-divergence*

$$I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = n \sum_{i,j=1}^{m} c_{ij}\ P_i P_j\ \phi\left(\frac{\hat{P}_i^{(n)}}{P_i}, \frac{\hat{P}_j^{(n)}}{P_j}\right).$$

Then $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ *converges in distribution to a chi-square distribution* $\chi^2_{R(\Sigma)}$ *with* $R(\Sigma)$ *degrees of freedom,* $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\mathrm{d}} \chi^2_{R(\Sigma)}$.

Section 5 gives an example using $I_{\overline{\Sigma},\phi}$ for a problem where the asymptotic distribution of the standard φ-divergence (4) is not necessarily a central chi-square distribution.

4 The $\tilde{I}_{\overline{\Sigma},\varphi}$-Divergence

As we have seen in the previous section, a basic requirement on ϕ is that $\phi_{xy} \neq 0$. The calculation of the constants c_{ij} becomes significantly easier if, at the same time, $\phi_{xx} = \phi_{yy} = 0$. These conditions can be satisfied by letting

$$\phi^{\varphi}(x, y) := 2\varphi\left(\frac{x+y}{2}\right) - \frac{\varphi(x) + \varphi(y)}{2}, \tag{9}$$

where $\varphi : [0, \infty) \to (-\infty, \infty]$ is a function which obeys exactly the same conditions as we have assumed in connection with the definition of the ordinary φ-divergence (4). All the conditions on ϕ in Theorem 3.1 are satisfied by ϕ^{φ} and the constants c_{ij} now calculate to $c_{ij} = \frac{2}{\varphi''}\overline{\Sigma}_{ij}$ for all $(i, j) \in \{1, \ldots, m\}^2$ and we get the following theorem:

Theorem 4.1 (The $\tilde{I}_{\overline{\Sigma},\varphi}$-Divergence) *Let* $\varphi : [0, \infty) \to (-\infty, \infty]$ *be a function with continuous second derivative on some interval* $I_\delta = (1 - \delta, 1 + \delta) \subset [0, \infty)$, *for which* $\varphi(1) = \varphi'(1) = 0$ *and* $\varphi'' := \varphi''(1) \neq 0$, *and let* φ *be arbitrary outside of* I_δ. *Let* ϕ^{φ} *be defined as in (9) and set*

$$\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \frac{2n}{\varphi''} \sum_{i,j=1}^{m} \overline{\Sigma}_{ij} P_i P_j \phi^{\varphi}\left(\frac{\hat{P}_i^{(n)}}{P_i}, \frac{\hat{P}_j^{(n)}}{P_j}\right).$$

On the conditions that $\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$ with covariance matrix Σ of rank $R(\Sigma)$, and that $\overline{\Sigma}$ is a generalized inverse of Σ, this statistic is asymptotically distributed chi-square with $R(\Sigma)$ degrees of freedom, $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\mathrm{d}} \chi^2_{R(\Sigma)}$.

The class of $\tilde{I}_{\overline{\Sigma},\varphi}$-divergences is a subclass of the $I_{\overline{\Sigma},\phi}$-divergences. Remarkably, it is large enough to comprise the classic Pearson case (1), the quadratic forms in generalized inverses (2), and the standard φ-divergence (4):

Remark 4.2 (Backward compatibility) *Let φ, ϕ^{φ}, and $\tilde{I}_{\overline{\Sigma},\varphi}$ be defined as in Theorem 4.1. The statistic $\tilde{I}_{\overline{\Sigma},\varphi}$ comprises the following statistics as special cases:*

i) choosing $\varphi(u) = (u-1)^2$ we get $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, i.e. the new statistic generalizes the quadratic form in a generalized inverse,
ii) choosing $\overline{\Sigma} = \Sigma^{-}_{\mathbf{P}} = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$, which is a generalized inverse of the covariance matrix $\Sigma_{\mathbf{P}}$ of the multinomial distribution $\mathcal{MN}(1, \mathbf{P})$, yields $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = I_{\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, i.e. the new statistic generalizes the φ-divergence I_{φ}. It can be shown that $R(\Sigma_{\mathbf{P}}) = m-1$, consistently. Finally
iii) choosing both $\varphi(u) = (u-1)^2$, and $\overline{\Sigma} = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$ yields $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, i.e. backward compatibility to Pearson's statistic.

