The Annals of Statistics
2016, Vol. 44, No. 2, 743–770
DOI: 10.1214/15-AOS1386

INFERENCE IN ADAPTIVE REGRESSION VIA THE KAC–RICE FORMULA

BY JONATHAN E. TAYLOR[*,1], JOSHUA R. LOFTUS[*,2] AND RYAN J. TIBSHIRANI[†,3]

Stanford University and Carnegie Mellon University†*

We derive an exact p-value for testing a global null hypothesis in a general adaptive regression setting. Our approach uses the Kac–Rice formula [as described in *Random Fields and Geometry* (2007) Springer, New York] applied to the problem of maximizing a Gaussian process. The resulting test statistic has a known distribution in finite samples, assuming Gaussian errors. We examine this test statistic in the case of the lasso, group lasso, principal components and matrix completion problems. For the lasso problem, our test relates closely to the recently proposed covariance test of Lockhart et al. [*Ann. Statist.* (2004) **42** 413–468].

In a few specific settings, our proposed tests will be less powerful than other previously known (and well-established) tests. However, it should be noted that the real strength of our proposal here is its generality. We provide a framework for constructing valid tests across a wide class of regularized regression problems, and as far as we can tell, such a unified view was not possible before this work.

1. Introduction. In this work, we consider the problem of finding the distribution of

$$\max_{\eta \in \mathcal{K}} \eta^T \varepsilon, \qquad \varepsilon \sim N(0, \Theta), \tag{1}$$

for a convex set $\mathcal{K} \subseteq \mathbb{R}^p$. In other words, we study a Gaussian process with a finite Karhunen–Loève expansion [1], restricted to a convex set in $\mathbb{R}^p$.

While this is a well-studied topic in the literature of Gaussian processes, our aim here is to describe an implicit formula for both the distribution of (1), as well as the almost surely unique maximizer

$$\eta^* = \operatorname*{argmax}_{\eta \in \mathcal{K}} \eta^T \varepsilon. \tag{2}$$

A main point of motivation underlying our work is the application of such a formula for inference in modern statistical estimation problems. We note that a similar

Received February 2015; revised September 2015.
[1]Supported by NSF Grant DMS-12-08857 and AFOSR Grant 113039.
[2]Supported by NIH Grant 5T32GMD96982.
[3]Supported by NSF Grant DMS-13-09174.
MSC2010 subject classifications. Primary 62M40; secondary 62J05.
Key words and phrases. Gaussian processes, convex analysis, regularized regression.

(albeit simpler) formula has proven useful in problems related to sparse regression [10, 15]. Though the general setting considered in this paper is ultimately much more broad, we begin by discussing the sparse regression case.

1.1. *Example*: *The lasso*. As a preview, consider the ℓ_1 penalized regression problem, that is, the lasso problem [19], of the form

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{3}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\lambda \geq 0$. Very mild conditions on the predictor matrix X ensure uniqueness of the lasso solution $\hat{\beta}$; see, for example, Tibshirani [21]. Treating X as fixed, we assume that the outcome y satisfies

$$y \sim N(X\beta_0, \Sigma), \tag{4}$$

where $\beta_0 \in \mathbb{R}^p$ is some fixed true (unknown) coefficient vector, and $\Sigma \in \mathbb{R}^{p \times p}$ is a known covariance matrix. In the following sections, we derive a formula that enables a test of a global null hypothesis H_0 in a general regularized regression setting. Our main result, Theorem 1, can be applied to the lasso problem in order to test the null hypothesis $H_0 : \beta_0 = 0$. This test involves the quantity

$$\lambda_1 = \|X^T y\|_\infty,$$

which can be seen as the first knot (i.e., critical value) in the lasso solution path over the regularization parameter λ [7]. Recalling the duality of the ℓ_1 and ℓ_∞ norms, we can rewrite this quantity as

$$\lambda_1 = \max_{\eta \in \mathcal{K}} \eta^T (X^T y), \tag{5}$$

where $\mathcal{K} = \{\eta : \|\eta\|_1 \leq 1\}$, showing that λ_1 is of the form (1), with $\varepsilon = X^T y$ (which has mean zero under the null hypothesis). Assuming uniqueness of the entries of $X^T y$, the maximizer η^* in (5) is

$$\eta_j^* = \begin{cases} \operatorname{sign}(X_j^T y), & |X_j^T y| = \|X^T y\|_\infty, \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, \ldots, p.$$

