Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

# Kernels, Degrees of Freedom, and Power Properties of Quadratic Distance Goodness-of-Fit Tests

Bruce G. LINDSAY, Marianthi MARKATOU, and Surajit RAY

In this article, we study the power properties of quadratic-distance-based goodness-of-fit tests. First, we introduce the concept of a *root kernel* and discuss the considerations that enter the selection of this kernel. We derive an easy to use normal approximation to the power of quadratic distance goodness-of-fit tests and base the construction of a *noncentrality index,* an analogue of the traditional noncentrality parameter, on it. This leads to a method akin to the Neyman-Pearson lemma for constructing optimal kernels for specific alternatives. We then introduce a *midpower analysis* as a device for choosing optimal degrees of freedom for a family of alternatives of interest. Finally, we introduce a new diffusion kernel, called the *Pearson-normal kernel*, and study the extent to which the normal approximation to the power of tests based on this kernel is valid. Supplementary materials for this article are available online.

KEY WORDS: Big data; High-dimensional testing; Midpower analysis; Optimal kernel construction; Pearson-normal kernel; Power lemma.

## 1. INTRODUCTION

There is a long literature related to the use of distance measures in statistics. These measures have been used to construct estimators via the minimum distance principle and to construct measures of goodness of fit for statistical models. Our interest in this article is in distances with a relatively simple quadratic structure that depends on the choice of a nonnegative definite kernel $K(s, t)$ on the sample space.

We will focus here on the goodness-of-fit aspect, and in particular on tools for studying power. We have a number of important new achievements to report, both theoretical and practical. On the theoretical side, we first show how the concept of *root kernel* leads to tools for building kernels with targeted power properties (see Section 2). Second, we show how the concept of *the surrogate power function* can greatly simplify the determination of power for these tests (see Section 3). Also in 3 we show how a simple summary of the surrogate power function, the *noncentrality index*, can be used to select kernels optimal for power. This results in a *midpower lemma* akin to the Neyman-Pearson lemma, which enables one to build kernels for specific problem. Next, this index will be used in Section 4 in a fundamentally new *midpower analysis,* as a way to select tuning parameters in kernel distances. This analysis replaces a possibly very difficult local-alternatives type of analysis. In this section, we also show how one can build a new extension of the normal kernel that mimics the Pearson chi-square kernel.

On the practical side, we use our tools to carefully study our main example, testing for multivariate normality in high dimensions. In Section 5, we study the structure of two kernels in high dimensions. Both kernels have bandwidths that need to be tuned. In Section 6 we use our surrogate power analysis, along with simulation, to show that there exist bandwidths that result in testing methods that work well in high-dimensional problems, and that a careful study of the sensitivity of the test will reveal the interaction between sample size and dimension in using such procedures.

We start with a formal definition. Let $\mathcal{X}$ be a sample space. The building block of a statistical distance is the function $K(s, t)$, a bounded, symmetric, nonnegative definite (NND) kernel defined on $\mathcal{X} \times \mathcal{X}$. (As we shall see soon, the construction of NND kernels is not hard.) We let $G$ be a null distribution whose fit we wish to assess.

Note that we allow the kernel $K$ to possibly depend on the null $G$, so that technically it should be written as $K_G$, although we will usually use the short form $K$ for conciseness.

*Definition 1.* Given a NND kernel function $K(s, t)$, the $K$-*based quadratic distance* between two probability measures F and G is defined as $\mathbb{D}_K(F, G) = \iint K(s, t)d(F - G)(s)d(F - G)(t)$.

The empirical distance $\mathbb{D}_K(\hat{F}, G)$, where $\hat{F}$ is a nonparametric estimator of the true $F$, is a goodness-of-fit measure. Note that we will hereafter use $\mathbb{D}$ instead of $\mathbb{D}_K$. In this setting it is also possible to construct kernels $K$ suitable to a variety of data settings, including multivariate problems and a mixture of discrete and continuous spaces. The simple structure of these distances makes it easy to estimate and easy to analyze. Lindsay et al. (2008), hereafter written as LMR (2008), clarified this structure and provided new theory for goodness-of-fit testing based on quadratic distances. The class of quadratic distance tests is central to goodness of fit, encompassing tests based on characteristic functions, density estimation, and the chi-squared tests. In the parametric world, score tests are of this class. The class also provides quadratic approximations to many other tests, such as those based on likelihood ratios.

Bruce G. Lindsay is Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *bgl@psu.edu*). Marianthi Markatou is Professor at the Departments of Biostatistics and Biomedical Informatics, SUNY at Buffalo, 3435 Main Street, 726 Kimball Hall, School of Public Health and Health Professions, and School of Medicine, Buffalo, NY 14260 (E-mail: *markatou@buffalo.edu*). Surajit Ray is Senior Lecturer, School of Mathematics and Statistics, University of Glasgow, Glasgow, G128QQ (E-mail: *Surajit.Ray@glasgow.ac.uk*). This research was supported by National Science Foundation grants ATM-0934739, DMS-04-05637, and DMS-05-04957. We thank the reviewers and AE for the suggestions and comments as they have improved our article.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

**Journal of the American Statistical Association**
**March 2014, Vol. 109, No. 505, Theory and Methods**
**DOI: 10.1080/01621459.2013.836972**

This recent article also introduced the idea of the spectral degrees of freedom (DOF) of the kernel $K$. This single-number summary was shown to provide a useful summary of the approximate chi-square distribution of the test under the null. The DOF are also relevant to power. Intuitively, tests with low DOF are more focused on a limited set of alternatives, and so have high power there, but would have weak power elsewhere. On the other hand, one would expect a test with high DOF to be more omnibus, with widespread but low power.

However, in the settings we consider here, the DOF are infinitely tuneable. We would like to better know how to make a rational decision about DOF. We will argue that the noncentrality index

$$\mathrm{NCI}(F) = \frac{\mathbb{D}(F, G)}{\sigma_n}$$

provides a useful one number summary of the power characteristics to be used in such analyses. Here $G$ is the null hypothesis, $F$ is the alternative, and $\sigma_n^2$ is the asymptotic variance of $\mathbb{D}(\hat{F}, G)$ under the null. Note that $G$ is viewed as a fixed quantity throughout this article and so deleted from the notation when it simplifies expressions.

This article cannot claim to be a final set of recommendations on the choice of the kernel $K$ or the DOF. In the end, the choice of distance kernel $K$ is a matter of design, with such important design factors as the data type, the dimension of the data, the ability to do explicit calculations, and the alternatives that are of highest interest. We will offer some further guidance in the discussion.

To make our wider purpose more concrete, however, we will illustrate our methods with a narrower design problem that arises in the following example:

*Example 1.* An example that we will follow through the article is a particular test based on kernel density estimation. Suppose that $n_h(x, y) = (2\pi h^2)^{-1/2} \exp((x - y)^2/2h^2)$ is the normal density, $\hat{f}_h(x) = n^{-1} \sum n_h(x, y_i)$ is the kernel density estimator, and $g$ is a null hypothesis density. Let $g^*(x) = \int g(y) n_h(x, y)dy$ be the kernel smoothed null hypothesis and let $f_\tau$ be the true density. Then a test of $H_0 : f_\tau = g$ could be based on the $L_2$ distance $\mathbb{D} = \int(\hat{f}_h(x) - g^*(x))^2 dx$. There is a natural multivariate extension to a normal kernel. This is a quadratic distance, whose kernel $K_h$ will be identified in the next section.

Much of the literature on quadratic goodness-of-fit testing is focused on specific kernels, or, if more general, does not consider the use of DOF as a tool in understanding power characteristics. There is also considerable literature on the choice of bandwidths in the context of density estimation where consideration of mean squared error has been paramount. Rather than detail all this literature here, we provide a review of some key articles in the online supplementary materials. There is also some new literature in the machine learning world relevant to this article that will be reviewed in Section 2.

## 2. QUADRATIC DISTANCES

We start with some basic concepts about quadratic distances. The canonical example of quadratic distance is the *Pearson's chi-squared distance* whose kernel is given by

$$K(r, s) = \sum_{i=1}^{m} \frac{I(r \in A_i) I(s \in A_i)}{G(A_i)}. \tag{1}$$

Here $I$ is the indicator function and $A_1, A_2, \ldots, A_m$ is a partitioning of the sample space into $m$ bins. The empirical distance is then

$$\sum_{i=1}^{m} \frac{(\hat{F}(A_i) - G(A_i))^2}{G(A_i)}.$$

To obtain the correct asymptotic theory, we must modify the kernel $K$ by centering it to obtain $K^{\mathrm{ctr}}$ (details in Section 2.1), in which case we can write $\mathbb{D}(F, G) = \iint K^{\mathrm{ctr}}(s, t) dF(s) dF(t)$. In this form it is clear that $\mathbb{D}(\hat{F}, G)$ is a $V$-statistic. The corresponding $U$-statistic that *unbiasedly* estimates the true distance $\mathbb{D}(F, G)$ is given by the expression

$$U_n = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} K^{\mathrm{ctr}}(x_i, x_j). \tag{2}$$

We will later use the simple form of $U_n$, which has an explicit mean and variance, to approximate the power properties of the test.

### 2.1 Construction of Quadratic Kernels Using Root Kernels

In a mathematical sense, the properties of the quadratic distance are completely determined by the kernel $K(r, s)$ and the model under investigation. However, in the previous literature on distance kernels, there has been little or no work on the construction of a nonnegative definite kernel $K$ that is designed to meet specific goals. Throughout this research we have found the concept of the root kernel to be fundamental to the analysis and construction of these distances. We introduce them here.

Let $k(x, t)$ be an arbitrary user-chosen kernel (possibly depending on $G$), where $x$ is in the sample space and $t$ is in some space, which we will call the *parameter space*. In some cases the parameter space is the sample space. This kernel $k$ will be called the *root kernel*. Let $du(t)$ be a user-selected measure on the parameter space (also possibly depending on $G$). The symmetric kernel $K(x, y) = \int k(x, t) k(y, t)\ du(t)$ is always nonnegative definite; it can be viewed as the functional analogue of a symmetric nonnegative definite matrix of the form $AA^T$. Note that both $k$ and $du$ can depend on $G$.

As an example, if one uses as the root kernel $k$ the normal kernel $\phi_{h^2}(x, t)$with mean $t$ and variance $h^2$, and $du(t) = dt$, then we are in the setting of Example 1. Then the distance kernel $K$ is a normal kernel $\phi_{2h^2}(x, y)$ with variance $2h^2$. We will call $h$ the bandwidth parameter.

This construction also offers a second interpretation to the quadratic distance. If we let $f^*(y) = \int k(y, t) dF(t)$ and $g^*(y) = \int k(y, t) dG(t)$, then we can, by reversal of orders of integration, rewrite the quadratic distance as the $L_2$ distance $\mathbb{D}(F, G) = \int (f^*(t) - g^*(t))^2\ du(t)$. If $k(s, t)$ is a smoothing kernel, as used in density estimation, then the quadratic distance is an $L_2$ distance between kernel smoothed versions of $F$ and $G$. Much of the literature on quadratic distances has started from this $L_2$ perspective, and so has focused on the root kernel $k$ rather than the distance kernel $K$.

If instead one uses $k(x,t) = e^{itx}$, so $f^*(t)$ is the characteristic function, and uses $\|f^*(t) - g^*(t)\|^2$ as the argument, one has a distance based on characteristic functions with distance kernel $K(x,y) = \int \cos(t(x-y))du(t)$. A variation on this characteristic function distance was used, for example, in the context of testing for independence, by Szekely, Rizzo, and Bakirov (2007).

It is less obvious, but this process is quite generally reversible. That is, if one starts with a nonnegative distance kernel $K(s,t)$, there generally exists a symmetric "square root" kernel $k(s,t)$ satisfying the relationship $\int k(s,r)k(t,r)\,du(r) = K(s,t)$, and so it can be viewed as the root kernel in an $L_2$ distance representation of the quadratic distance.

Thus, rather than selecting kernels $K$, we will focus on selecting a root kernel $k(x,t)$. Here are three goals in this selection:

- First, it should give good *power properties*. This is the main focus of this article.
- Second, it should provide a well-behaved null distribution. We will later give an example in which an attempt to grab more power in the tails, via the Pearson-normal kernel, degrades the null distribution.
- A third, more practical, consideration is that $K$ should be explicitly computable for the given $k$. Generally speaking, we think the class of easy-to-compute kernels is rich enough to suit almost every purpose.