In a series of papers, C.R. Rao [26, 27] advocated the use of Hellinger distance in statistical analyses. An open problem is to do similar investigations for the new $I_{\overline{\Sigma},\phi}$-divergence measures and to show whether the power of a test can be optimized by choosing certain loss-functions with respect to the process the data is sampled from.

5 Application to Markov Chains

In the following we consider a finite chain $(S, \mathbb{P})$ with state space $S = \{1, \ldots, m\}$ and transition probabilities $\mathbb{P} = (p_{ij})_{(i,j)\in S^2}$ such that the process $(X_l)_{l\in\mathbb{N}}$, $X_l \in S$, fulfills for all $(i, j_{l-1}, \ldots, j_1) \in S^l$

$$P[X_l = i | X_{l-1} = j_{l-1}, \ldots, X_1 = j_1] = P[X_l = i | X_{l-1} = j_{l-1}] = p_{ij_{l-1}}.$$

We denote the elements of the n'th power of $\mathbb{P}$ by p_{ij}^n. Our aim is to show that $\tilde{I}_{\overline{\Sigma},\varphi}$ generalizes standard serial tests for independent sequences to tests on the frequency vector in ergodic chains $(S,\mathbb{P})$.

Let us assume that the chain $(S,\mathbb{P})$ is irreducible and aperiodic in the following. The finiteness of S implies positive recurrence and thus the existence of a stable distribution $\mathbf{P}$ which is unique by the irreducibility. Furthermore, ergodicity, i.e. $\lim_{n\to\infty} p_{ij}^n = P_j$, is guaranteed by aperiodicity. A Central Limit Theorem for the frequency count in such chains is e.g. given in [29, Theorem 42.VII]: Let $C_i^{(n)}$ denote the number of visits of the process (X_l) in state i during the first n steps, $C_i^{(n)} = \#\{1 \le l \le n : X_l = i\}$, and put the vectors $C^{(n)} = (C_1^{(n)}, \ldots, C_m^{(n)})$ and $\hat{\mathbf{P}}^{(n)} = \frac{1}{n} C^{(n)}$.

In order to represent asymptotic expectation and covariances we use the following abbreviations: denote the minor of the determinant $A(\lambda)$ which equals $(-1)^{i+j}$ times the determinant of the minor of $(\lambda I_m - A)_{ji}$ by $A_{ij}(\lambda)$, where $(\lambda I_m - A)_{ji}$ arises from deleting the j-th row and i-th column in $(\lambda I_m - A)$. Further denote the so-called principal minor of second order of the determinant $A(\lambda)$, namely the determinant of the principal minor of $(\lambda I_m - A)_{ij|ij}$, by $A_{ij|ij}(\lambda)$. For arbitrary $(i,j) \in S^2$ put

$$Q_{ij} = \frac{\mathbb{P}_{ij|ij}(1)}{\sum_{k\in S} \mathbb{P}_{kk}(1)}, \quad Q_i = \sum_{j\neq i} Q_{ij}, \text{ and } Q = \sum_{i\in S} Q_i,$$

and let the matrix $V = (v_{ij})_{(i,j)\in S^2}$ be defined by

$$v_{ii} = P_i(1-P_i) + 2P_i(P_i(1-\frac{Q}{2}) + Q_i - 1) \tag{10}$$

$$v_{ij} = -P_iP_j + P_i(P_j(1-\frac{Q}{2}) + Q_j) + P_j(P_i(1-\frac{Q}{2}) + Q_i) - Q_{ij} \tag{11}$$

for $(i,j) \in S^2, i \neq j$. From [29, Chapter 4] we summarize without proof the following