Let j^* denote the maximizing index, so that $|X_{j^*}^T y| = \|X^T y\|_\infty$, and also $s^* = \operatorname{sign}(X_{j^*}^T y)$, $\Theta_{jk} = X_j^T \Sigma X_k$. To express our test statistic, we define

$$\mathcal{V}_{\eta^*}^- = \max_{\substack{s \in \{-1,1\}, k \neq j \\ 1 - s\Theta_{j^*k}/\Theta_{j^*j^*} > 0}} \frac{s(X_k - \Theta_{j^*k}/\Theta_{j^*j^*} X_{j^*})^T y}{1 - s\Theta_{j^*k}/\Theta_{j^*j^*}},$$

$$\mathcal{V}_{\eta^*}^+ = \min_{\substack{s \in \{-1,1\}, k \neq j \\ 1 - s\Theta_{j^*k}/\Theta_{j^*j^*} < 0}} \frac{s(X_k - \Theta_{j^*k}/\Theta_{j^*j^*} X_{j^*})^T y}{1 - s\Theta_{j^*k}/\Theta_{j^*j^*}}.$$

Then under $H_0 : \beta_0 = 0$, we prove that

$$\frac{\Phi(\mathcal{V}^+_{\eta^*}/\Theta^{1/2}_{j^*j^*}) - \Phi(\lambda_1/\Theta^{1/2}_{j^*j^*})}{\Phi(\mathcal{V}^+_{\eta^*}/\Theta^{1/2}_{j^*j^*}) - \Phi(\mathcal{V}^-_{\eta^*}/\Theta^{1/2}_{j^*j^*})} \sim \text{Unif}(0,1), \tag{6}$$

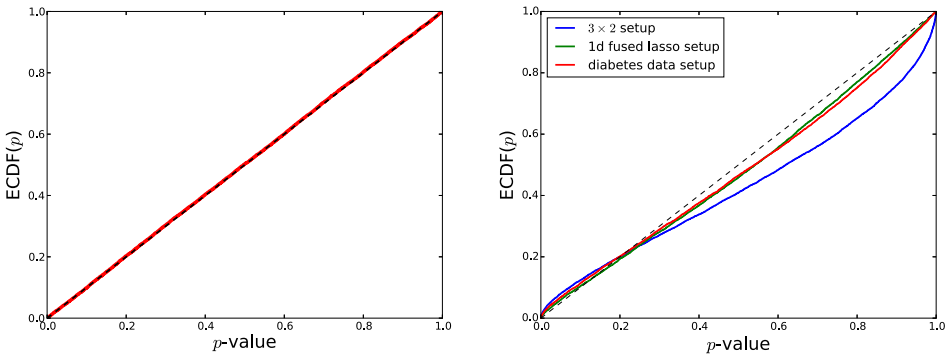
where Φ is the standard normal cumulative distribution function. This formula is somewhat remarkable, in that it is exact—not asymptotic in n, p—and relies only on the assumption of normality for y in (4) (with essentially no real restrictions on the predictor matrix X). As mentioned above, it is a special case of Theorem 1, the main result of this paper.

We refer to the above test statistic (6) as the *Kac–Rice test* for the lasso. It may seem complicated, but when the predictors are standardized, $\|X_j\|_2 = 1$ for $j = 1, \ldots, p$, and the observations are independent with (say) unit marginal variance, $\Sigma = I$, then $\mathcal{V}^-_{\eta^*}$ is equal to the second knot λ_2 in the lasso path and $\mathcal{V}^+_{\eta^*}$ is equal to ∞. Therefore, (6) simplifies to

$$\frac{1 - \Phi(\lambda_1)}{1 - \Phi(\lambda_2)} \sim \text{Unif}(0,1). \tag{7}$$

This statistic measures the relative sizes of λ_1 and λ_2, with values of $\lambda_1 \gg \lambda_2$ being evidence against the null hypothesis.

Figure 1(a) shows the empirical distribution function of a sample of 20,000 p-values from (6), over lasso problems with a variety of different predictor matrices,



(a) Kac-Rice test (b) Covariance test

FIG. 1. *The left panel shows the empirical distribution function of a sample of* 20,000 *p-values from* (6)*, coming from a variety of different lasso setups. The agreement with uniform here is excellent. The right panel shows the empirical distribution function of a sample of* 10,000 *covariance test p-values, computed using an* Exp(1) *approximation, using three different lasso setups. The* Exp(1) *approximation is generally conservative, whereas the Kac–Rice test is exact.*

all under the global null model $\beta_0 = 0$. In particular, for each sample, we drew the matrix X uniformly at random from the following cases:[4]

- small case: X is 3×2, with values (in row-major order) equal to $1, 2, \ldots, 6$;
- fat case: X is $100 \times 10{,}000$, with columns drawn from the compound symmetric Gaussian distribution having correlation 0.5;
- tall case: X is $10{,}000 \times 100$, with columns drawn from the compound symmetric Gaussian distribution having correlation 0.5;
- lower triangular case: X is 500×500, a lower triangular matrix of 1's (the lasso problem here is effectively a reparametrization of the 1-dimensional fused lasso problem [20]);
- diabetes data case: X is 442×10, the diabetes data set studied in Efron et al. [7].