As an example of the third item, using $k(x,t) = e^{itx}$ and $du(t)$ as $\phi_\tau(t,0)dt$, we get the family of normal kernels back.

$$K_\tau(s,t) = \int e^{it(x-y)}\phi_\tau(t,0)dt = e^{-\tau^2(x-y)^2}.$$

This example also illustrates the point that a single distance kernel $K$ will have many root kernel representations via the choice of the measure $du(t)$.

From the computational point of view, a particularly nice family of root kernels are based on Markov diffusions. This is because the Markov property dictates that the transition kernel $K_t(x,y)$ of the process, which indicates the chances of going from state $x$ to state $y$ in time $t$, satisfies what we will call the *diffusion equation*:

$$K_{2t}(x,y) = \int K_t(x,z)K_t(z,y)\,du(z),$$

for time parameter $t$. Hence, the root kernel and distance kernel are in the same family. See Kondor and Lafferty (2002) and Lafferty and Lebanon (2005) for the use of diffusion kernels in machine learning.

Finally, it is important to understand that the distance kernel $K$ that generates a particular distance is not unique. The calculations for the basic theory require finding the $G$-centered kernel

$$K^{\text{ctr}}(s,t) = K(s,t) - K(s,G) - K(G,t) + K(G,G),$$

where we used the convention that when a distribution, here $G$, replaces an argument, it means we have integrated the argument out with $G$. For example, $K(s,G) = \int K(s,t)dG(t)$ and $K(G,G) = \iint K(s,t)dG(s)dG(t)$. Note that $K^{\text{ctr}}$ generally depends on $G$ even if $K$ does not. It is important to note that $K^{\text{ctr}}$ gives exactly the same distance measure as $K$. The distinction is that $K^{\text{ctr}}$ must be used in the theoretical calculations.

We note, as a new result, that if $K$ has root kernel $k$, then the centered kernel can be expressed as

$$K^{\text{ctr}}(x,y) = \int (k(x,t) - k(G,t))(k(y,t) - k(G,t))\,du(t).$$

That is, the root of the centered kernel $K^{\text{ctr}}$ is $k^{\text{ctr}}(x,t) = k(x,t) - k(G,t)$.

## 2.2 Distributional Properties of Quadratic Distance

Suppose we have a given parametric model $\{G_\theta : \theta \in \Theta\}$ and a single sample $x_1, \ldots, x_n$ from an unknown distribution $F$. We now wish to test whether $F$ is in the family $G_\theta$. In this section, we bring forth the facts that will enable us to create a simple power analysis in the next section.

We start with the simple null hypothesis, where $\{G_\theta\}$ is actually $G$, a single probability measure.

*Simple Null Hypothesis.* Quite generally there exists a spectral decomposition of the kernel $K^{\text{ctr}}$ that depends on $G$. Suppose the kernel $K(s,t)$ satisfies the condition $\iint K(s,t)dG(s)dG(t) < \infty$, where $G$ is the null probability measure. The spectral decomposition theorem then gives a decomposition of the form

$$K^{\text{ctr}}(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t), \tag{3}$$

where the $\lambda_j$'s and $\phi_j$'s are the eigenvalues and the normalized eigenfunctions of $K$ under measure $G$. However, in most real multivariate cases it is quite hard to determine the eigenvalues and eigenfunctions, so we believe it is very important to focus on approximation methods that avoid this difficulty.

If the decomposition has eigenvalues $\lambda_i$, then under the null hypothesis the limiting distribution of the $V$-statistic is $n\mathbb{D}(\hat{F}, G) \to \sum_j \lambda_j Z_j^2$. The limiting distribution of the $U$-statistic is

$$nU_n \to \sum_j \lambda_j \left(Z_j^2 - 1\right).$$

We denote the last distribution as $\chi^2(\lambda^*)$, with $\lambda^* = (\lambda_1, \lambda_2, \ldots)$.

This limiting result is not particularly useful, given the dependence on infinitely many eigenvalues. However, we can learn about their key properties by integration of $K^{\text{ctr}}$. The sum of the eigenvalues is the trace of $K$ under $G$, namely

$$\operatorname{tr}(K^{\text{ctr}}) = \sum_{j=1}^{\infty} \lambda_j = \int K^{\text{ctr}}(s,s)dG(s).$$

The trace of $(K^{\text{ctr}})^2$, with respect to $G$, is the sum of squared eigenvalues,

$$\operatorname{tr}((K^{\text{ctr}})^2) = \sum_{j=1}^{\infty} \lambda_j^2 = \iint (K^{\text{ctr}})^2(s,t)dG(s)dG(t) < \infty.$$

As a new result, we note that these fundamental quantities can be represented using root kernel as

$$\operatorname{tr}(K^{\mathrm{ctr}}) = \int \operatorname{var}_G(k^{\mathrm{ctr}}(X, t))du(t),$$

and

$$\operatorname{tr}((K^{\mathrm{ctr}})^2) = \iint \operatorname{cov}_G(k^{\mathrm{ctr}}(X, t), k^{\mathrm{ctr}}(X, s))du(s)du(t).$$

Given these quantities, we have the following definition.

*Definition 2.* The DOF, under measure $G$, of a kernel $K$ are defined to be

$$\operatorname{DOF}(K) = \frac{[\operatorname{tr}(K^{\mathrm{ctr}})]^2}{\operatorname{tr}((K^{\mathrm{ctr}})^2)} = \frac{(\sum \lambda_j)^2}{\sum \lambda_j^2}. \tag{4}$$

For the standard Pearson chi-squared distance there are a finite number of positive eigenvalues, all equal, and $\operatorname{DOF}(K^{\mathrm{ctr}})$ is just the usual DOF. In other cases, $\operatorname{DOF}(K^{\mathrm{ctr}})$ represents the DOF of the Satterthwaite approximation of $\chi^2(\lambda^*)$. LMR (2008) showed that for large DOF both the Satterthwaite approximation and the $\chi^2(\lambda^*)$ are close to normal, a result we will use later. Note that $\operatorname{DOF}(K^{\mathrm{ctr}}) = \operatorname{DOF}(a \cdot K^{\mathrm{ctr}})$ for any constant $a$.

For our later interpretation, it is important to note from the article by LMR (2008) that the standardized kernel $K_{\mathrm{std}} = K^{\mathrm{ctr}} \cdot \operatorname{tr}(K^{\mathrm{ctr}})/\operatorname{tr}((K^{\mathrm{ctr}})^2)$ has the "chi-squared" property of $\operatorname{tr}(K_{\mathrm{std}}) = \operatorname{tr}(K_{\mathrm{std}}^2) = \operatorname{DOF}(K) = \operatorname{DOF}(K_{\mathrm{std}})$. Such a rescaled distance provides exactly the same inference as $K$. This means that we can, without loss of generality, interpret DOF as being the mean and half of the variance of the standardized distance statistics.

*Under the Alternative.* Under the alternative, the $U$-statistic has mean $D(F_\tau, G)$, where $F_\tau$ is the true distribution, and variance given as follows:

*Proposition 1.* The exact variance of $U_n$ under the true distribution $F_\tau$ is given by

$$\begin{aligned}\operatorname{var}(U_n(G)) = {} & \frac{2}{n(n-1)} E_{F_\tau}[K^{\mathrm{ctr}}(X_1, X_2)]^2 \\ & + \frac{4}{n} E_{F_\tau}[(\Delta(X))^2],\end{aligned} \tag{5}$$

where $X_1, X_2$ are independent replicates from $F_\tau$, $K$ is the centered kernel with respect to $G$, and

$$\Delta(x) = K^{\mathrm{ctr}}(x, F_\tau) - K^{\mathrm{ctr}}(x, G) + K^{\mathrm{ctr}}(F_\tau, G) - K^{\mathrm{ctr}}(F_\tau, F_\tau).$$

*Proof.* Given in the online supplementary materials. □

To simplify our notation hereafter, we will write

$$\operatorname{var}_{F_\tau}(U_n) = a_n \varpi + b_n \upsilon,$$

where $\varpi = E_{F_\tau}[K_G(X_1, X_2)]^2$, and $\upsilon = E_{F_\tau}[(\Delta(X))^2]$, $a_n = \frac{2}{n(n-1)}$, $b_n = \frac{4}{n}$.

*Composite Null Hypothesis.* We now suppose that one has estimated the parameter $\theta$ under the null with some estimator $\hat{\theta}$ and that one proposes to use $D(\hat{F}, G_{\hat{\theta}})$ as a test statistic for the null model $\{G_\theta\}$. The preceding results no longer directly apply. Due to its dependence on $\hat{\theta}$, the distance measure $D(\hat{F}, G_{\hat{\theta}})$ is no longer a simple quadratic function of $\hat{F}$ and the corresponding $U_n$ is not an unbiased estimator of distance. For example, we know from the chi-squared example that the DOF are reduced by the number of parameters estimated.

One can, however, hope to appeal to the preceding theory by approximating $D(\hat{F}, G_{\hat{\theta}})$ with a quadratic distance having a modified kernel $K^*$. See LMR (2008) for the special case when $\hat{\theta}$ is the maximum likelihood estimator. We here offer a new generalization of this result to other estimators. We assume that the estimator can be expressed as functional $\hat{\theta}(F)$ that are consistent for $\theta$ in $\{G_\theta\}$ in the sense that $\hat{\theta}(G_\theta) = \theta$ for all $\theta$. This includes the minimum distance estimator based on $K$.

The simplest way to derive such an approximation is via a formal von Mises expansion. This will in fact identify the correct structure, although in any particular case such an expansion cannot guarantee that the remainder is stochastically small, which must be done by other means.

To enable this calculation, we assume that the vector $g(\varepsilon) = \hat{\theta}(F_\varepsilon)$, where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon \hat{F}$, has a linear first-order expansion of the form

$$\begin{aligned} g(0) &= \theta(F), \\ \frac{d}{d\varepsilon} g(\varepsilon)|_{\varepsilon=0} &= \int \theta'(x) d\hat{F}(x), \end{aligned} \tag{6}$$

so that $\theta'(x)$ is the vector influence function for $\hat{\theta}$.

Let the centered version of root kernel $k(x, t)$ be $k^{\mathrm{ctr}}(\theta, x, t)$, where we express the dependence on $G_\theta$ through the first argument $\theta$. Define the *estimation-centered root kernel* to be

$$\begin{aligned} v_F(\theta, x, t) = {} & k^{\mathrm{ctr}}(\theta, x, t) - k^{\mathrm{ctr}}(\theta, F, t) \\ & + \theta'(x)^T [\nabla_\theta k^{\mathrm{ctr}}(\theta, F, t)]. \end{aligned}$$

*Proposition 2.* Suppose that $\hat{\theta} = \theta(\hat{F})$ is an estimator of vector $\theta$ with vector influence function $\theta'(x)$. Let $h(\varepsilon)$ be the formal von Mises approximation to the testing problem (see Appendix). Then the first two formal von Mises derivatives of the empirical distance $\mathbb{D}(\hat{F}, G)$ are

$$h'(0) = 2 \iint v_F(\theta, x, t)(k^{\mathrm{ctr}}(\theta, F, t))du(t)d\hat{F}(x)$$

and

$$h''(0) = 2 \iint \left( \int v_F(\theta, x, t) v_F(\theta, y, t) du(t) \right) d\hat{F}(x) d\hat{F}(y).$$

In particular, $h'(0) = 0$ under the null hypothesis, and so the *estimation-centered root kernel* $v_G(\theta, x, t)$ is an approximate root kernel for the null hypothesis.

*Proof.* Given in the Appendix. □

*Remark 1.* We recognize that virtually every real application of goodness-of-fit testing involves nuisance parameters $\theta$ in the null hypothesis. However, there are sufficient complexities in the quadratic power problem, so our purpose here will be to take the first, and most important, step of verifying that our surrogate power analysis works in a simple null hypothesis.