Lemma 5.1 (CLT for finite ergodic chains) *Let $(S,\mathbb{P})$ be finite, irreducible and aperiodic, and let V be defined as above, then the stable distribution $\mathbf{P} = (P_1,\ldots,P_m)$ can be written $P_i = \mathbb{P}_{ii}(1)(\sum_{k=1}^m \mathbb{P}_{kk}(1))^{-1}$, $i \in S$, and the sequence $\hat{\mathbf{P}}^{(n)}$ fulfills the Central Limit Theorem*

$$\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, V) \qquad \text{as } n \to \infty.$$

For a (large) sample size n and a given chain model $(S,\mathbb{P})$, the hypothesis H_0: $(X_l)_{1\le l\le n}$ is a sample path of $(S,\mathbb{P})$ can now be tested against the alternative hypothesis (H_1: $(X_l)_{1\le l\le n}$ is not a sample path of $(S,\mathbb{P})$), by applying $\tilde{I}_{\overline{\Sigma},\varphi}$ to the normalized counter vector $\hat{\mathbf{P}}^{(n)}$. The parameterization of the test statistic is done by the above formulas for $\mathbf{P}$ and V, where we choose $\overline{\Sigma}$ to be a generalized inverse of V.

Note the correspondence to the multinomial case where the counter vector $C^{(n)}$ is computed from an i.i.d. sequence (X_l) which can be modeled by so-called independent chains having a transition matrix $\mathbb{P}$ consisting of equal, strictly positive lines, $\mathbf{P}$, say. In such a case, (10) and (11) reduce to the variances and covariances of a multinomial distribution $\mathcal{MN}(1,\mathbf{P})$ with parameters 1 and $\mathbf{P}$, i.e. $v_{ii} = P_i(1-P_i)$, and $v_{ij} = -P_iP_j$ for $j \neq i$. From Remark 4.2 (ii) we know that $\tilde{I}_{\overline{\Sigma},\varphi}$ becomes I_φ for the appropriate generalized inverse $\Sigma^-_{\mathbf{P}}$, so that $\tilde{I}_{\overline{\Sigma},\varphi}$ may be seen as generalization of serial (i.e. based on counting-) tests on independent processes to processes with finite memory. The additional terms in (10) and (11) are corrections to the variances and covariances which arise from the correlation within ergodic but not necessarily independent chains.

Example 5.2 *Consider the Markov chain defined by the state space* $S=\{1,2\}$ *and transition matrix* $\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 \\ p & q \end{pmatrix}$, $\quad p+q=1, \quad 0<p<1$ *The stable distribution is* $\mathbf{P} = (p/(1/2+p), 1/(1+2p))$. *For* $p=1/2$ *we get a model for the fair memoryless coin; in this case,* $C^{(n)}$ *is distributed multinomial with parameters* n *and* $(1/2,1/2)$, *and with* $\hat{\mathbf{P}}^{(n)} = \frac{1}{n}C^{(n)}$, *any of the statistics (4) is asympotically distributed chi-square as* n *goes to infinity.*

Now consider the case $p=1/4$, *so that* $\mathbf{P}=(1/3,2/3)$. *The asymptotic covariance matrix is easily computed to* $V = \begin{pmatrix} \frac{10}{27} & -\frac{10}{27} \\ -\frac{10}{27} & \frac{10}{27} \end{pmatrix}$, *which is not given by a multinormal model* $\Sigma_{\mathbf{P}}$. *Also,* $diag({P_1}^{-1},{P_2}^{-1}) = diag(3,2/3)$ *is no generalized inverse of* V. *The asymptotic distribution of any of the statistics (4) thus needs not to be a central chi-square. It is easy to check that* $V^- := \begin{pmatrix} \frac{27}{40} & -\frac{27}{40} \\ -\frac{27}{40} & \frac{27}{40} \end{pmatrix}$ *fulfills* $VV^-V=V$ *so that letting* $\overline{\Sigma}=V^-$, *one gets the desired asymptotic chi-square distribution with 1 degree of freedom of any* $I_{\overline{\Sigma},\phi}$- *or* $\tilde{I}_{\overline{\Sigma},\varphi}$*-divergence measure. An example is given by the generalized I-divergence*

$$\tilde{I}_{V^-,\varphi}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) = 2n\sum_{i,j=1}^{2} v^-_{ij}P_iP_j\phi^\varphi\left(\frac{\hat{P}^{(n)}_i}{P_i},\frac{\hat{P}^{(n)}_j}{P_j}\right),$$

where $\varphi(u)=\varphi_1(u)=1-u+u\ln u$, *and* v^-_{ij} *are the elements of the matrix* V^-.

