Biometrika (2022), **109**, 3, *pp.* 605–609 https://doi.org/10.1093/biomet/asac034

Rejoinder: 'Multi-scale Fisher's independence test for multivariate dependence'

BY S. GORSKY

Department of Mathematics and Statistics, University of Massachusetts Amherst, 710 N. Pleasant Street, Amherst, Massachusetts 01003, U.S.A.
sgorsky@umass.edu

AND L. MA

Department of Statistical Science, Duke University, Box 90251, Durham, North Carolina 27708, U.S.A.
li.ma@duke.edu

1. INTRODUCTION

We wish to start by expressing our gratitude to the Editorial Board for giving us the opportunity to discuss our article, and to all of the discussants for their very insightful, thought-provoking contributions. These discussions provide a more comprehensive review of the vast relevant literature along with more in-depth views on several popular approaches for multivariate dependency testing, including computationally efficient approximation to kernel methods (Schrab et al., 2022), the k-nearest-neighbour-based mutual information test (Berrett, 2022) and the binary-expansion-test-based framework (Lee et al., 2022a). The discussions also complement our paper with additional numerical experiments that shed light on the statistical and computational strengths and weaknesses of the various approaches, and suggest interesting future research directions such as testing functional dependency (Lee et al., 2022a).

We focus our response on three important practical considerations raised in the discussions: (i) the ability to characterize the dependency structure, (ii) the statistical power of the multi-scale Fisher's independence test, MULTIFIT, on various dependencies and (iii) the connection between MULTIFIT and other methods.

2. A FEATURE SELECTION PERSPECTIVE ON DEPENDENCY LEARNING

As Berrett (2022) pointed out, the existing literature on multivariate dependency testing has mostly focused on formulating the decision problem as testing a single composite null hypothesis of independence, without much attention given to learning the nature of the dependency. By dividing the dependency structures into a collection of log odds ratios defined on various 2×2 tables, MULTIFIT essentially treats the inference on multivariate dependency from a perspective of feature selection through multiple testing control. By selecting the important tables with strong empirical evidence of a log odds ratio away from 0, this perspective allows the procedure to identify and summarize the structure of the underlying dependency.

Many high-dimensional statistical inference problems including regression modelling have long been treated from the feature selection perspective by the statistical community. Penalization, shrinkage and various multiple testing strategies have drawn much research effort in these contexts. After all, in high-dimensional applications, the practitioners are often more interested in identifying important structures in the data than in verifying if there exists any interesting structure at all. We believe the same is true in applied contexts of multivariate dependency testing involving a large number of variables. In fact, as the dimensionality of random vectors grows, it is often harder to find pairs of variables that are strictly