*Two-Sample Quadratic Tests.* The quadratic distance formulation lends itself naturally to the comparison of any two distributions $F$ and $G$, no matter how they arise. Thus, it is not surprising that there is some literature relating them to the two-sample problem where $F$ and $G$ are estimated by independent samples. There is a significant literature in machine learning, represented by Gretton et al. (2012) and Poczos, Ghahramani, and Schneider (2012).

This is relevant to our present purpose for two reasons. First, they have found that kernel-based two-sample comparisons were more powerful than other distance methods considered (Friedman and Rafsky 1979; Anderson et al. 1994; Biau and Gyorfi 2005; Sriperumbudur et al. 2010). This substantiates that we are on a useful track. Second, if we cannot explicitly calculate the integrals in $\mathbb{D}(\hat{F}, G)$, but we can simulate, then we can easily turn the goodness-of-fit problem into a two-sample problem. We take a large sample from $G$, create the empirical $\hat{G}$, and compute $\mathbb{D}(\hat{F}, \hat{G})$ instead of $\mathbb{D}(\hat{F}, G)$. There is, of course, a theoretical difference in that the true distribution $G$ for the simulation is known, so one can simulate with arbitrary accuracy. We leave the details of this extension to future work.

## 3. POWER CONSIDERATIONS

One challenge in undertaking a study of power for quadratic distances is that the distributional theory is rather complex. This leaves two obvious options: numerical approximation and simulation. A simulation approach usually leads to little insight, so we construct a surrogate power function in this section. In our numerical section, we will compare it with simulated power curves.

### 3.1 The Surrogate Power Function

Under a general alternative $F_\tau$, the $U$-statistics given in (2) provides us with a simple relationship $E(U_n) = \mathbb{D}(F_\tau, G)$. A second nice feature of the $U$-statistic version of the test statistic is that we have an explicit formula for the exact variance under an arbitrary alternative (see (5)).

Our power approximation is based on two observations:

- Under the null hypothesis, the limiting distribution for $nU_n$ is asymptotically $\chi^2(\lambda^*)$. We know that this distribution is approximately normal for large DOF. (And, as we will later see, the best power tends to occur for large DOF.)
- Under a fixed alternative $F_\tau$, the limiting distribution for $U_n$ is again normal, with asymptotic variance $\frac{4}{n} E_{F_\tau}[(\Delta(X))^2]$. This follows from standard $U$-statistic theory.

We therefore propose to use a normal approximation under both null and alternative, but when we create the approximation we will use the *exact* mean and variance of the $U$-statistic under the chosen alternative $F_\tau$.

We first use the normal approximation under the null to write the rejection region as $\{U_n > z_\alpha \sqrt{a_n \varpi_0}\}$, where $\varpi_0$ is the variance calculated under the null hypothesis and $z_\alpha$ is the normal critical value for a size $\alpha$ test. (This critical value could be replaced by a simulated value if one desires more accuracy.) The power can then be approximated using asymptotic normality under the alternative $F_\tau$, with $\mathbb{D} = \mathbb{D}(F_\tau, G)$, using the *surrogate power function*

$$\beta(F_\tau) = P_{F_\tau}(U_n > z_\alpha \sqrt{a_n \varpi_0}) = P\left( Z > \frac{z_\alpha \sqrt{a_n \varpi_0} - \mathbb{D}}{\sqrt{a_n \varpi + b_n \upsilon}} \right).$$

We note that the surrogate power function smoothly interpolates between the null and alternative variance functions.

### 3.2 The Midpower Manifold and NCI

Our next observation is that the $\beta(F_\tau) = 0.5$ when $z_\alpha \sqrt{a_n \varpi_0} - \mathbb{D} = 0$. This makes for an even simpler analytic calculation than the full surrogate power, and so we will focus on power 0.5. Further, because it is equivalent asymptotically and makes calculations simpler, hereafter we will replace $a_n = 2/(n)(n-1)$ with $a_n = 2/n^2$.

We therefore define the *noncentrality index* to be

$$\mathrm{NCI}(F) := \left( n \frac{\mathbb{D}(F, G)}{\sqrt{2\varpi_0}} \right),$$

noting that $\{F : \mathrm{NIC}(F) = z_\alpha$ gives a manifold of alternatives $F_\tau$ where the surrogate power $\beta(F_\tau)$ of the size $\alpha$ test is 0.5. Call this the *midpower manifold.*

We will compare the power structure of kernels by analysis of this surrogate midpower manifold. As an additional heuristic justification for NCI, we note that in the partitioned chi-square case it equals the traditional noncentrality parameter used in an analysis of local alternatives.

Note that the kernel $K$ is implicitly involved in the numerator and denominator of NCI. Indeed, it clarifies the key issue: power at $F$ is a tradeoff between increasing the sensitivity of the distance $\mathbb{D}$ to the target distribution $F$ and minimizing the inflation of the null variance. After standardizing the distance, the squared denominator equals the DOF, so we need to control DOF. Under an alternative distribution $F$, one can write the numerator as

$$\mathbb{D}(F, G) = \sum_j \lambda_j E_F^2[\phi_j(X)].$$

Thus, the distance, and so the numerator, is determined by the eigenmoments $E_F[\phi_j(X)]$, whose nonzero values signify departures from the null, as well as the eigenvalues $\lambda_j$, as they determine which eigenmoment failures are weighted most heavily. For the normal kernel, the eigenmoments are damped Hermite polynomial moments (LMR 2008).

*Remark 2.* We note that Gourieroux and Tenreiro (2001) studied the power properties of kernel-based goodness-of-fit tests assuming that the bandwidth parameter $h_n \to 0$, under a sequence of local alternatives. Here, we have offered a simpler approximation to the power that is approximately valid when there is no bandwidth, or the bandwidth $h$ is fixed. This is important because we will see that in a testing situation, it is highly likely that the best selected bandwidths do *not* converge to zero as $n$ goes to infinity. Indeed, as $d$ increases, they may become infinite.

*Remark 3.* One might also think that a local alternatives (Neyman 1937; Pitman 1948; McManus 1991) approach to our problem would somehow offer a more useful or accurate power approximation. We think Occam's razor is the best guide. Note that the midpower manifold clearly depends on the sample size $n$, and *always* represents "local" alternatives that are of interest at $n$. If one were to use a noncentral chi-square, or noncentral $\chi^2(\lambda^*)$, approach to this problem, we believe that one would end up with essentially the same analysis, provided that one simplified the noncentral problem further and used suitable normal approximations for the noncentral distribution calculations.

## 3.3 The Midpower Lemma

We will first use our surrogate power analysis to show how one can build a kernel that is targeted toward midpower for a particular family of alternatives. This analysis will then be applied to better understand the strengths of the Gaussian kernel. We will also use the method to build a new class of kernels more similar to the chi-square.

Let $F_\theta$ be a family of alternatives to the simple null hypothesis $G$. Let $W$ be a prior probability measure, with $w(\theta)$ its density on $\theta$, that assigns weights to the values of the parameter. The weighting density should be highest at those alternatives where one wishes to focus on the midpower of the test. That is, as a substitute for maximizing the full surrogate power function, we consider the problem of choosing a centered kernel $K(x, y)$ that maximizes the following integrated noncentrality index over this family of alternatives:

$$\mathrm{NCI}_w(K) = \frac{\left(\int \mathbb{D}(F_\theta, G)dW(\theta)\right)}{\mathrm{var}_G^{1/2}(K)}.$$

Note that this problem has built into it the tradeoff between the sensitivity of distance to $\{F_\theta\}$ and the variance of the test statistic under the null.

This optimality problem has a simple solution $B(x, y)$ that should have good midpower properties over $F_\theta$, with its greatest emphasis on those alternatives with high weight $w(\theta) = dW(\theta)$. The uncentered version of $B$ is

$$B^*(x, y) = \int \frac{f_\theta(x)}{g(x)} \frac{f_\theta(y)}{g(y)} dW(\theta).$$

This kernel has the root kernel $\frac{f_\theta(x)}{g(x)}$ under measure $du(\theta) = dW(\theta)$. We can consider $B^*$ to be the ratio of the $(x, y)$ density $f_W(x, y) = \int f_\theta(x) f_\theta(y) dW(\theta)$ to the null density $g(x)g(y)$. If we center this kernel using measure $G$, we find the centered likelihood ratio kernel

$$\begin{aligned} B(x, y) &= \left(\int_\theta \left[\frac{f_\theta(x)}{g(x)} - 1\right]\left[\frac{f_\theta(y)}{g(y)} - 1\right] dW(\theta)\right) \\ &= B^*(x, y) - \frac{\int f_\theta(x)dW(\theta)}{g(x)} - \frac{\int f_\theta(y)dW(\theta)}{g(y)} + 1. \end{aligned}$$

The last expression makes it easy to show that $\mathrm{tr}(B) = \mathrm{tr}(B^*) - 1$, where

$$\mathrm{tr}(B^*) = \int \left[\int \frac{f_\theta(x)^2}{g(x)} dx\right] dW(\theta).$$

We will call the following the *midpower lemma.*

*Lemma 1* (Midpower Lemma). Suppose that $\mathrm{tr}(B^2) = \iint B^2(x, y)dG(x)dG(y)$ is finite. The objective function $\mathrm{NCI}_w(K)$ is maximized uniquely by kernels equivalent to $B$, and has the value

$$\mathrm{NCI}_w(B) = \sqrt{\iint B^2(x, y)dG(x)dG(y)}.$$

*Proof.* To solve the above optimization problem, we rewrite the numerator of $\mathrm{NCI}_w(K)$ as

$$\begin{aligned} \int D(F_\theta, G)dW(\theta) &= \int_\theta \left(\int_y \int_x K(x, y)\left[\frac{f_\theta(x)}{g(x)} - 1\right]\right. \\ &\qquad \times \left.\left[\frac{f_\theta(y)}{g(y)} - 1\right] dG(x)dG(y)\right) dW(\theta) \\ &= \int_y \int_x B(x, y)K(x, y)dG(x)dG(y). \end{aligned}$$

The optimization problem is then to find the centered kernel $K$ that maximizes

$$\mathrm{NCI}_w(K) = \frac{\left[\iint B(x, y)K(x, y)dG(x)dG(y)\right]}{\sqrt{\iint K^2(x, y)dG(x)dG(y)}}.$$

The result now follows from the Cauchy–Schwarz inequality. □

The following corollary shows the resemblance between the midpower lemma and the Neyman-Pearson lemma.

*Corollary 1.* The maximum noncentrality index for testing simple $g$ versus simple $f$ is attained by the centered rank one kernel $B(x, y) = [\frac{f(x)}{g(x)} - 1][\frac{f(y)}{g(y)} - 1]$. That is, the likelihood ratio $\frac{f(y)}{g(y)}$ is a root kernel for the distance.

When one is creating the weighting prior $w$, one might need to take some care to ensure that $\mathrm{tr}(B^2)$ is finite. In particular, note that for $\mathrm{tr}(B)$ to be finite it is necessary that the inner integral

$$h(\theta) = \int \frac{f_\theta(x)^2}{g(x)} dx$$

be finite with probability 1 under $W$. That is, one must choose the family of alternatives $F_\theta$ so the elements have a finite chi-square distance from $g$. Of course, this condition is not sufficient, and so one also needs to check whether $h(\theta)$ has a finite integral under $dW$. However, the finiteness of $\mathrm{tr}(B^*)$ can be shown to suffice for other key quantities to be finite, such as $\int \mathbb{D}_B(F_\theta, G)dW(\theta)$ and $\mathrm{tr}(B^2)$, and hence DOF.

*Reverse Interpretation of a Kernel.* In this section we use the midpower lemma in reverse to show that every kernel $K$ has an implicit set of alternatives against which it is optimal. This in turn leads to a new interpretation of the Gaussian kernel.

*Proposition 3.* Let $g(x)$ be the null density. Let $k(x, t)$ be a root kernel for $K$. Let the family $F_\theta$ of alternatives to $g$ have the form

$$f(x, \theta) = g(x)k(x, \theta)/g_h^*(\theta),$$

where $g^*(\theta) = \int g(x)k(x, \theta)dx$ is the normalizing constant. Let the prior weighting measure be $dW(\theta) = Cg^*(\theta)^2 du(\theta)$, where $C$ is the normalizing constant for $dW$. Then the optimal kernel $B$ is simply the kernel $K$.