As is seen in the plot, the agreement with uniform is very strong.

In their proposed *covariance test*, Lockhart et al. [10] show that under the global null hypothesis $H_0 : \beta_0 = 0$,

$$\lambda_1(\lambda_1 - \lambda_2) \xrightarrow{d} \mathrm{Exp}(1) \qquad \text{as } n, p \to \infty, \tag{8}$$

assuming standardized predictors, $\|X_j\|_2 = 1$ for $j = 1, \ldots, p$, independent errors in (4) with unit marginal variance, $\Sigma = I$, and a condition to ensure that λ_2 diverges to ∞ at a sufficient rate.

In finite samples, using Exp(1) as an approximation to the distribution of the covariance test statistic seems generally conservative, especially for smaller values of n and p. Figure 1(b) shows the empirical distribution function from 10,000 covariance test p-values, in three of the above scenarios. [The predictors were standardized before applying the covariance test, in all three cases; this is not necessary, as the covariance test can be adapted to the more general case of unstandardized predictors, but was done for simplicity, to match the form of the test as written in (8).] Even though the idea behind the covariance test can be conceivably extended to other regularized regression problems (outside of the lasso setting), the Exp(1) approximation to its distribution is generally inappropriate, as we will see in later examples. Our test, however, naturally extends to general regularization settings, allowing us to attack problems with more complex penalties such as the group lasso and nuclear norm penalties.

It is important now to emphasize the main goal of this paper. We provide a general recipe for constructing global tests in regularized regression problems. We do not claim these tests to be most powerful across all settings; indeed, in certain settings, our test will have less power than other well-studied alternative approaches. The strength of our proposal is that it applies so broadly, and furthermore, that it

[4]Here, we only drew the matrix X at random across the different cases in order to illustrate that this test holds for many different designs. For each fixed choice of design, the plot would look exactly the same.

leads to selective tests, beyond the global null, for many of these problems as well (pursued in subsequent papers after this one). Power and selective inference are the topics of Sections 6.2 and 6.3 of the discussion.

The rest of this paper is organized as follows. In Section 2, we describe the general framework for regularized regression problems that we consider, and a corresponding global null hypothesis of interest; we also state our main result, Theorem 1, which gives an exact p-value for this null hypothesis. The next two sections are then dedicated to proving Theorem 1. Section 3 characterizes the global maximizer (2) in terms of the related Gaussian process and its gradient. Section 4 applies the Kac–Rice formula to derive the joint distribution of the maximum value of the process (1) and its maximizer (2), which is ultimately used to derive the (uniform) distribution of our proposed test. In Section 5, we broadly consider practicalities associated with our test statistic, revisit the lasso problem and examine the group lasso, principal components and matrix completion problems as well. We end with a discussion in Section 6.

2. General regularized regression problems. We examine a class of regularized least squares problems of the form

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \cdot \mathcal{P}(\beta), \tag{9}$$

with outcome $y \in \mathbb{R}^n$, predictor matrix $X \in \mathbb{R}^{n \times p}$, and regularization parameter $\lambda \geq 0$. We assume that the penalty function $\mathcal{P}$ satisfies

$$\mathcal{P}(\beta) = \max_{u \in C} u^T \beta, \tag{10}$$

where $C \subseteq \mathbb{R}^p$ is a convex set, that is, $\mathcal{P}$ is the support function of C. This serves as a very general penalty, as we can represent any seminorm (and hence any norm) in this form with the proper choice of set C. We note that the solution $\hat{\beta} = \hat{\beta}_\lambda$ above is not necessarily unique (depending on the matrix X and set C) and the element notation used in (9) reflects this. A standard calculation, which we give in Section A.1 of the supplementary document [17], shows that the fitted value $X\hat{\beta}$ is always unique, and hence so is $\mathcal{P}(\hat{\beta})$.

Now define

$$\lambda_1 = \min\bigl\{\lambda \geq 0 : \mathcal{P}(\hat{\beta}_\lambda) = 0\bigr\}.$$

This is the smallest value of λ for which the penalty term in (9) is zero; any smaller value of the tuning parameter returns a nontrivial solution, according to the penalty. A straightforward calculation involving subgradients, which we give in Section A.2 of the supplement [17], shows that

$$\lambda_1 = \mathcal{Q}\bigl(X^T(I - P_{XC^\perp})y\bigr), \tag{11}$$

where $\mathcal{Q}$ is the dual seminorm of $\mathcal{P}$, that is, $\mathcal{Q}(\beta) = \max_{v \in C^\circ} v^T \beta$, the support function of the polar set C° of C. This set can be defined as $C^\circ = \{v : v^T u \leq 1 \text{ for all } u \in C\}$, or equivalently,

$$C^\circ = \{v : \mathcal{P}(v) \leq 1\},$$

the unit ball in $\mathcal{P}$. Furthermore, in (11), we use $P_{XC^\perp}$ to denote the projection operator onto the linear subspace

$$XC^\perp = X\{v : v \perp C\} \subseteq \mathbb{R}^n.$$

We recall that for the lasso problem (3), the penalty function is $\mathcal{P}(\beta) = \|\beta\|_1$, so $\mathcal{Q}(\beta) = \|\beta\|_\infty$; also $C = \{u : \|u\|_\infty \leq 1\}$, which means that $C^\perp = 0$, and $P_{XC^\perp} = 0$. Hence $\lambda_1 = \|X^T y\|_\infty$, as claimed in Section 1.1.