For elaborate examples concerning tests for uniform pseudorandom generators see [34, 33] (see [5, 12] for an introduction to such generators and tests for their randomness).

6 Appendix

Again, denote convergence in probability by $\xrightarrow{\mathrm{p}}$ and convergence in distribution by $\xrightarrow{\mathrm{d}}$. We show that under the assumptions of Theorem 3.1, both remainder terms, $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ and $U(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ converge in probability to zero so that the asymptotic distribution of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ is equal to that of $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$.

As to the second term, recall that $U(\hat{\mathbf{P}}, \mathbf{P}) \neq 0$ implies that $\hat{\mathbf{P}} \notin K_\delta(\mathbf{P})$. The asymptotic normality of $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P})$ implies the convergence in probability of $\hat{\mathbf{P}}^{(n)}$ to $\mathbf{P}$ with respect to the maximum norm in $\mathbb{R}^m$ and consequently the convergence of $U(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\mathrm{p}} 0$ as n goes to infinity. As to the first remainder term, we have $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \frac{n}{2} \sum_{i,j=1}^{m} c_{ij}\ P_i P_j\ R_{ij}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, where $R_{ij}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \epsilon_{xx}^{(n)}(i,j) Q_i^2 + 2\epsilon_{xy}^{(n)}(i,j) Q_i Q_j + \epsilon_{yy}^{(n)}(i,j) Q_j^2$. Discarding constants we get $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\mathrm{p}} 0$ provided that (i), (ii), and (iii) hold for all $i, j \in \{1, \ldots, m\}$:

(i) $\epsilon_{xx}^{(n)}(i,j)\ \ n\left(\hat{P}_i^{(n)} - P_i\right)^2 \xrightarrow{\mathrm{p}} 0$,

(ii) $\epsilon_{xy}^{(n)}(i,j)\ \ n\left(\hat{P}_i^{(n)} - P_i\right)\left(\hat{P}_j^{(n)} - P_j\right) \xrightarrow{\mathrm{p}} 0$, and

(iii) $\epsilon_{yy}^{(n)}(i,j)\ \ n\left(\hat{P}_j^{(n)} - P_j\right)^2 \xrightarrow{\mathrm{p}} 0$.

We will prove these by showing that each $\epsilon^{(n)}(i,j)$ converges in probability to zero, and that the remaining terms are tight sequences of random variables, so that the products (i)–(iii) themselves converge.

So, let i, j, and n be arbitrary but fixed and let $\epsilon > 0$. By the continuity of $\phi_{xy}(\cdot,\cdot)$ at $(1,1)$ there exist $\alpha = \alpha(\epsilon) > 0$ and $\beta = \beta(\epsilon) > 0$ such that $\max\{\alpha, \beta\} \leq \delta$, where $\delta > 0$ is given by the assumptions in Theorem 3.1, and

$$|a_{ij}^{(n)} - 1| < \alpha \wedge |b_{ij}^{(n)} - 1| < \beta \Rightarrow \left|\epsilon_{xy}^{(n)}(i,j)\right| < \epsilon. \tag{12}$$

Note that α and β do not depend on n. From (6) we have $|a_{ij}^{(n)} - 1| \leq |\frac{\hat{P}_i^{(n)}}{P_i} - 1|$ and $|b_{ij}^{(n)} - 1| \leq |\frac{\hat{P}_j^{(n)}}{P_j} - 1|$ such that applying (12) and (6) we get

$$P\left[\left|\epsilon_{xy}^{(n)}(i,j)\right| < \epsilon\right] \quad \geq \quad P\left[|a_{ij}^{(n)} - 1| < \alpha \wedge |b_{ij}^{(n)} - 1| < \beta\right] \geq$$

$$\geq \quad P\left[|\frac{\hat{P}_i^{(n)}}{P_i} - 1| < \alpha \wedge |\frac{\hat{P}_j^{(n)}}{P_j} - 1| < \beta\right].$$