*Proof.* This is an easy calculation. □

We can apply this result to the normal kernel with root $\phi_h(x, t)$ to get some insight into its implicit weighting of alternatives. This alternative distribution with density $f(x, \theta) = g(x)\phi_h(x, \theta)/g_h^*(\theta)$, converges to point mass at $\theta$ as $h \to 0$. That is, for $h$ small, this alternative is very similar to putting mass 1 in a small neighborhood of $\theta$. Under continuity of $g$ the weight function $w_h(\theta)$ converges to $C_0 g^2(\theta)$ where $C_0^{-1} = \int g^2(\theta)d\theta$, as $h \to 0$.

Our conclusion is that the normal kernel with a *small* bandwidth $h$ has its midpower tuned particularly to a family of highly concentrated alternatives at all possible locations $\theta$, where we assign larger weight to alternatives at those $\theta$ for which $g(\theta)^2$ is large. The mean noncentrality for this scheme is

$$C_t \sqrt{\operatorname{var}_G\left(\phi^{\text{ctr}}_{\sqrt{2}h}\right)},$$

where $\operatorname{var}_G(\phi^{\text{ctr}}_{\sqrt{2\text{h}}})$ is the variance of the centered normal kernel.

One might compare the optimality properties of the normal kernel with those of binned chi-squared kernel. In this case one could use as alternatives to $G$ the family of densities $f(x, \theta) = I(x \in A_\theta)g(x)/G(A_\theta)$, where $\theta$ is the bin index. If one uses the masses $G(A_\theta)$ for $w(\theta)$, then the optimal kernel is the chi-square kernel. The key difference between the chi-squared and Gaussian kernels is that the weights for the normal are not proportional to $g(\theta)$, but rather proportional to $g(\theta)^2$, and so the normal kernel puts less weight in the tails of $g$ than the chi-square.

*Pearson Normal Kernel.* The possible limitations of the Gaussian kernel suggest that one use the midpower lemma to improve it. In the process we will arrive at a new kernel with a number of interesting mathematical properties.

The following is a new method of constructing kernels. Let $k_h(x, t)$ be a "basis" kernel with bandwidth parameter $h$. As in the preceding case, let the class of parametric alternatives have the form $k_h(x, \theta)g(x)/g^*(\theta)$. If one applies the midpower lemma using as a weighting density $g_h^*(t)$ instead of $g_h^*(t)^2$, in the spirit of the chi-squared construction, one arrives at the following *Pearsonized kernel*:

$$K(x, y) = \int \frac{k_h(x, \theta)k_h(y, \theta)}{g_h^*(t)} d\theta. \tag{7}$$

After reversing orders of integration, we obtain

$$\mathbb{D}(F, G) = \int \frac{(f_h^*(t) - g_h^*(t))^2}{g_h^*(t)} dt.$$

This distance mimics the "squared standardized deviates" feature of the chi-squared distance while having DOF that are continuously adjustable through the bandwidth $h$. The general form of this kernel was given by Seo and Lindsay (2008) who studied this kernel as a quadratic equivalent of the corresponding doubly smoothed Kullback–Leibler discrepancy,

$$\int f_h^*(x) \log(f_h^*(x)/g_h^*(x)) dx.$$

The Pearsonized kernels described above always have a simple centering operation $K^{\text{ctr}}(x, y) = K(x, y) - 1$. This makes them particularly attractive for an eigenanalysis as the centered and uncentered kernels have the same eigenfunctions. They differ only on the eigenvalue for the constant eigenfunction 1.

We apply this new methodology to our problem to arrive at a new kernel, the Pearson normal kernel.

*Example 2* (The Pearson-normal kernel). If the basis kernel is $\phi_{hI}(x, t)$, a $d$-dimensional multivariate normal with variance $h^2 I$, and the null distribution function $G$ is multivariate standard normal, the *Pearson-normal kernel* can be calculated as

$$K_h(r, s) = \frac{(h^2 + 1)^d}{h^d (h^2 + 2)^{d/2}} \exp\left[-\frac{(r - s)^T (r - s)}{2h^2(h^2 + 2)} + \frac{r^T s}{h^2 + 2}\right].$$

This kernel is a product, $K_h(\mathbf{x}, \mathbf{y}) = \prod K_h(x_i, y_i)$, over the univariate kernels

$$K_h(r, s) = \frac{h^2 + 1}{h(h^2 + 2)^{1/2}} \exp\left[-\frac{(r - s)^2}{2h^2(h^2 + 2)} + \frac{rs}{h^2 + 2}\right].$$

This kernel, which is new to our knowledge, has some interesting mathematical properties. As a function of $r$ and $s$ the Pearson-normal kernel has considerable superficial similarity to the normal kernel. The following proposition shows that it can be interpreted as a ratio of normal densities.

*Proposition 4.* In the univariate case the Pearson-normal kernel can be represented as

$$K_h(x, y) = \frac{f_\rho(x, y)}{n_1(x, 0)n_1(y, 0)},$$

where $f_\rho$ is a bivariate normal density function for variables $X, Y$ with means 0 and variances 1, and covariance $\rho = 1/(h^2 + 1)$.

*Proof.* Given in the online supplementary materials. □

## 4. THE MIDPOWER ANALYSIS

In the preceding section, we have shown that the optimality properties of a kernel have value in the interpretation of the kernel and even in construction of kernels. However, it still only provides weak guidance on the selection of the bandwidth parameters $h$ in a kernel like the normal or Pearson normal. We therefore develop a more precise strategy that could be useful in comparing the sensitivity of a fixed class of kernels. We denote this class as $\{K_h\}$ where $h$ is any index of the kernels, whether a bandwidth, a collection of bandwidths, or some other index. We call $h$ bandwidth here for simplicity.

For each bandwidth $h$ there is a midpower manifold $B_h$, surrounding the null hypothesis $G$ in the space of distributions that corresponds to those $F$ that have power 0.50. We cannot expect that there will be any value of $h$ that is superior to all others, in the sense of having its manifold inside all the others. In fact, we will have to set a statistically meaningful criterion to narrow our choices. We illustrate with one possibility that we like because of its ease in interpretation. It gives a way to select $h$ as a way to measure the sensitivity of the test.

### 4.1 Midpower in a Target Alternative Family

Our first step is to create one or more families of *target* alternative hypotheses $\{F_\delta\}$, each with a scalar parameter $\delta \geq 0$. We think of $\{F_\delta\}$ as a one-dimensional curve of alternatives through the space of all alternatives. These families should be constructed so as to suit the purposes of the tester; they might be "most plausible," "most interesting," or be based on the "deviations that are likely to negate the conclusions we would draw if we used the null model." We assume that $\delta = 0$ corresponds to the null, and that as $\delta$ increases the departure from the null becomes more severe. Hence, $\{F_\delta\}$ includes local as well as remote alternatives. As we shall see in the next section, this format can easily be extended to a larger family of alternatives $\{F_{\theta,\delta}\}$, with scalar $\delta$ and vector $\theta$, together with a weighting prior $W$ on $\theta$.

Assuming each target family $F_\delta$ is smooth in $\delta$, then there exists, for each $h$, a smooth power curve $\beta_h(\delta)$ that depends on $n$ and on the DOF through the tuning parameter $h$. For each bandwidth, we define the *midpower sensitivity* of the bandwidth $h$ by

$$\delta_{\text{mid}}(h) = \arg\min\{\delta : \beta_h(\delta) = 0.5\}.$$

This is the smallest value of $\delta$ where the test achieves power 0.5. Note that this step depends on $\delta$ being univariate. Otherwise $\{\delta : \beta_h(\delta) = 0.5\}$ would be a set of $\delta$ values.

If the power curve is monotonically increasing in $\delta$, then the midpower alternative $\delta_{\text{mid}}(h)$ divides the values of $\delta$ into two intervals where the chances of success in rejecting the alternative are below or above 0.5. We think the 50–50 nature of the midpower alternative makes $\delta_{\text{mid}}(h)$ a natural measure of the overall sensitivity of the test based on bandwidth $h$. If a test with bandwidth $h_1$ has a midpower alternative $\delta_{\text{mid}}(h_1)$ smaller than $\delta_{\text{mid}}(h)$, we will say $h_1$ is *more sensitive* to the alternative family than $h$.

Next we define the *midpower optimal bandwidth* $h^*_{\text{midopt}} = \arg\min_h(\delta_{\text{mid}}(h))$. This is the bandwidth that generates the most sensitive test among all bandwidths. Note that such a test need not be *globally optimal* for $F_\delta$, in the sense that the power curve satisfies $\beta_{h^*}(\delta) \geq \beta_h(\delta)$ for all $\delta$. However, if a globally optimal test exists, it is also midpower optimal. Note that there is nothing in this analysis that restricts $h$ to be univariate.

Insisting on midpower optimality ensures that our focus is on the range of alternatives where the power function is well away from size $\alpha$, that is, cases of very low power, and 1, that is, cases with very high power.

This concludes the selection of the best $h$. As a summary of how well the whole family of kernels performed for the alternatives under consideration, define $\delta^*_{\text{sens}} = \delta_{\text{mid}}(h^*_{\text{midopt}})$to be the *midpower sensitivity over h* of the family of kernels $K_h$ in family $F_\delta$.

Clearly, different target families $F_\delta$ could generate very different $h^*_{\text{midopt}}$, in which case a compromise would have to be made among the resulting bandwidths. In our worked examples, this problem has been very mild. Another solution would be to combine tests with different bandwidths, a subject that we will not tackle here.

### 4.2 Midpower Analysis in Quadratic Distances

A special advantage of using the midpower analysis in the quadratic distance problem is that $h^*_{\text{midopt}}$ and $\delta^*_{\text{sens}}$ can be straightforward to obtain from the surrogate power curve. As we noted earlier, the power approximation is 0.5 when the noncentrality index

$$\text{NCI}(\delta, h) := \left(n\frac{\mathbb{D}(F_\delta, G)}{\sqrt{2\varpi_0}}\right)$$

equals $z_\alpha$. Here $\mathbb{D} = \mathbb{D}(F_\delta, G)$ depends both on $\delta$ in the alternative and $h$, through the kernel $K_h$. However, $\varpi_0$, the variance of the distance statistic under the null, depends only on $h$, not on $\delta$. If we wish to include nuisance parameters $\theta$ in the analysis, thereby creating a larger family of alternatives $\{F_{\theta,\delta}\}$, we could here replace $\mathbb{D}(F_\delta, G)$ with $\int \mathbb{D}(F_{\theta,\delta}, G)dW_\delta(\theta)$.

The special structure of the noncentrality index suggests the following shortcut to doing the midpower analysis. We start by finding the upper envelope of the noncentrality index. We first fix $\delta$, and then maximize the noncentrality index over $h$. Call the resulting bandwidth $h^*_{\text{env}}(\delta)$, and let the resulting *upper envelope* be

$$\text{NCI}^{\text{env}}(\delta) := \sup_h \text{NCI}(\delta, h) = \text{NCI}(\delta, h^*_{\text{env}}(\delta)).$$

Let us then find the smallest $\delta$, say $\delta_{\min}$ for which $\text{NCI}^{\text{env}}(\delta) = z_\alpha$, assuming continuity of the envelope. Values of $\delta$ smaller than $\delta_{\min}$ cannot have power equal to 0.5 for any $h$ because their noncentrality index is too small. However, the bandwidth $h = h^*_{\text{opt}}(\delta_{\min})$ does attain power 0.5 at $\delta_{\min,}$ and so it must be the midoptimal bandwidth $h^*_{\text{midopt}}$. Finally $\delta_{\min}$ is the midoptimal sensitivity $\delta^*_{\text{sens}}$.

## 5. KERNELS IN HIGHER DIMENSIONS

We have now built up a collection of tools for analyzing power and making kernel selections. Before we apply them to the problem of testing for multivariate normality, we must first develop a greater understanding of how kernels work in higher dimensions. We will focus on the Gaussian and Pearson-normal as they are the keys to our detailed study.