Having just given an example of a seminorm in which C is of full dimension p, so that $C^\perp = \{0\}$, we now consider one in which C has dimension less than p, so that $C^\perp$ is nontrivial. In a generalized lasso problem [22], the penalty is $\mathcal{P}(\beta) = \|D\beta\|_1$ for some chosen penalty matrix $D \in \mathbb{R}^{m \times n}$. In this case, it can be shown that the dual seminorm is $\mathcal{Q}(\beta) = \min_{D^T z = \beta} \|z\|_\infty$. Hence, $C = \{u : \min_{D^T z = u} \|z\|_\infty \leq 1\}$, and $C^\perp = \text{null}(D)$, the null space of D. In many interesting cases, this null space is nontrivial; for example, if D is the fused lasso penalty matrix, then its null space is spanned by the vector of all 1's.

2.1. *A null hypothesis.* As in the lasso case, we assume that y is generated from the normal model

$$y \sim N(X\beta_0, \Sigma), \tag{12}$$

with X considered fixed. We are interested in the distribution of (1) in order to test the following hypothesis:

$$H_0 : \mathcal{P}(\beta_0) = 0. \tag{13}$$

This can be seen a global null hypothesis, a test of whether the true underlying coefficient vector β_0 has a trivial structure, according to the designated penalty function $\mathcal{P}$.

Assuming that the set C contains 0 in its relative interior, we have $\mathcal{P}(\beta) = 0 \iff P_C \beta = 0$, where P_C denotes the projection matrix onto $\text{span}(C)$. Therefore we can rewrite the null hypothesis (13) in a more transparent form, as

$$H_0 : P_C \beta_0 = 0. \tag{14}$$

Again, using the lasso problem (3) as a reference point, we have $\text{span}(C) = \mathbb{R}^p$ for this problem, so the above null hypothesis reduces to $H_0 : \beta_0 = 0$, as in Section 1.1. In general, the null hypothesis (14) tests $\beta_0 \in C^\perp$, the orthocomplement of C.

Recalling that

$$\lambda_1 = \max_{v \in C^\circ} v^T X^T (I - P_{XC^\perp}) y,$$

one can check that, under H_0, the quantity λ_1 is precisely of the form (1), with $\mathcal{K} = C^\circ$, $\varepsilon = X^T(I - P_{XC^\perp})y$, and $\Theta = \mathrm{Cov}(\varepsilon) = X^T(I - P_{XC^\perp})\Sigma(I - P_{XC^\perp})X$ [as $\mathbb{E}(\varepsilon) = X^T(I - P_{XC^\perp})X\beta_0 = 0$ when $\beta_0 \in C^\perp$].

2.2. *Statement of main result and outline of our approach.* We now state our main result.

THEOREM 1 (Kac–Rice test). *Consider the general regularized regression problem in* (9), *with* $\mathcal{P}(\beta) = \max_{u \in C} u^T\beta$ *for a closed, convex set* $C \subseteq \mathbb{R}^p$ *containing* 0 *in its relative interior. Denote* $\mathcal{K} = C^\circ = \{v : \mathcal{P}(v) \le 1\}$, *the polar set of* C, *and assume that* $\mathcal{K}$ *can be stratified into pieces of different dimensions, that is,*

$$\mathcal{K} = \bigcup_{j=0}^{p} \partial_j \mathcal{K}, \tag{15}$$

where $\partial_0\mathcal{K}, \ldots, \partial_p\mathcal{K}$ *are smooth disjoint manifolds of dimensions* $0, \ldots, p$, *respectively. Assume also assume that the process*

$$f_\eta = \eta^T X^T (I - P_{XC^\perp})y, \qquad \eta \in \mathcal{K}, \tag{16}$$

is Morse for almost every $y \in \mathbb{R}^n$. *Finally, assume that* $y \in \mathbb{R}^n$ *is drawn from the normal distribution in* (12).