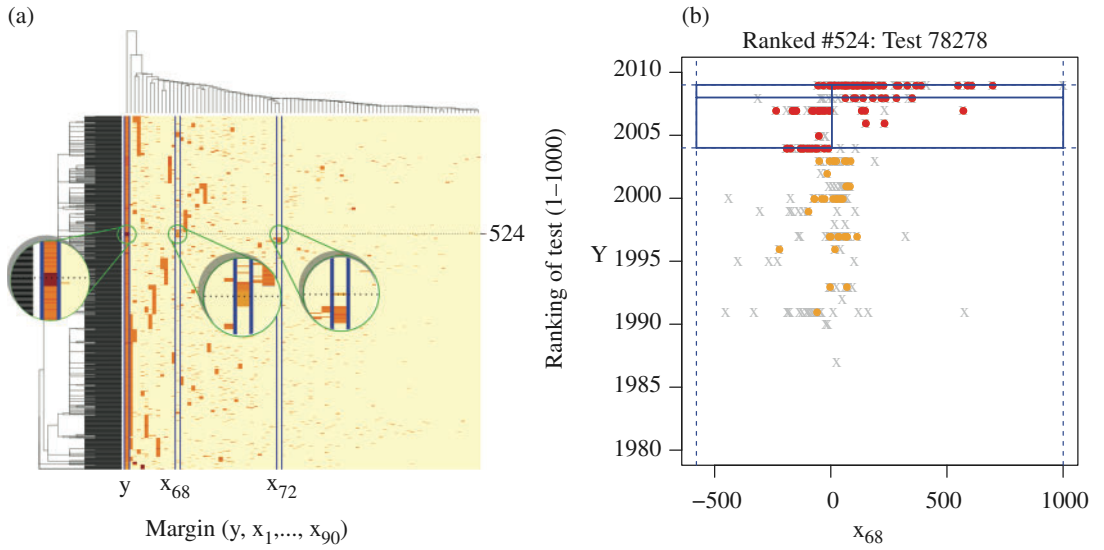

Fig. 1. Million Song Dataset. (a) A heatmap showing the margins involved in the dependencies identified by MULTIFIT in the 1000 2×2 tables with the most significant p-values. The darkness of the cells corresponds to the number of times a margin was divided to give rise to the corresponding 2×2 table. The x margins are ordered in decreasing level of importance, in terms of the number of times they are divided in the top 1000 tables. The horizontal dotted line highlights a test (ranked 524) and the vertical solid lines mark two relevant covariates x_{68} and x_{72} involved in that test. (b) Visualization of the table with the 524th most significant p-value. The two margins shown are y and x_{68}. This test is conditioned on $-727.5 \leqslant x_{72} < 35.91$. There is a notable shift to higher values of x_{68} when the year of release is in 2008–09 compared to 2004–07.

independent than otherwise, and rejecting the null hypothesis of independence thus becomes a matter of sample size without much scientific significance.

For example, when analysing the Million Song Dataset (Bertin-Mahieux et al., 2011), the practitioner probably has little doubt in expecting that some musical features are associated with the year of release of a song. The question of interest is which features those are and how they are dependent. To investigate this, we ran MULTIFIT on a subset of 250 observations from the Million Song Data with parameters $p^* = 10^{-10}$ and $R^* = 2$. The heatmap in Fig. 1(a) illustrates the relevance of the features, each corresponding to a column, in terms of the frequencies of their division in forming the top 1000 2×2 tables, each corresponding to a row, with the most extreme p-values under MULTIFIT. Figure 1(b) visualizes one of these top 2×2 tables (ranked 524), which involves two x margins, x_{68} and x_{72}.

3. STATISTICAL POWER AGAINST VARIOUS ALTERNATIVES

Berrett (2022) and Lee et al. (2022a) both pointed to the important fact that no test can be uniformly powerful against all fixed alternatives (Zhang, 2019). The implication is that, when applying dependency tests, the practitioner should consider carefully the relevance and appropriateness of the types of dependencies on which various tests are effective according to the application at hand.

The discussants shared insights and provided numerical evidence on the types of alternatives against which MULTIFIT may or may not be powerful compared to other methods. Berrett (2022) in particular commented, based on a sufficient condition to ensure large-sample consistency, that MULTIFIT is most powerful against large-scale dependencies that reside in coarse resolutions. On the other hand, Lee et al. (2022a) noted empirical evidence showing that MULTIFIT actually outperforms other approaches on alternatives that involve highly local dependencies.

To resolve this seeming paradox, we note that the effectiveness of the coarse-to-fine adaptive screening strategy employed by MULTIFIT hinges on the key assumption that the underlying dependency structure,

whether local or not, incurs some deviation away from independence on coarse scales. It is important to recognize that this is different from assuming that the dependency structure mainly resides in coarse scales and thus is of a global nature. In fact, the dependency can be mostly local, but 'spills over' into coarser resolutions to some degree. Such a spillover can be rather weak, incurring log odds ratios only slightly away from 0 in coarse scales, as the empirical evidence in coarser scales only needs to be strong enough to lead the procedure to zoom into the relevant regions in the sample space and select the relevant cuboids as 'suspects' for future tests at finer scales, where eventually convincing statistical evidence may arise on small tables with stronger dependencies. In other words, the p-value threshold for identifying candidate cuboids for further testing in finer resolutions is much less stringent than what is needed to declare statistical significance against the null. Therefore, as observed by Lee et al. (2022a), MULTIFIT can still have good power against highly localized dependencies, provided that they have some spillover into coarser scales, which is what Berrett (2022) indicated.

As shown by Berrett (2022) and Schrab et al. (2022), it is not too difficult to construe simulation scenarios in which there are no spillover effects whatsoever. For example, in the sinusoidal scenario considered by the discussants, the dependency can be periodic with frequencies perfectly matching that of the partition grid lines that MULTIFIT uses to construct the 2×2 tables, thereby perfectly dissipating the spillover into the first scale. The power of MULTIFIT suffers unsurprisingly in these cases. While we acknowledge that it is possible in some settings to encounter such data that have almost perfectly localized dependency without any spillover into coarse resolutions, we believe that in a vast range of real-world applications involving natural data generative processes, the spillover assumption is reasonable.