As a smoothing kernel in higher dimensions, the normal $\phi_{hI}(x, y)$ is one of a class of kernels of product form

$$\phi_{hI}(x, y) = \prod_{i=1}^{d} \phi_h(x_i, y_i).$$

More generally, one can always create a multivariate kernel by taking a product of univariate kernels. This structure is flexible and fast, and adaptable to different data types.

The multivariate Gaussian also has the feature of being invariant under orthornormal transformations of the variables: if $\Gamma$ is an orthonormal matrix, then $\phi_{hI}(\Gamma x, \Gamma y) = \phi_{hI}(x, y)$. If the model structure makes orthonormal invariance a desirable feature, then one should be cautious of the product formulation as $K_{\text{prod}}(x, y) = \prod K(x_i, y_i)$ is orthonormal invariant for only a few base kernels $K$. (One can show that the product of Pearson normal kernels is also invariant.) In general, a test will be invariant if the kernel $K(x, y)$ is a function of vectors $x, y$ through $x^T x,\ y^T x$, and $y^T y$ alone.

However, in a general problem in which the data vector could have coordinates with range restrictions or some variables discrete, or even other types of variables, then orthonormal invariance seems to be a less relevant criterion. In these settings, creating a multivariate kernel simply by using a product of univariate kernels has some advantages. First, univariate kernels are readily available, or easily constructed, for many data types. Second, the choice of any bandwidth parameters can be done on a coordinate-by-coordinate basis.

The midpower lemma gives us some insight into the narrow circumstances under which product kernels might be an ideal construction. If the null hypothesis $G$ has the product density $\Pi_i g_i(x_i)$, if the family of interesting alternatives has product structure $\Pi f_{\theta_i}(x_i)$, and if the weighting density $w(\theta)$ has product structure $\Pi w_i(\theta_i)$, then the ideal kernel is simply a product of the best kernels for each univariate problem:

$$B^*(x, y) = \prod_i \int \frac{f_{\theta_i}(x_i) f_{\theta_i}(y_j)}{g(x_i) g(y_i)} dW_i(\theta_i).$$

If, however, the null model has independence structure but the prior on interesting alternatives do not, then the optimality theory would lead us away from the product construction. We give a simple illustration in the following section.

### 5.1 Example of Kernel Construction

We now provide an example of optimal but nonproduct kernel. Let the model $G$ be a bivariate normal with mean vector 0 and variance-covariance matrix the identity $I$. The family of alternatives consists of bivariate normals with mean vector $\mu$ and the same identity variance-covariance matrix. Write $\mu_1 = \rho \cos\theta$, $\mu_2 = \rho \sin\theta$, where $0 \le \theta \le 2\pi$, $0 < \rho < \infty$. For fixed $\rho$, let the prior $w(\theta)$ be uniform on $(0, 2\pi)$. Then

$$f_\mu(x) = \frac{1}{2\pi} \exp\left[-\frac{x_1^2 + x_2^2 + \rho^2}{2} + \rho(x_1 \cos\theta + x_2 \sin\theta)\right],$$

and

$$\begin{aligned}\int_0^{2\pi} f_\mu(x) dW(\mu) &= \int_0^{2\pi} \frac{1}{2\pi} \exp\left[-\frac{x_1^2 + x_2^2 + r^2}{2}\right] \\ &\quad \times \exp\left[-\rho(x_1 \cos\theta + x_2 \sin\theta)\right] \frac{1}{2\pi} d\theta \\ &= \frac{1}{2\pi} I_0\left[\rho\sqrt{x_1^2 + x_2^2}\right] \exp\left[-\frac{x_1^2 + x_2^2 + \rho^2}{2}\right],\end{aligned}$$

where $I_0$ is the modified Bessel function of the first kind. Therefore, the centered kernel B can be calculated as

$$\begin{aligned}B(x, y) = \exp\left(-\frac{\rho^2}{2}\right) \Bigg( I_0\left[\rho\left\{\left(x_1^2 + y_1^2\right)^2 + \left(x_2^2 + y_2^2\right)^2\right\}^{1/2}\right] \\ - I_0\left[\rho\sqrt{x_1^2 + x_2^2}\right] - I_0\left[\rho\sqrt{y_1^2 + y_2^2}\right]\Bigg) + 1.\end{aligned}$$

This new kernel has $\rho$ as the tuning parameter, allowing one to tune the kernel, by midpower analysis, to the best power for the given sample size. In keeping with the symmetry of the prior, the distance depends on the data only through the data radii $\sqrt{x^2 + y^2}$.

### 5.2 Product Kernels and the Binning Index

Both the normal and Pearson normal kernels are examples of what we call product kernels:

$$K(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{d} K_j(x_j, y_j).$$

If each kernel $K_j$ is nonnegative definite, with root kernel $k_j$, then it is clear that the product $K$ is nonnegative definite as well. We have already shown that such a kernel might not be ideal for dependent data. We here show how having such a product structure can potentially simplify the calculations need for finding DOF and the noncentrality index NCI.

A key example of such a construction would be a binned chi-square, where the bins are products of coordinate-wise bins, and where one specifies a *product null*: that is, the null hypothesis $G$ specifies that the individual variables $X_i$ are independent with distribution $G_i$. Then the kernel has the form given in (1). For such a kernel, the DOF calculation is well known. It is the total number of bins, minus 1. The product structure implies the total number of bins is the product of the number of bins used for each coordinate. If we used a fixed number of bins per coordinate, say $\mathbb{B}$, then the DOF would be DOF $= \mathbb{B}^d - 1$.

That is, with a cross-product binning strategy, the DOF would be exponential in $d$. To make comparisons with chi-square binning strategies, for a general quadratic distance we define its *binning index* to be

$$\mathbb{B} = (\mathrm{DOF} + 1)^{1/d}.$$

Thus, $\mathbb{B} = 2$, for instance, would correspond to a chi-squared test with two bins for each coordinate. This would be the minimal number needed to gain information on the fit in all coordinate directions using a chi-squared test.

It is an unfortunate fact that the DOF calculation for the uncentered product kernel, which is wrong, can be much simpler algebraically than for the centered kernel, which is right. The simpler calculation still provides important insight as to the effect of dimension. Under the product null $G = \prod G_i$, we have $\mathrm{tr}(K^2) = \prod \mathrm{tr}_{G_i}(K_i^2)$ and $\mathrm{tr}(K) = \prod \mathrm{tr}_{G_i}(K_i)$, and so we get the result that $\mathrm{DOF}(K) = \prod \mathrm{DOF}(K_i)$. That is, ignoring centering, the $\mathrm{DOF}(K)$ would grow exponentially in $d$ just as in the binning chi-squared.

In the online supplementary materials we offer an argument as to why $\mathrm{DOF}(K)$ is quantitatively close to $\mathrm{DOF}(K^{\mathrm{ctr}})$ in general. In particular, exponential growth of the DOF with $d$ is to be expected.

### 5.3 DOF for the Normal Kernel

Returning now to our example, we can see if we hold the bandwidth fixed, we see an exponential growth in the DOF.

*Example 1,* continued. We return to study the DOF function in Example 1. The following propositions offer the connection between the exact DOF and data dimension for the case of a multivariate normal kernel with covariance matrix $\Sigma_h$ applied to a null hypothesis of standard multivariate normality.

*Proposition 5.* If the kernel is a multivariate normal kernel with covariance matrix $\Sigma_h$, and the model is a multivariate normal with covariance matrix $V$, then the theoretical DOF of the *centered* kernel are given by the expression

$$\begin{aligned}\mathrm{DOF} = &((\mid \Sigma_h \mid^{-1/2} - \mid \Sigma_h + 2V \mid^{-1/2})^2) \\ &/ (\mid \Sigma_h \mid^{-1/2} \mid \Sigma_h + 4V \mid^{-1/2} - 2 \mid \Sigma_h + V \mid^{-1/2} \\ &\times \mid \Sigma_h + 3V \mid^{-1/2} + \mid \Sigma_h + 2V \mid^{-1}).\end{aligned}$$

*Proof.* Given in the online supplementary materials. □

*Corollary 2.* If $\Sigma_h = h^2 I$ and $V = I$, where $I$ is the identity matrix, the DOF of the *centered* kernel are given by the expression

$$\mathrm{DOF} = \frac{(h^{-d} - (h^2+2)^{-d/2})^2}{h^{-d}(h^2+4)^{-d/2} - 2(h^2+1)^{-d/2}(h^2+3)^{-d/2} + (h^2+2)^{-d}}$$
$$= \frac{\left(1 - \left[\frac{h^2}{h^2+2}\right]^{d/2}\right)^2}{\left[\frac{h^2}{h^2+4}\right]^{d/2} - 2\left[\frac{h^2}{h^2+1}\right]^{d/2}\left[\frac{h^2}{h^2+3}\right]^{d/2} + \left[\frac{h^2}{h^2+2}\right]^{d}}.$$

It follows that for fixed $d$, the binning index satisfies

$$(h/2)\mathbb{B}_h \to 1 \text{ as } h \to 0.$$

The DOF of the *uncentered* kernel satisfies

$$\mathrm{DOF}_{\mathrm{unc}}^{1/\mathrm{d}} = \left(1 + \frac{4}{h^2}\right)^{1/2}.$$

*Proof.* Given in the online supplementary materials. □

The corollary provides some heuristic insights into the relationship of DOF and the bandwidth. If we think of DOF as equating to a number of chi-squared bins, then this last approximation for $\mathbb{B}_h$ indicates that $2/h$, for a standard normal, can be interpreted as the number of bins in that dimension, with $h = 1$ thereby being roughly equivalent to a split into two bins on that coordinate. For small $h$ the binning index based on the uncentered kernel, which is much simpler, gives an equivalent approximation.

## 5.4 Properties of the Pearson Normal Kernel

Since the Pearson-normal kernel is new, we start by gaining a deeper understanding of its structure. One very nice feature in the testing of multivariate normality is its simple and elegant eigenanalysis.

*Theorem 1.* The univariate Pearson-normal kernel $K$ has a spectral decomposition under the null standard normal density represented by

$$K_\rho(x, y) = \sum \frac{\rho^k H_k(x) H_k(y)}{k!},$$

where $H_k(x)$ is the $k$th Hermitian polynomial and $\rho = 1/(1 + h^2)$.

*Proof.* Given in the online supplementary materials. □

We note that the spectral decomposition has the same eigenfunctions for every value of $\rho$, and hence for every value of $h^2$. The following corollary is a consequence of this fact.

*Corollary 3.* There is a diffusion equation for the Pearson-normal kernel given as

$$\int K_{\rho_1}(x, z) K_{\rho_2}(z, y) n_1(z, 0) dz = K_{\rho_1 \rho_2}(x, y).$$

Thus, the Pearson-normal kernel is our second example of a diffusion kernel. Additionally, the following corollary provides the DOF associated with the Pearson-normal kernel.

*Corollary 4.* Let the distribution $G$ be the uncorrelated multivariate normal with zero mean vector. Then, in the univariate case, the DOF of the Pearson-normal kernel are given as

$$\mathrm{DOF}(K) = \frac{1+\rho}{1-\rho} = 1 + 2h^{-2},$$

where $\rho = 1/(1 + h^2)$, and that of the centered kernel are the same

$$\mathrm{DOF}(K^{\mathrm{ctr}}) = \frac{[(1-\rho)^{-1} - 1]^2}{(1-\rho^2)^{-1} - 1} = \frac{\rho^2/(1-\rho)^2}{\rho^2/(1-\rho^2)} = 1 + 2h^{-2}.$$

In the multivariate case, the DOF of the centered kernel are given by the formula

$$\mathrm{DOF}(K^{\mathrm{ctr}}) = \frac{[(1-\rho)^{-d} - 1]^2}{(1-\rho^2)^{-d} - 1}$$
$$= \frac{[1 - (1-\rho)^d]^2}{\{1-\rho\}^d \times \{(1+\rho)^{-d} - (1-\rho)^d\}}.$$

It follows that as $\rho$ goes to 1 (and $h^2 \to 0$) the binning index satisfies

$$(1-\rho)\mathbb{B}_h \to 2.$$

Since the binning index is therefore approximately $2h^{-2}$for $h$ small, unlike the Gaussian $2h^{-1}$, we can see that the test has much larger DOF for small $h$ than does the Gaussian kernel.