Now, consider testing the null hypothesis $H_0 : \mathcal{P}(\beta_0) = 0$ [*equivalently,* $H_0 : P_C\beta_0 = 0$, *since we have assumed that* $0 \in \mathrm{relint}(C)$]. *Define* $\Lambda_\eta = G_\eta^{-1}H_\eta$ *for* G_η, H_η *as in* (29), (30), $\mathcal{V}_\eta^-$, $\mathcal{V}_\eta^+$ *as in* (24), (23), *and* σ_η^2 *as in* (32). *Finally, let* η^* *denote the almost sure unique maximizer of the process* f_η *over* $\mathcal{K}$,

$$\eta^* = \operatorname*{argmax}_{\eta \in \mathcal{K}} f_\eta,$$

and let $\lambda_1 = f_{\eta^*}$ *denote the first knot in the solution path of problem* (9). *Then under* H_0,

$$\mathbb{S} = \frac{\int_{\lambda_1}^{\mathcal{V}_{\eta^*}^+} \det(\Lambda_{\eta^*} + zI)\phi_{\sigma_{\eta^*}^2}(z)\,dz}{\int_{\mathcal{V}_{\eta^*}^-}^{\mathcal{V}_{\eta^*}^+} \det(\Lambda_{\eta^*} + zI)\phi_{\sigma_{\eta^*}^2}(z)\,dz} \sim \mathrm{Unif}(0, 1), \tag{17}$$

where ϕ_{σ^2} *denotes the density function of a normal random variable with mean* 0 *and variance* σ^2.

The quantity (17) is the *Kac–Rice pivot* evaluated at $\mu = 0$. Lemma 4 shows it is a pivotal quantity for the mean μ near η^*, derived via the Kac–Rice formula. Here we give a rough explanation of the result in (17), and the approach we take to prove it in Sections 3 and 4. The next section, Section 2.3, discusses the assumptions behind Theorem 1; in summary, the assumption that $\mathcal{K}$ separates as in (15) allows

us to apply the Kac–Rice formula to each of its strata, and the Morse assumption on the process f_η in (16) ensures the uniqueness of its maximizer η^*. These are very weak assumptions, especially considering the strength of the exact, nonasymptotic conclusion in (17).

Our general approach is based on finding an implicit formula for $\mathbb{P}(\lambda_1 > t)$ under the null hypothesis H_0, where λ_1 is the first knot in the solution path of problem (9) and can be written as

$$\lambda_1 = \max_{\eta \in \mathcal{K}} f_\eta,$$

where $f_\eta = \eta^T X^T (I - P_{XC^\perp}) y$, the process in (16). Our representation for the tail probability of λ_1 has the form

$$\mathbb{P}(\lambda_1 > t) = \mathbb{E}\big(\mathbb{Q}(1_{(t,\infty)})\big). \tag{18}$$

Here, $\mathbb{Q}$ is a random distribution function and $1_{(t,\infty)}$ is the indicator function for the interval (t, ∞). The distribution $\mathbb{Q}$ depends on η^*, a maximizer of the process f_η, almost surely unique by the Morse assumption. This maximizer satisfies

$$\eta^* \in \partial \mathcal{Q}\big(X^T (I - P_{XC^\perp}) y\big) \subseteq \mathcal{K},$$

with $\partial \mathcal{Q}$ the subdifferential of the seminorm $\mathcal{Q}$. Under the assumption that $\mathcal{K} = \bigcup_{j=0}^{p} \partial_j \mathcal{K}$, the main tool we invoke is the Kac–Rice formula [1], which essentially enables us to compute the expected number[5] of global maximizers occuring in each stratum $\partial_j \mathcal{K}$. This leads to the distribution of λ_1, in Theorem 2, as well as the representation in (18), with $\mathbb{Q}$ given in an explicit form. Unfortunately, computing tail probabilities $\mathbb{P}(\lambda_1 > t)$ of this distribution involve evaluating some complicated integrals over $\mathcal{K}$ that depend on X, Σ, and hence the quantity λ_1 as a test statistic does not easily lend itself to the computation of p-values. We therefore turn to the survival function $\mathbb{S}$ associated with the measure $\mathbb{Q}$, and our main result is that, when carefully evaluated, this (random) survival function can be used to derive a test of H_0, as expressed in (17) in Theorem 1 above.

2.3. *Discussion of assumptions.* In terms of the assumptions of Theorem 1, we require that C contains 0 in its relative interior so that we can write the null hypothesis in the equivalent form $H_0 : P_C \beta_0 = 0$, which makes the process f_η in (16) have mean zero under H_0. We additionally assume that C is closed in order

[5]Note that for almost every realization, under the Morse assumption, there is generically only one maximizer overall and hence the number of them is either 0 or 1. We use the term "number of global maximizers" when applying the Kac–Rice formula as it applies to counting different types of points. In our applications of it, however, there is only ever 0 or 1 such points. Similar arguments were used to establish the accuracy of the expected Euler characteristic approximation for the distribution of the global maximum of a smooth Gaussian process in Taylor et al. [18].

to guarantee that f_η has a well-defined (finite) maximum over $\eta \in \mathcal{K} = C^\circ$. See Section A.3 of the supplement [17].