Now let n go to infinity. The asymptotic normality of $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P})$ implies $\frac{\hat{P}_i^{(n)}}{P_i} - 1 \xrightarrow{\mathrm{P}} 0$ and $\frac{\hat{P}_j^{(n)}}{P_j} - 1 \xrightarrow{\mathrm{P}} 0$. Thus,

$$P\left[|\frac{\hat{P}_i^{(n)}}{P_i} - 1| < \alpha \wedge |\frac{\hat{P}_j^{(n)}}{P_j} - 1| < \beta\right] \to 1$$

for all $\alpha, \beta > 0$. We thereby have shown that for arbitrary i and j and for every $\epsilon > 0$, $P[|\epsilon_{xy}^{(n)}(i,j)| < \epsilon] \to 1$ as $n \to \infty$. A similar calculation yields $\epsilon_{xx}^{(n)}(i,j) \xrightarrow{\mathrm{P}} 0$ and $\epsilon_{yy}^{(n)}(i,j) \xrightarrow{\mathrm{P}} 0$.

It remains to check the tightness of the remaining terms. For the sake of simplicity we let $I_n := \sqrt{\frac{n}{\Sigma_{ii}}}(\hat{P}_i^{(n)} - P_i)$ and $J_n := \sqrt{\frac{n}{\Sigma_{jj}}}(\hat{P}_j^{(n)} - P_j)$. Then clearly $I_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$, $J_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$. As for (i) and (iii), the fact that I_n^2 is asymptotically distributed chi-square yields the desired tightness of $n\left(\hat{P}_i^{(n)} - P_i\right)^2$. Regarding (ii), we immediately get tightness as a consequence of the asymptotic normality of the factors. ∎

Acknowledgment: The author wishes to thank his thesis' supervisor Ferdinand Österreicher from the Department of Mathematics, University of Salzburg, Austria, for the strong support and many hints for a concise presentation, and Peter Hellekalek from the same department, who is head of the author's supporting FWF project P-11143-MAT.

References

[1] M.S. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc.*, Ser **B**(28):131–142, 1966.

[2] D.E. Boekee. *A generalization of the Fisher information measure.* Delft University Press, Delft, 1977.

[3] N. Cressie and T. Read. Multinomial Goodness-of-fit Tests. *J. R. Statist. Soc. B*, **46**(3):440–464, 1984.

[4] I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, **8**:85–108, 1963.

[5] P. Hellekalek. On the assessment of random and quasi-random point sets. In P. Hellekalek and G. Larcher, editors, *Random and Quasi-Random Point Sets*,

volume 138 of *Lecture Notes in Statistics*, pages 49–108. Springer, New York, 1998.

[6] A. B. Israel and T. Greville. *Generalized Inverses: Theory and Applications.* Wiley Interscience Publications. Wiley and Sons, New York, 1974.

[7] P. Kafka, F. Österreicher, and I. Vincze. On powers of f-divergences defining a distance. *Studia Sci. Math. Hungar.*, 26:415–422, 1991.

[8] S. Kotz et al., editors. *Minimum discrimination information (MDI) estimation*, volume 5 of *Wiley Interscience Publication*, pages 527–529. John Wiley, 1985.

[9] S. Kullback. *Information Theory and Statistics.* John Wiley, New York, 1959.

[10] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[11] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer, New York, 1986.

[12] P. L'Ecuyer and P. Hellekalek. Random number generators: Selection criteria and testing. In P. Hellekalek and G. Larcher, editors, *Random and Quasi-Random Point Sets*, volume 138 of *Lecture Notes in Statistics*, pages 223–266. Springer, New York, 1998.

[13] F. Liese and I. Vajda. *Convex Statistical Distances.* Teubner-Texte zur Mathematik, Band *95*. Teubner, 1987.