Recall that the Pearsonized kernels were designed to increase testing sensitivity for model deficiencies in regions of low null density. When one does this, there is a risk of inflating testing variability. Our first glimpse of this comes in the binning index calculation above. We will also show its disappointing behavior in our null simulation study. Our heuristic explanation for this is that in the pursuit of improved sensitivity to alternatives, we lost ground in both the null variance and in the null sampling behavior. The following theorem gives a partial explanation for the latter phenomenon through the heavy tails of the null sampling distribution.

*Theorem 2.* Let $X_1, X_2, \ldots, X_n$ be independent $d$-dimensional random vectors. Then the statistic $U_n$ formed by using the $d$-dimensional Pearson-normal kernel can be expressed as

$$U_n = \frac{1}{n(n-1)} \cdot \frac{(1+h^2)^d}{h^d(h^2+2)^{d/2}}$$
$$\times \sum_{i=1}^{n} \sum_{j \neq i}^{n} \exp\left[\frac{1}{2(h^2+2)} W_{ij} - \frac{1}{2h^2} Y_{ij}\right],$$

where

$$W_{ij} = \sum_{l=1}^{d} \left(\frac{X_{il} + X_{jl}}{\sqrt{2}}\right)^2,$$

and

$$Y_{ij} = \sum_{l=1}^{d} \left(\frac{X_{jl} - X_{il}}{\sqrt{2}}\right)^2.$$

Under the normal null hypothesis $W_{ij}$, $Y_{ij}$ are independent random variables from a chi-squared distribution with $d$ DOF. It follows that the third and higher moments of the distribution of the statistic $U_n$ do not exist.

*Proof.* Given in the online supplementary materials. □

## 6. TESTING FOR MULTIVARIATE NORMALITY: A DETAILED STUDY

We have now developed a set of tools for analysis of the power of a quadratic test. We have also identified the key structures of the Gaussian and Pearson-normal kernel. We now delve into a more detailed study of power properties when testing the null hypothesis that $G$ is standard multivariate normal. As noted earlier, this setting was chosen because we can carry out all necessary calculations analytically and thereby avoid simulation or numerical integration errors.

### 6.1 Bowman and Foster Tests

To give further context to our study, we bring up some background. Bowman and Foster (1993) studied tests of multivariate normality. Their *integrated squared error* statistic was defined as

$$\mathrm{ISE} = \int [f(x) - \hat{f}(x)]^2 dx,$$

where $f(x)$ indicates a multivariate normal distribution with zero mean vector and covariance matrix $(1+h^2)I_d$, and $\hat{f}(x)$ indicates a density estimator constructed by using the normal kernel. This distance is identical to the quadratic distance test we will be using. They indicate that they chose the smoothing parameter based on

$$h = \left(\frac{4}{(d+2)n}\right)^{\frac{1}{d+4}}, \tag{8}$$

where $d$ is the dimension of the normal model. This is the optimal tuning parameter for MSE when the true density is normal (Bowman and Foster 1993, p. 535).

Bowman and Foster (1993) reported, via simulation, excellent power results for the ISE test procedure. It outperformed four competitors in power in a crossed study design that included two distributional types, a bimodal normal mixture and a gamma(2,1) distribution, data dimensions from $d = 1$ up to 6, and sample sizes $n$ of 25, 50, and 100. It is therefore of interest to see if their chosen bandwidth is similar to the optimal bandwidth for testing.

### 6.2 The Study Design

One of our main goals here was to evaluate the quality of the surrogate power function. To do so, we accompanied our theoretical analysis with a painstaking simulation study. We also sought to evaluate the role of dimension of the data in the choice of the binning index, so our design goes up to 16 dimensions.

To accomplish these ends with greater clarity, we designed our study as follows:

- First, we assumed that the null normal model has known parameters. This removes from our investigation the complicating question of the role of parameter estimation in a composite null. Simulation of the size under the null can then be easily compared with the surrogate power approximation.
- We used a midpower analysis with a variety of alternative families to see how stable the choice of bandwidth was over alternatives and sample sizes and how the choice depends on data dimension. We also compared the surrogate power with simulated power.
- We chose to assess power using target alternatives based on mixture models. Certainly our methods could also be used to evaluate this test, and select bandwidth parameters, when the alternatives, based on scientific needs, were heavier tailed than normal or even marginally normal with dependency structures. However, the midpower lemma suggests that one might wish to use a different kernel distance in those cases. We used midpower sensitivity analysis to measure how sensitive the possible tests were to these mixture alternatives.
- We considered only normal kernels and Pearson normal kernels with bandwith matrix $hI$, so that the issue of best power could be resolved with the choice of a single parameter. If the null variance matrix $\Sigma$ were unknown, but $\Sigma$ was known under a fixed null hypothesis, we would consider a bandwidth matrix of the form $h\Sigma$. This would create a test procedure that was affine invariant, and so have a single distribution for the compound null hypothesis.

Our midpower study had the following design parameters. The dimensions $d$ were 2, 4, 8, and 16. The sample sizes were 200, 500, 750, and 1000. We did work with larger samples as well, looking at sample sizes of 10,000, 100,000, or larger for dimensions 8 and 16. All programs were written in C++.

Our choice of bandwidths $h$ were dictated by the need to find the midoptimal bandwidth for each midpower problem. These results could have been reported in terms of DOF($h$), a more universal measure, but the near exponential growth in optimal DOF with dimension made it more useful to express the bandwidths in terms of the binning index $\mathbb{B}(h) = (\mathrm{DOF}(h)+1)^{1/d}$. Note that for the normal kernel, the approximation given in Theorem 1 means that the binning index is approximately $\mathbb{B}(h) \approx 2/h$ for the normal kernel. For the Pearson-normal, the binning index is approximately $2/h^2$.

The null hypothesis was standard multivariate normal. In our study design the alternatives were chosen to be multicoordinate mixtures in which all the coordinates are mutually independent. As noted earlier, this is ideal for the product kernels under consideration. Some of the coordinates were set to be standard normal, while the remaining coordinates, $m$ in number, were nonnormal. We varied $m$ as a design parameter. Each of the nonnormal coordinates was a mixture of univariate normals, either of the symmetric form $0.5N(a, 1-a^2) + 0.5N(-a, 1-a^2)$ for $a$ between 0 and 1, or the asymmetric form $0.75N(a, 1-3a^2) + 0.25N(-3a, 1-3a^2)$ for $a$ between 0 and $\sqrt{1/3}$. For a fixed value of $m$, and with the symmetric or asymmetric nature of the mixture fixed, the parameter $a$ will be treated as the univariate parameter for the midpower analysis ($\delta$ in our earlier notation).

We let the number of nonnormal coordinates be $m = 1, 2, 4$, and 8 in $d = 8$ dimensions, but considered only $m = 1$ in $d = 2$ and 4 dimensions.

*Remark 4*. Note that the mixtures in each coordinate are designed so that the mean of $X$ is zero and the variance is 1, regardless of $a$. When this is so, the first and second moments no longer contain information about the alternative. By doing so we are emulating the situation where the first and second moments are estimated under the null, and so contain no lack-of-fit information.

In the alternatives we considered, the symmetric mixture alternatives also agree with the null in the third moment, but the asymmetric do not. Thus, we expect to see some difference in the midpower analysis that depends on $m$ and on the symmetry of the mixture.

### 6.3 The Null Hypothesis Distribution

We created a Monte Carlo sample to find the null distribution for all of our design settings and used it to find critical values for the normal kernel test and Pearson normal test. Given that we will later do a simulated power, we felt it necessary to understand the true size of the test based on $z_\alpha$.

To identify the empirical 95th quantile of the distribution of the test statistic under the null hypothesis, we generated $r = 1000$ replicate samples from a multivariate normal distribution with mean vector zero and identity covariance matrix. For each sample, we computed the distance between the empirical cumulative distribution function $\hat{F}$ and the $\text{MVN}_d(\mathbf{0}, I)$ model. We then order the values of the test statistic from smallest to largest. The empirical cutoff value was the 95th quantile of the above list of numbers, and the standard error for estimating the nominal 0.05 level was 0.00218. To compute the simulated power of the tests, we simply counted the number of model rejections in our set of $r = 1000$ replicates at the given settings.

For the normal kernel, the results were very consistent over $d$, $\mathbb{B}(h)$, and $n$. If one used the normal theory critical value $z_\alpha$ with $\alpha = 0.05$, the simulated sizes of the distance test ranged from 0.05 to 0.08. After accounting for the simulation error involved, the results were largely consistent with the existence of a constant size of 0.065 across all cases. Thus, for the following power analysis, it would be more reasonable to say the power curves were for tests with true size of about 0.065. We think this should have little effect on the choice of the best DOF, however, or the relative sensitivity of the tests.

For the Pearson-normal kernel, however, we found some disturbing results. We found that repeated simulations gave very different estimated critical values. This is due to the wide dispersion of the largest-order statistics. We think an explanation for this is, as we noted earlier, that the third moment of the test statistic fails to exist under the null hypothesis. We were not aware of this initially, as the first and second moments do exist, and so all of our theoretical calculations could be carried out. Just the same, the normal distribution for the null hypothesis distribution is a very, very bad approximation in the tails, where the critical value is determined.

### 6.4 Theoretical and Simulated Power

We start with the surrogate power function given in Section 3.1. We first contrast two rather different viewpoints of power in our problem. One viewpoint takes a fixed alternative $a$, and plots $\beta_h(a)$ as a function of $h$. (In our case, we replaced $h$ with the binning index $\mathbb{B}(h)$ so as to relate it to the number of chi-squared bins.) The right end of these plots, with larger $\mathbb{B}$, represents small bandwidths, and so greater variance in the test statistics. These curves have a mound shape showing that for bandwidths that are either too small or too large, the power is quite low, and that it is important to select bandwidth/DOF carefully (see Figure 1).

We can also view $\beta_h(a)$ as a function of $a$, as needed to do the midpower analysis. For a fixed binning index, we can plot this curve as a function of $a$. We can then overlay the curves for various values of $\mathbb{B}$ on a common plot, and then choose the first one to cross 0.5 as the midoptimal binning index. Such a plot is found in Figure 2. All of these plots showed that the midoptimal bandwidth was optimal more widely than the power 0.50.

One of our key goals was to demonstrate that these theoretical plots are useful guides for choosing bandwidths. We therefore carried out a large number of simulations at a variety of alternatives, values of binning index $\mathbb{B}$, and sample sizes. The plots for the normal kernel were consistent in showing that the power approximation was slightly larger than the simulated power, with slightly greater error for $\mathbb{B}$ small. (This is consistent with the enlargement of size that we found.) Based on these plots, we feel very confident in saying that choosing $\mathbb{B}$ based on the power

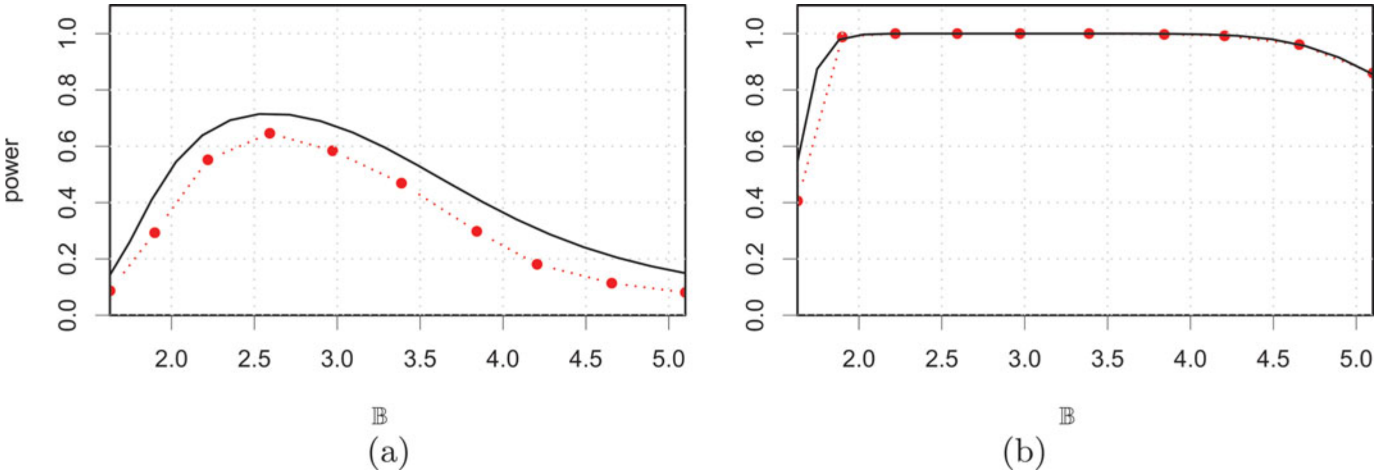

Figure 1. Theoretical power (solid black) overlaid on simulated power (dotted green) for coordinate mixtures with alternative $a = 0.8$. The power curve is plotted against the binning index $\mathbb{B}$ for (a) $n = 200$ in eight dimensions, and (b) $n = 750$ in two dimensions.