As indicated above, whether a dependency structure induces spillover into coarse resolutions depends on the partition grid along which the cuboids are constructed. For simplicity of exposition, Gorsky & Ma (2022) described a partition grid that always divides exactly in the middle of a margin when creating the 2×2 tables. Allowing the partition grid to instead deviate away from these perfectly centred dyadic lines can often restore the spillover phenomenon. For example, in the sinusoidal example, if we adopt a partition grid along the corresponding dyadic quantiles of beta(1, 1.3), instead of those of beta(1, 1), which corresponds to always dividing in the middle, then the power of MULTIFIT can be improved, as shown in Fig. 2. Such a perturbation on the partition grid can be readily accomplished by applying a cumulative distribution function transform of beta(1, 1.3) to some or all margins of the observations, possibly after marginal rank transforms, before applying MULTIFIT. We remark that the finite-sample validity of MULTIFIT is preserved under such perturbations since they can be viewed as marginal transforms of the data.

4. CONNECTION TO OTHER METHODS FOR MEASURING DEPENDENCY

MULTIFIT and its univariate predecessor Fisher exact scanning (Ma & Mao, 2019) bear resemblance to multi-resolution methods such as wavelet analyses. In essence, the collection of local odds ratios defined on cuboids at different scales and locations provide a multi-scale decomposition of the underlying copula structure. In contrast, while binary-expansion-test, BET-based, methods (Zhang, 2019; Lee et al., 2022b) also take a multi-scale approach, they treat each resolution holistically rather than in the location-specific fashion of MULTIFIT and Fisher exact scanning. Consequently, BET-based methods are generally more powerful when the underlying dependency spreads over large portions of the sample space, especially if the structure displays periodicity or long-range symmetry, which is confirmed in Lee et al. (2022a). Location-specific approaches such as MULTIFIT and Fisher exact scanning tend to be advantageous for alternatives involving localized dependencies, sharp changes and heterogeneity, or lack of periodicity and symmetry across the sample space.

We applaud the insightful comment from Lee et al. (2022a) that one could view MULTIFIT as a member of a general class of tests in the BET framework. However, we remark that this connection appears to exist only for the nonadaptive version of MULTIFIT involving exhaustive testing on all cuboids up to a maximum resolution. Both the computational practicality and statistical efficacy of MULTIFIT hinge on its adaptivity in the coarse-to-fine scanning based on the spillover assumption. It will be exciting to see