Apart from these rather minor assumptions on C, the main requirements of the theorem are: the polar set $C^\circ = \mathcal{K}$ can be stratified as in (15), the process f_η in (16) is Morse, and y follows the normal distribution in (12). Overall, these are quite weak assumptions. The first assumption, on $\mathcal{K}$ separating as in (15), eventually permits us to apply to the Kac–Rice formula to each stratum $\partial_j \mathcal{K}$. We remark that many convex sets possess such a decomposition; see Adler and Taylor [1]. In particular, note that such an assumption does not limit our consideration to polyhedral $\mathcal{K}$: a set can be stratifiable but still have a boundary with curvature (e.g., as in $\mathcal{K}$ for the group lasso and nuclear norm penalties).

Further, the property of being a Morse function is truly generic; again, see Adler and Taylor [1] for a discussion of Morse functions on stratified spaces. If f_η is Morse for almost every y, then its maximizers are almost surely isolated, and the convexity of $\mathcal{K}$ then implies that f_η has an almost surely unique maximizer η^*. From the form of our particular process f_η in (16), the assumption that f_η is Morse can be seen as a restriction on the predictor matrix X (or more generally, how X interacts with the set C). For most problems, this only rules out trivial choices of X. In the lasso case, for example, recall that $f_\eta = (X\eta)^T y$ and $\mathcal{K}$ is equal to the unit ℓ_1 ball, so $f_\eta^* = \|X^T y\|_\infty$, and the Morse property requires $|X_j^T y|$, $j = 1, \ldots, p$ to be unique for almost every $y \in \mathbb{R}^p$. This can be ensured by taking X with columns in general position (a weak condition that also ensures uniqueness of the lasso solution; see Tibshirani [21]).

The assumption of normally distributed errors in the regression model (12) is important for the work that follows in Sections 3 and 4, which is based on Gaussian process theory. Note that we assume a known covariance matrix Σ, but we allow for a dependence between the errors (i.e., Σ need not be diagonal). Empirically, the (uniform) distribution of our test statistic under the null hypothesis appears to quite robust against nonnormal errors in many cases of interest; we discuss this in Section 6.1.

Lastly, we make a note about the role of convexity. Aside from computing $\mathcal{V}_\eta^-, \mathcal{V}_\eta^+$ (see Section A.6 of the supplementary document [17]), the convexity of the parameter space $\mathcal{K}$ is not necessary; we only need local convexity as described in Adler and Taylor [1], that is, we only need to assume that the support cone of $\mathcal{K}$ is locally convex everywhere. This is essentially the same as positive reach [8]. To be clear, while convexity is used in connecting the Kac–Rice test to the regularized regression problem in (11) (i.e., it establishes an equality between left- and right-hand sides), the right-hand side is well-defined even if C° is not convex. That is, the set $\mathcal{K}$ in (1) need not be convex. In fact, the issue of convexity is only important for computational purposes, not theoretical purposes. In this sense, the current paper provides an exact conditional test based on the global maximizer of a smooth Gaussian field on a fairly arbitrary set. This is an advance in the theory of smooth Gaussian fields as developed in [1] and will be investigated in future work.

2.4. *Notation.* Rewrite the process f_η in (16) as

$$f_\eta = \eta^T X^T (I - P_{XC^\perp})(I - P_{XC^\perp}) y = \eta^T \widetilde{X}^T \widetilde{y}, \qquad \eta \in \mathcal{K},$$

where $\widetilde{X} = (I - P_{XC^\perp})X$ and $\widetilde{y} = (I - P_{XC^\perp})y$. The distribution of $\widetilde{y}$ is hence $\widetilde{y} \sim N(\widetilde{X}\beta_0, \widetilde{\Sigma})$, where $\widetilde{\Sigma} = (I - P_{XC^\perp})\Sigma(I - P_{XC^\perp})$. Furthermore, under the null hypothesis $H_0 : P_C \beta_0 = 0$, we have $\widetilde{y} \sim N(0, \widetilde{\Sigma})$. For convenience, in Sections 3 and 4, we will drop the tilde notation, and write $\widetilde{y}, \widetilde{X}, \widetilde{\Sigma}$ as simply y, X, Σ, respectively. To be perfectly explicit, this means that we will write the process f_η in (16) as

$$f_\eta = \eta^T X^T y, \qquad \eta \in \mathcal{K},$$

where $y \sim N(X\beta_0, \Sigma)$, and the null hypothesis is $H_0 : y \sim N(0, \Sigma)$. Notice that when $\operatorname{span}(C) = \mathbb{R}^p$, we have exactly $\widetilde{y} = y$, $\widetilde{X} = X$, $\widetilde{\Sigma} = \Sigma$, since $P_{XC^\perp} = 0$. However, we reiterate that replacing $\widetilde{y}, \widetilde{X}, \widetilde{\Sigma}$ by y, X, Σ in Sections 3 and 4 is done purely for notational convenience, and the reader should bear in mind that the arguments themselves do not portray any loss of generality.