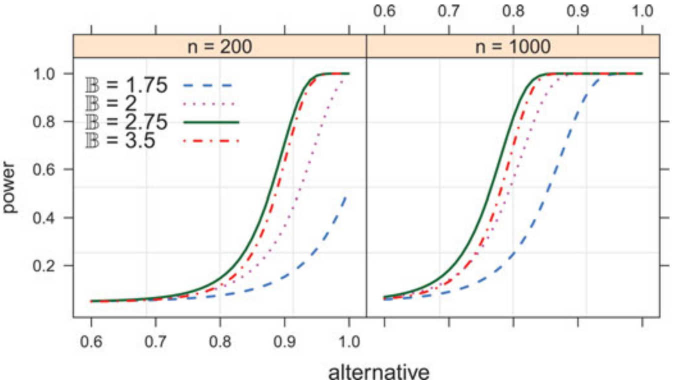

Figure 2. Power curves for symmetric eight-dimensional ($m = 2$ mixtures) plotted against alternatives (x-axis) for four distinct 𝔹 values near the optimal value with sample sizes 200 and 1000.

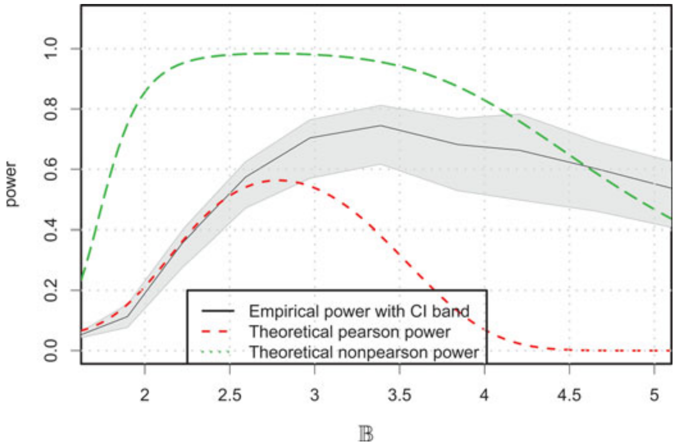

Figure 3. Comparison of theoretical power with normal kernel (dotted green) and Pearson kernel (dotted red) with the empirical power given by the solid line along with its 95% confidence band.

approximation gives a value that would lead to true midpower being very nearly optimal.

For the Pearson-normal kernel, when we carried out the simulations under the alternatives, the statistic was somewhat better behaved than under the null, and so we found it reasonable to show a family of empirical power curves as they depended on the bin index and on the endpoints of the nonparametric confidence interval for the 0.05 quantile (see Figure 3).

This figure shows that the normal kernel has a significantly higher power curve than the Pearson kernel, no matter the chosen critical value. It also shows that the surrogate curve and the simulated curves for the Pearson kernel agreed very well for larger bandwidths, but failed for smaller ones, possibly reflecting the overall challenges that arise in using this kernel.

The Pearson-normal kernel had an unstable null distribution and relatively poor sensitivity against mixture alternatives, leading us to conclude that the normal kernel was superior for such targets. Although we do not investigate it further here, it could be superior for other target alternatives, such as heavy-tailed distributions.

### 6.5 The Role of $h$, DOF, and 𝔹

We concentrate hereafter on the Gaussian kernel due to its superior performance. Table 1 summarizes, for the symmetric alternatives, the results of our search for the value of 𝔹 that gives optimal power, expressed as a function of the data dimension, sample size, and alternative family considered. Additionally, Table 2 and Table S-1 in the online supplementary materials present results, for the same symmetric alternatives, for large sample sizes and for dimension $d = 16$, respectively. From these three tables, we can draw the following general observations.

The optimal binning indices/DOF have the nice property that they depend very little on sample size, and so we can summarize their behavior based on other indicators. (See also Figure S-1 and Table S-1 in the online supplementary materials.) From Table 1 it seems that for a fixed dimension $d$, there is little dependence on the $m$, the number of nonnormal coordinates; this becomes even more clear in the large samples of Table 2. This is a good feature, as the value of $m$ would not be known in advance. Overall, the dimension $d$ of the data, which is known to the user, appears to be the single, most important factor in determining the optimal value of 𝔹.

Table 3 presents simulation results for the asymmetric data case for sample sizes 200, 500, 750, and 1000 and dimensions 2, 4, and 8. Here again we see that sample size has a relatively weak effect on optimal bin width. Since $m = d$ in all these examples, there is no information on the effect of the number of nonnormal coordinates. Once again dimension, a user-known quantity, is the key determinant in the optimal choice.

Table 1. Optimum 𝔹 for symmetric mixture alternatives

| | | $n = 200$ | | $n = 500$ | | $n = 750$ | | $n = 1000$ | |
|---|---|---|---|---|---|---|---|---|---|
| Dimension | $m$ | 𝔹 | DOF | 𝔹 | DOF | 𝔹 | DOF | 𝔹 | DOF |
| 8 | 1 | 2.875 | 4671 | 2.750 | 3268 | 2.750 | 3268 | 2.750 | 3268 |
| | 2 | 2.750* | 3268 | 2.750 | 3268 | 2.625 | 2256 | 2.625* | 2256 |
| | 4 | 2.625 | 2256 | 2.625 | 2256 | 2.5 | 1526 | 2.5 | 1526 |
| | 8 | 2.5 | 1526 | 2.5 | 1526 | 2.5 | 1526 | 2.5 | 1526 |
| 4 | 1 | 3.756 | 198 | 3.505 | 150 | 3.505 | 150 | 3.505 | 150 |
| | 2 | 3.505 | 150 | 3.505 | 150 | 3.26 | 112 | 3.26 | 112 |
| | 4 | 3.26 | 112 | 3.26 | 112 | 3.26 | 112 | 3.001 | 81 |
| 2 | 1 | 5.099 | 25 | 4.58 | 20 | 4.58 | 20 | 4.58 | 20 |
| | 2 | 4.58 | 20 | 4.58 | 20 | 4.58 | 20 | 4.58 | 20 |

*Cases plotted in Figure 2.

Table 2. Optimum $\mathbb{B}$ for symmetric mixture alternatives (eight dimensions) with larger $n$'s

| | $n = 10e4$ | | $n = 10e5$ | | $n = 10e6$ | | $n = 10e7$ | |
|---|---|---|---|---|---|---|---|---|
| $m$ | $\mathbb{B}$ | DOF | $\mathbb{B}$ | DOF | $\mathbb{B}$ | DOF | $\mathbb{B}$ | DOF |
| 1 | 2.5 | 1526 | 2.5 | 1526 | 2.375 | 1011 | 2.375 | 1011 |
| 2 | 2.5 | 1526 | 2.375 | 1011 | 2.375 | 1011 | 2.375 | 1011 |
| 4 | 2.5 | 1526 | 2.375 | 1011 | 2.375 | 1011 | 2.375 | 1011 |
| 8 | 2.5 | 1526 | 2.375 | 1011 | 2.375 | 1011 | 2.375 | 1011 |

If we compare Table 1 (with $m = d$) with Table 3, we can see that the optimal bandwidth does depend mildly on the type of alternative; one needs a slightly smaller bandwidth (and larger binning index) for power against asymmetric mixtures. One might wish to choose a compromise bandwidth in this setting.

## 6.6 Comparison With Bowman–Foster

One of our key points in this article is that if kernel density estimators are used in testing, the old rules about selecting bandwidth are no longer valid. Indeed, it can be seen from our analyses that for the alternatives we have chosen, the optimal binning index converged to a constant. This means that the bandwidth $h$ does not go to zero, in contrast with the standard MSE theory. In this section we briefly compare our results with the results we would obtain if using the optimal MSE bandwidth for the problem.

Table S-2 of the online supplementary materials presents a comparison of our DOF calculation with that by Bowman and Foster (1993). We convert the bandwidth parameter suggested by Bowman and Foster to DOF for comparison and called it the BF-DOF. The table provides a numerical comparison over the cases we consider. In Figure 4, we show how the power is reduced for the BF-DOF in two special cases.

It is particularly striking how much larger the BF-DOF is in higher dimensions, corresponding to smaller bandwidths. For $n = 10{,}000$ the binning index is about twice as large, putting it into the range of very low power. There is a simple explanation. Our DOF calculation is based on the variability of the test statistic, not its bias. The fact is that the null model and the data have the same kernel smoothing and so there is no bias. However, the BF-DOF is based on the mean squared error and it is clear that to reduce bias squared in higher dimensions, a very small bandwidth is required.

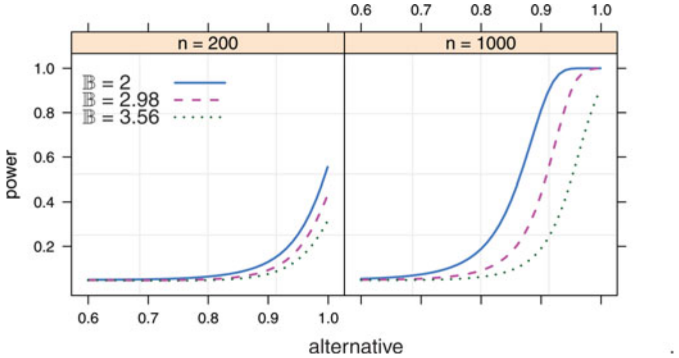

Figure 4. Comparison of power using Bowman–Foster DOF or its implied $\mathbb{B}$ (2.98 for $n = 200$ and 3.56 for $n = 1000$) versus the optimal $\mathbb{B}$ (2 for all sample sizes) for 16 dimensions with $m = 2$.

## 6.7 A Sensitivity Analysis

To compare the sensitivity of the tests across our targeted alternative families, we created the following measure of the disagreement between null and alternative. Given a finite mixture of multivariate normals $\sum_k \pi_k n_V(x, \mu_k)$ in which the component distributions have a common covariance matrix $V$, we can write the variance of the random vector $X$ as

$$\mathrm{var}(X) = \mathrm{var}(D) + V,$$

where $D$ is the discrete distribution with mass $\pi_k$ at $\mu_k$. Thus, we can measure the strength of the mixing distribution, relative to the normal errors, via the *root variance ratio*

$$M^* = \sqrt{\mathrm{tr}(\mathrm{var}(D) \cdot V^{-1})}.$$

This measure is zero if there is no mixing, and otherwise it compares how much of the total variance of $X$ is due to mixing and how much is due to normal errors.

For a second interpretation, note that if there are just two mixture components, so the density is $\pi_1 n_V(x, \mu_1) + \pi_2 n_V(x, \mu_2)$, then

$$M^* = \sqrt{\pi_1 \pi_2}\sqrt{(\mu_1 - \mu_2)' V^{-1}(\mu_1 - \mu_2)},$$

where the second square root is the Mahalanobis distance between the two component distributions. We note that for the symmetric case, $M^* \leq 1$ means the density for $X$ is unimodal whereas for $M^* > 1$, it is bimodal (Ray and Lindsay 2005). In our class of mixture alternatives, we will transform the parameter $a$ in each coordinate to the corresponding value of $M^*$. We will consider it a benchmark for *high* sensitivity if the test is sensitive enough to detect a unimodal alternative. That is, if the midoptimal $M^*$ is 1 or smaller. We will call the sensitivity *inadequate* if the test cannot detect $M^* = 2$, which is a well-separated mixture.