[14] K. Matusita. Decision rules based on the distance for problems of fit. *Ann. Math. Statist.*, 26:631–640, 1955.

[15] K. Matusita. Distance and decision rules. *Ann. Inst. Statist. Math.*, 16:305–320, 1964.

[16] M. Menéndez, D. Morales, L. Pardo, and I. Vajda. Divergence-based estimation and testing of statistical models of classification. *Journal of Multivariate Analysis*, **54**(2):329–354, 1995.

[17] M. L. Menéndez, D. Morales, L. Pardo, and M. Salicrú. Asymptotic behaviour and statistical applications of divergence measures in multinomial populations: a unified study. *Statistical Papers*, 36:1–29, 1995.

[18] D. Morales, L. Pardo, M. Salicrú, and M. L. Menéndez. Asymptotic properties of divergence statistics in a stratified random sampling and its applications to test statistical hypotheses. *Journal of Statistical Planning and Inference*, **38**:201–222, 1994.

[19] D. Morales, L. Pardo, and I. Vajda. Asymptotic divergence of estimates of discrete distributions. *Journal of Statistical Planing and Inference*, **48**:347–369, 1995.

[20] M. Z. Nashed, editor. *Generalized Inverses and Applications*, New York, 1976. Academic Press.

[21] J. Neyman. Contributions to the theory of the χ^2 test. Proceedings of the First Berkley Symposium on Mathematical Statistics and Probability, 1949.

[22] F. Österreicher. The construction of least favourable distributions is traceable to a minimal perimeter problem. *Studia Sci. Math. Hungar.*, 17:341–351, 1982.

[23] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and its statistical applications. *Annals of the Institute of Statistical Mathematics*, 1999. to appear.

[24] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Magazine Series*, **50** (5):157–172, 1900.

[25] C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications.* Wiley Series in Probability and Mathematical Statistics. Wiley and Sons, 1971.

[26] C.R. Rao. The use of hellinger distance in graphical displays of contingency table data. In *Proc. 5-th Tartu Conference, Estonia: New Trends in Probabiility and Statistics, Vol.3: Multivariate Statistics and Matrices in Statistics*, pages 143–162, 1995.

[27] C.R. Rao. An alternative to correspondence analysis using hellinger distance. In *Proc. International Symposium on Contemporary Multivariate Analysis and Its Applications, Hong Kong*, pages A11–A29, 1997.

[28] T. Read and N. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer Series in Statistics. Springer Verlag, New York, 1988.

[29] V. Romanovsky. *Discrete Markov Chains.* Wolters-Noordhoff Publishing, Groningen, Netherlands, 1970.

[30] M. Salicrú, D. Morales, M. L. Menéndez, and L. Pardo. On the applications of divergence type measures in testing statistical hypotheses. *Journal of Multivariate Analysis*, **51**:372–391, 1994.

[31] I. Vajda. x^α-divergence and generalized Fischer's information. *Trans. 6th Prague Conference on Information Theory*, pages 873–886, 1973.

[32] I. Vincze. On the concept and measure of information contained in an observation. In J. Gani and V.F. Rohatgi, editors, *Contributions to Probability*, pages 207–214. Academic Press, 1981.

[33] S. Wegenkittl. Monkeys, gambling, and return times: Assessing pseudorandomness. In P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 625–631. IEEE Press, 1999.

[34] S. Wegenkittl and M. Matsumoto. Getting rid of correlations among pseudorandom numbers: Discarding versus tempering. *ACM Trans. Modeling and Computer Simulation*, **9**(3):282–294, 1999.

[35] K. Zografos, K. Ferentinos, and T. Papaioannou. ϕ-divergence statistics: Sampling properties and multinomial goodness of fit and divergence tests. *Commun. Statist. Theory MEth.*, 19(5):1785–1802, 1990.

Author's address

Dr. Stefan Wegenkittl Institut für Mathematik, Universität Salzburg, Hellbrunnerstr. 34, A-5020 Salzburg, Austria
e-mail : Stefan@u24.at