We will write $\mathbb{E}_0$ to emphasize that an expectation is taken under the null distribution $H_0 : y \sim N(0, \Sigma)$.

3. Characterization of the global maximizer. Near any point $\eta \in \mathcal{K}$, the set $\mathcal{K}$ is well approximated by the support cone $S_\eta \mathcal{K}$, which is defined as the polar cone of the normal cone $N_\eta \mathcal{K}$. The support cone $S_\eta \mathcal{K}$ contains a largest linear subspace—we will refer to this $T_\eta \mathcal{K}$, the *tangent space* to $\mathcal{K}$ at η. The tangent space plays an important role in what follows.

We study the process f_η in (16), which we now write as $f_\eta = \eta^T X^T y$ over $\eta \in \mathcal{K}$, where $y \sim N(X\beta_0, \Sigma)$, and with the null hypothesis $H_0 : y \sim N(0, \Sigma)$ (see our notational reduction in Section 2.4). We proceed as in Chapter 14 of [1], with an important difference being that here the process f_η does not have constant variance. Aside from the statistical implications of this work that have to do with hypothesis testing, another goal of this paper is to derive analogues of the results in Adler and Taylor [1], Taylor et al. [18], for Gaussian processes with nonconstant variance. For each $\eta \in \mathcal{K}$, we define a modified process

$$\widetilde{f}_z^\eta = f_z - z^T \alpha_{\eta, X, \Sigma}(\nabla f_{|T_\eta \mathcal{K}}), \qquad z \in \mathcal{K},$$

where $\alpha_{\eta, X, \Sigma}(\nabla f_{|T_\eta \mathcal{K}})$ is the vector that, under $H_0 : y \sim N(0, \Sigma)$, computes the expectation of f_z given $\nabla f_{|T_\eta \mathcal{K}}$, the gradient restricted to $T_\eta \mathcal{K}$, that is,

$$z^T \alpha_{\eta, X, \Sigma}(\nabla f_{|T_\eta \mathcal{K}}) = \mathbb{E}_0(f_z | \nabla f_{|T_\eta \mathcal{K}}).$$

To check that such a representation is possible, suppose that the tangent space $T_\eta \mathcal{K}$ is j-dimensional, and let $V_\eta \in \mathbb{R}^{p \times j}$ be a matrix whose columns form an orthonormal basis for $T_\eta \mathcal{K}$. Then $\nabla f_{|T_\eta \mathcal{K}} = V_\eta V_\eta^T X^T y$, and a simple calculation using the

properties of conditional expectations for jointly Gaussian random variables shows that

$$\mathbb{E}_0(f_z|\nabla f_{|T_\eta\mathcal{K}}) = z^T X^T P_{\eta,X,\Sigma} y,$$

where

$$P_{\eta,X,\Sigma} = \Sigma X V_\eta \big(V_\eta^T X^T \Sigma X V_\eta\big)^\dagger V_\eta^T X^T, \tag{19}$$

the projection onto XV_η with respect to Σ (and $A^\dagger$ denoting the Moore–Penrose pseudoinverse of a matrix A). Hence, we gather that

$$\alpha_{\eta,X,\Sigma}(\nabla f_{|T_\eta\mathcal{K}}) = X^T P_{\eta,X,\Sigma} y,$$

and our modified process has the form

$$\widetilde{f}_z^\eta = f_z - z^T X^T P_{\eta,X,\Sigma} y = (Xz)^T (I - P_{\eta,X,\Sigma}) y. \tag{20}$$

The key observation, as in Taylor et al. [18] and Adler and Taylor [1], is that if η is a critical point, that is, $\nabla f_{|T_\eta\mathcal{K}} = 0$, then

$$\widetilde{f}_z^\eta = f_z \qquad \text{for all } z \in \mathcal{K}. \tag{21}$$