Tables 4 and 5 present sensitivity results for the high-dimensional cases. In the symmetric cases with $d = 8$, there is a clear increase in sensitivity as either $m$ increases or $n$

Table 3. Optimum $\mathbb{B}$ for asymmetric mixture alternatives

| | $n = 200$ | | $n = 500$ | | $n = 750$ | | $n = 1000$ | |
|---|---|---|---|---|---|---|---|---|
| Dimension | $\mathbb{B}$ | DOF | $\mathbb{B}$ | DOF | $\mathbb{B}$ | DOF | $\mathbb{B}$ | DOF |
| 8 | 3.00 | 6567 | 2.875 | 4671 | 2.875 | 4671 | 2.75 | 3268 |
| 4 | 4.004 | 256 | 3.755 | 198 | 3.755 | 198 | 3.755 | 198 |
| 2 | 4.590 | 30 | 5.099 | 25 | 5.099 | 25 | 5.099 | 25 |

Table 4. Sensitivity analysis for symmetric case with dimension 8

| $m$ | $n = 200$ | $n = 750$ | $n = 1000$ |
|---|---|---|---|
| 1 | 2.35 | 1.49 | 1.38 |
| 2 | 1.76 | 1.25 | 1.17 |
| 4 | 1.43 | 1.07 | 1.02 |
| 8 | 1.25 | 0.95 | 0.85 |

Table 5. Sensitivity analysis for asymmetric case in dimensions 8 and 16

| $d$ | $n = 200$ | $n = 500$ | $n = 750$ | $n = 1000$ | $n = 10{,}000$ | $n = 1{,}00{,}000$ |
|---|---|---|---|---|---|---|
| 8 | 1.70 | 1.25 | 1.12 | 1.04 | 0.61 | 0.39 |
| 16 | 4.42 | 1.91 | 1.58 | 1.41 | 0.75 | 0.47 |

increases. However, high sensitivity was not attained even at $n = 1000$ when $m$ was fixed at one. Indeed at that sample size, we had this degree of sensitivity only for $m = 8$. On the other hand, we did not have inadequate sensitivity except when $n = 200$ and $m = 1$. In the asymmetric cases, Table 5, where $m$ was fixed at $d$, we can see the interplay between sensitivity of the test, dimension $d$, and sample size $n$. Very poor sensitivity occurred only for $n = 200$ and $d = 16$, but high sensitivity only occurred for sample sizes above $n = 1000$.

### 6.8 Discussion of the Analysis

Determining whether a distribution is a normal or a mixture of normals is actually quite hard when the two components are not well separated. This difficulty is compounded in higher dimensions, when, for $m = 1$, one is searched through $d$-dimensional space for the one coordinate with a signal. Thus, we do not find our sensitivity results disappointing, especially in the light of earlier results that ascribed good power to the normal kernel test relative to other tests. Certainly the midpower lemma suggests it should do well at detecting lumps in the normal distribution. Our comparison of the normal and Pearson normal kernels suggests that attempts to increase sensitivity of the distance $\mathbb{D}(F, G)$ could backfire by causing greater variance under the null and so decrease the noncentrality index. One route around this might be to build a kernel more specifically tuned to the mixture problem.

## 7. CONCLUDING DISCUSSION

Our focus in this article has been on building a set of straightforward tools that a methodologist could use to build kernel-based goodness-of-fit procedures for a wide variety of problems. One goal was to make the analyses as simple as possible, especially emphasizing the avoidance of the need to work with infinite-dimensional noncentral distributions. We wanted to go beyond the rather crude guidance given by the DOF of the test, and indeed also better understand how one could choose DOF in higher dimensions. We think we have, through the study of the normal bandwidth problem, given some evidence that a simple surrogate power function approach will work more insightfully than simulation studies.

In the process of creating this analysis, it became clear that it is imperative that one focus on local alternatives, the ones that give extra power where it is needed. We have done so by identifying the midpower alternative, and shown that this yields a lovely reduction of the surrogate power function to a single number summary, the noncentrality index. This reduction enabled us to take a further step, through the midpower lemma, and create a new tool for building kernels with power targeted to a class of alternatives.

The midpower lemma is just a starting point, however, for building a good kernel-based procedure. Just as in nonparametric density estimation, there is a need to have procedures that have a bandwidth parameter that allows one to tune the procedure to the changing landscape as the sample size changes. We have therefore provided a recipe, the midpower analysis, for choosing the best bandwidth parameter $h$ when considering a one parameter family of alternatives. If one uses careful thought as to the alternatives that are truly of interest, then one can gain some insights about the behavior of the test through changes in dimension and sample size. In our example, we were interested in the ability to detect multimodal mixtures, and in particular how sensitive the normal kernel test would be.

Along the way we converted DOF to the binning index to better account for the effects of dimension. In our selected examples, we found that the data dimension was the main factor in choosing the optimal DOF. Our analysis showed that if traditional rules from the density estimation literature are used to obtain $h$ (and hence the DOF), the resulting tests will have suboptimal power. A detailed simulation study verified that our proposed analysis, based on a number of simplifications, gave reasonable answers without the burden of an extensive simulation.

We think of this article as offering guidance to a methodologist who might be working on a particular goodness-of-fit problem. Indeed, we intend to use it ourselves in more focused investigations. As to someone who is seeking advice for a test for multivariate normality, our results directly give guidance as to the sensitivity of the test for mixture models. We also think that the normal kernel test is currently the best option for high-dimensional testing of multivariate normality. However, we do think that the very rich set of possibilities for kernel construction, including the tools we describe here, will lead to even better methodologies in the future. For example, the works by Hui and Lindsay (2010) and Lindsay and Yao (2012) point to the promise of a matrix extension of the quadratic distance methodology that is useful for testing for normality and independence. This allows for additional diagnostics for the lack-of-fit problem, including finding linear combinations of the data that fit poorly.

## APPENDIX: PROOF OF PROPOSITION 2

To carry out this expansion, we let $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\hat{F}$, then create a function of $\varepsilon$ defined by $h(\varepsilon) = D(F_\varepsilon, G_{\hat{\theta}(F_\varepsilon)})$, noting that for a consistent estimator, with $\hat{\theta}(G_\theta) = \theta$, then $h(0) = D(F, G_{\theta(F)})$ and $h(1)$ is our test statistic $D(\hat{F}, G_{\hat{\theta}})$. We will use the one-term approximation

$$h(1) - h(0) \approx \varepsilon h'(0)|_{\varepsilon=1},$$

when $h'(0)$ is not zero. Otherwise we will use

$$h(1) - h(0) \approx \varepsilon h'(0) + 2^{-1}\varepsilon^2 h''(0)|_{\varepsilon=1} = 2^{-1}h''(0).$$

We have

$$\begin{aligned}\frac{d}{d\varepsilon}\mathbb{D}(F_\varepsilon, G_{\hat{\theta}(F_\varepsilon)}) &= \frac{d}{d\varepsilon}\int [k^{\text{ctr}}(\theta(F_\varepsilon), X, t)dF_\varepsilon(X)]^2 du(t)\\ &= 2\int \left[\int [k^{\text{ctr}}(\theta(F_\varepsilon), X, t)dF_\varepsilon(X)]\right]\\ &\quad \times \left[\frac{d}{d\varepsilon}\int k^{\text{ctr}}(\theta_\varepsilon, X, t)dF_\varepsilon(X)\right] du(t),\end{aligned}$$

where

$$\frac{d}{d\varepsilon}\int k^{\mathrm{ctr}}(\theta_\varepsilon, X, t)\,dF_\varepsilon(X) = \int \theta_\varepsilon'^{T}\,\nabla_\theta k^{\mathrm{ctr}}(\theta(F_\varepsilon), Y, t)\,dF_\varepsilon + \int k^{\mathrm{ctr}}(\theta(F_\varepsilon), X, t)\,d(\hat{F} - F).$$

This gives

$$h'(0) = 2\iint v_F(x, t)[k^{\mathrm{ctr}}(\theta, F, t)]\,du(t)\,d\hat{F}(x).$$

Since $k^{\mathrm{ctr}}(\theta, F, t) = k^{\mathrm{ctr}}(\theta, G_\theta, t) = 0$ under the model, we take one more term in the expansion to find

$$\frac{d^2}{d\varepsilon^2} D(F_\varepsilon, G_{\hat{\theta}(F_\varepsilon)})|_{\varepsilon=0} = 2\int \left[\frac{d}{d\varepsilon}\int k^{\mathrm{ctr}}(\theta_\varepsilon, X, t)\,dF_\varepsilon(X)\right]^2 du(t)|_{\varepsilon=0} = 2\int\left[\int v(\theta, X, t)\,d\hat{F}(x)\right]^2 du(t).$$

## SUPPLEMENTARY MATERIALS

The supplement to the article contains an extensive literature review and proofs of several results that appear in this article. It also contains a section on the relationship between the centered and uncentered DOF. Additionally, the supplementary materials contains tables and figures related to the simulation studies presented in Section 6 of the main article.

*[Received October 2011. Revised March 2013.]*

## REFERENCES

Anderson, N. H., Hall, P., and Titterington, D. M. (1994), "Two-Sample Test Statistics for Measuring Discrepancies Between Two Multivariate Probability Density Functions Using Kernel-Based Density Estimates," *Journal of Multivariate Analysis*, 50, 41–54. [399]

Biau, G., and Gyorfi, L. (2005), "On the Asymptotic Properties of a Nonparametric l1-Test Statistic of Homogeneity," *IEEE Transactions on Information Theory*, 51, 3965–3973. [399]

Bowman, A. W., and Foster, P. J. (1993), "Adaptive Smoothing and Density-Based Tests for Multivariate Normality," *Journal of the American Statistical Association*, 88, 529–537. [405,408]

Friedman, J., and Rafsky, L. (1979), "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests," *The Annals of Statistics*, 7, 697–717. [399]

Gourieroux, C., and Tenreiro, C. (2001), "Local Power Properties of Kernel Based Goodness of Fit Tests," *Journal of Multivariate Analysis*, 78, 161–190. [399]

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012), "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, 13, 723–773. [399]

Hui, G., and Lindsay, B. (2010), "Projection Pursuit via White Noise Matrices," *Sankhya B*, 72, 123–153. [409]

Kondor, R. I., and Lafferty, J. D. (2002), "Diffusion Kernels on Graphs and Other Discrete Input Spaces," in *ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, pp. 315–322. [397]

Lafferty, J., and Lebanon, G. (2005), "Diffusion Kernels on Statistical Manifolds," *Journal of Machine Learning Research*, 6, 129–163. [397]

Lindsay, B. G., Markatou, M., Ray, S., Yang, K., and Chen, S.-C. (2008), "Quadratic Distances on Probabilities: A Unified Foundation," *The Annals of Statistics*, 36, 983–1006. [395]

Lindsay, B. G., and Yao, W. (2012), "Fisher Information Matrix: A Tool for Dimension Reduction, Projection Pursuit, Independent Component Analysis, and More," *Canadian Journal of Statistics*, 40, 712–730. [409]

McManus, D. A. (1991), "Who Invented Local Power Analysis?" *Econometric Theory*, 7, 265–268. [399]

Neyman, J. (1937), "'Smooth Test' for Goodness of Fit," *Scandinavian Actuarial Journal*, 20, 149–199. [399]

Pitman, E. J. G. (1948), "Lectures on Non-Parametric Inference," Columbia University, Spring Semester. [399]

Poczos, B., Ghahramani, Z., and Schneider, J. (2012), "Copula-Based Kernel Dependency Measures," in *Proceedings of the 29th International Conference on Machine Learning*, New York: Omnipress, pp. 775–782. [399]

Ray, S., and Lindsay, B. (2005), "The Topography of Multivariate Normal Mixtures," *The Annals of Statistics*, 33, 2042–2065. [408]

Seo, B., and Lindsay, B. G. (2010), "A Computational Strategy for Doubly Smoothed MLE Exemplified in the Normal Mixture Model," *Computational Statistics and Data Analysis*, 54, 1930–1941. [401]

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Scholkopf, B., and Lanckriet, G. R. G. (2010), "Hilbert Space Embeddings and Metrics on Probability Measures," *Journal of Machine Learning Research*, 11, 1517–1561. [399]

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [397]