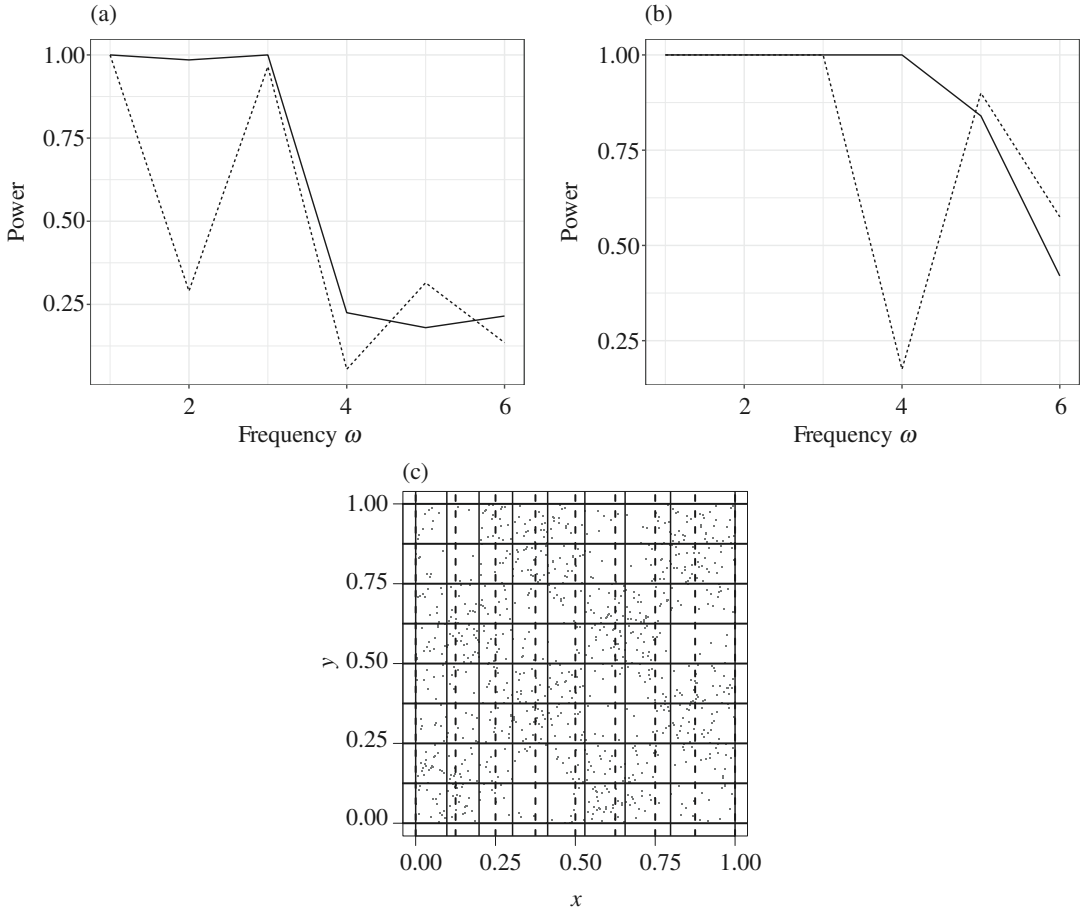

Fig. 2. Estimated power of MULTIFIT with (a) $R^* = 1$, (b) $R^* = 2$, and sample size $n = 8000$ under the sinusoidal scenario over a range of frequencies under the centred dyadic partition (dashed line) versus under a perturbed partition defined along the dyadic quantiles of beta(1, 1.3) (solid line). (c) An illustration of the centred dyadic partition (dashed line) and perturbed partition (solid line) on a sample after a marginal rank transform.

if this standard, adaptive version of MULTIFIT can also be viewed inside some general class, which may facilitate further theoretical analysis.

Finally, we conjecture that the multi-scale divide-and-conquer approach based on testing the local odds ratios that MULTIFIT and Fisher exact scanning employ might have a deeper connection to mutual information-based approaches (Kraskov et al., 2004; Kinney & Atwal, 2014; Berrett & Samworth, 2019). In § 4 of Ma & Mao (2019), an empirical analysis on a microbiome dataset from the American Gut Project (McDonald et al., 2015) showed that the ranking of two empirical estimates of the mutual information, namely the maximal information coefficient (Reshef et al., 2011) and the k-nearest-neighbour-based mutual information (Kraskov et al., 2004), are both strongly correlated with the ranking of the p-value on testing the null of independence under Fisher exact scanning. It would be interesting to explore this connection further.

REFERENCES

BERRETT, T. B. (2022). Discussion of 'Multi-scale Fisher's independence test for multivariate dependence'. *Biometrika* **109**, 589–92.

BERRETT, T. B. & SAMWORTH, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika* **106**, 547–66.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B. & Lamere, P. (2011). The million song dataset. In *Proc. 12th Int. Conf. Music Info. Retrieval (ISMIR 2011)*, University of Miami: Miami, Florida pp. 591–6.

Gorsky, S. & Ma, L. (2022). Multi-scale Fisher's independence test for multivariate dependence. *Biometrika*, **109**, 569–87.

Kinney, J. B. & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proc. Nat. Acad. Sci.* **111**, 3354–9.

Kraskov, A., Stögbauer, H. & Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E* **69**, 066138.

Lee, D., El-Zaatari, H., Kosorok, M. R., Li, X. & Zhang, K. (2022a). Discussion of 'Multi-scale Fisher's independence test for multivariate dependence'. *Biometrika* **109**, 593–6.

Lee, D., Zhang, K. & Kosorok, M. R. (2022b). The binary expansion randomized ensemble test (BERET). *Statist. Sinica*, to appear.

Ma, L. & Mao, J. (2019). Fisher exact scanning for dependency. *J. Am. Statist. Assoc.* **114**, 245–58.

McDonald, D., Hornig, M., Lozupone, C., Debelius, J., Gilbert, J. A. & Knight, R. (2015). Towards large-cohort comparative studies to define the factors influencing the gut microbial community structure of ASD patients. Microb. Ecol. Health Dis., https://doi.org/10.3402/mehd.v26.26555.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. & Sabeti, P. C. (2011). Detecting novel associations in large data sets. Science **334**, 1518–24.

Schrab, A., Jitkrittum, W., Szabó, Z., Sejdinovic, D. & Gretton, A. (2022). Discussion of 'Multi-scale Fisher's independence test for multivariate dependence'. *Biometrika* **109**, 597–603.

Zhang, K. (2019). BET on independence. *J. Am. Statist. Assoc.* **114**, 1620–37.

[*Received on* 30 *May* 2022. *Editorial decision on* 6 *June* 2022]