Similar to our construction of $\alpha_{\eta,X,\Sigma}(\nabla f_{|T_\eta\mathcal{K}})$, we define $C_{X,\Sigma}(\eta)$ such that

$$\begin{aligned}\mathbb{E}_0\big(\widetilde{f}_z^\eta|\widetilde{f}_\eta^\eta\big) &= \frac{z^T X^T (I - P_{\eta,X,\Sigma})\Sigma(I - P_{\eta,X,\Sigma}^T) X\eta}{\eta^T X^T (I - P_{\eta,X,\Sigma})\Sigma(I - P_{\eta,X,\Sigma}^T) X\eta} \cdot \widetilde{f}_\eta^\eta \\ &= z^T C_{X,\Sigma}(\eta) \cdot \widetilde{f}_\eta^\eta\end{aligned} \tag{22}$$

and after making three subsequent definitions,

$$\mathcal{V}_\eta^- = \max_{z\in\mathcal{K}: z^T C_{X,\Sigma}(\eta) < 1} \frac{\widetilde{f}_z^\eta - z^T C_{X,\Sigma}(\eta)\cdot \widetilde{f}_\eta^\eta}{1 - z^T C_{X,\Sigma}(\eta)}, \tag{23}$$

$$\mathcal{V}_\eta^+ = \min_{z\in\mathcal{K}: z^T C_{X,\Sigma}(\eta) > 1} \frac{\widetilde{f}_z^\eta - z^T C_{X,\Sigma}(\eta)\cdot \widetilde{f}_\eta^\eta}{1 - z^T C_{X,\Sigma}(\eta)}, \tag{24}$$

$$\mathcal{V}_\eta^0 = \max_{z\in\mathcal{K}: z^T C_{X,\Sigma}(\eta) = 1} \widetilde{f}_z^\eta - z^T C_{X,\Sigma}(\eta)\cdot \widetilde{f}_\eta^\eta, \tag{25}$$

we are ready to state our characterization of the global maximizer η.

LEMMA 1. *A point $\eta \in \mathcal{K}$ maximizes f_η over a convex set $\mathcal{K}$ if and only if the following conditions hold*:

$$\nabla f_{|T_\eta\mathcal{K}} = 0, \qquad \widetilde{f}_\eta^\eta \ge \mathcal{V}_\eta^-, \qquad \widetilde{f}_\eta^\eta \le \mathcal{V}_\eta^+ \quad \textit{and} \quad \mathcal{V}_\eta^0 \le 0. \tag{26}$$

The same equivalence holds true even when $\mathcal{K}$ is only locally convex.

PROOF. In the forward direction ($\Rightarrow$), note that $\nabla f_{|T_\eta \mathcal{K}} = 0$ implies that we can replace $\widetilde{f}_z^\eta$ by f_z (and $\widetilde{f}_\eta^\eta$ by f_η) in the definitions (24), (23), (25), by the key observation (21). As each $z \in \mathcal{K}$ is covered by one of the cases $C_{X,\Sigma}(\eta) < 1$, $C_{X,\Sigma}(\eta) > 1$, $C_{X,\Sigma}(\eta) = 1$, we conclude that

$$f_\eta \geq f_z \qquad \text{for all } z \in \mathcal{K},$$

that is, the point η is a global maximizer.

As for the reverse direction ($\Leftarrow$), when η is the global maximizer of f_η over $\mathcal{K}$, the first condition $\nabla f_{|T_\eta \mathcal{K}} = 0$ is clearly true (provided that $\mathcal{K}$ is convex or locally convex), and the other three conditions follow from simple manipulations of the inequalities

$$f_\eta \geq f_z \qquad \text{for all } z \in \mathcal{K}. \qquad \square$$

REMARK 1. The above lemma does not assume that $\mathcal{K}$ decomposes into strata, or that f_η is Morse for almost all w, or that $y \sim N(X\beta_0, \Sigma)$. It only assumes that $\mathcal{K}$ is convex or locally convex, and its conclusion is completely deterministic, depending only on the process f_η via its covariance function under the null, that is, via the terms $P_{\eta,X,\Sigma}$ and $C_{X,\Sigma}(\eta)$.

We note that, under the assumption that f_η is Morse over $\mathcal{K}$ for almost every $y \in \mathbb{R}^p$, and $\mathcal{K}$ is convex, Lemma 1 gives necessary and sufficient conditions for a point $\eta \in \mathcal{K}$ to be the almost sure unique global maximizer. Hence, for convex $\mathcal{K}$, the conditions in (26) are equivalent to the usual subgradient conditions for optimality, which may be written as

$$\nabla f_\eta \in N_\eta \mathcal{K} \quad \Longleftrightarrow \quad \nabla f_{|T_\eta \mathcal{K}} = 0, \nabla f_{|(T_\eta \mathcal{K})^\perp} \in N_\eta \mathcal{K},$$

where $N_\eta \mathcal{K}$ is the normal cone to $\mathcal{K}$ at η and $\nabla f_{|(T_\eta \mathcal{K})^\perp}$ is the gradient restricted to the orthogonal complement of the tangent space $T_\eta \mathcal{K}$.

Recalling that f_η is a Gaussian process, a helpful independence relationship unfolds.

LEMMA 2. *With $y \sim N(X\beta_0, \Sigma)$, for each fixed $\eta \in \mathcal{K}$, the triplet $(\mathcal{V}_\eta^-, \mathcal{V}_\eta^+, \mathcal{V}_\eta^0)$ is independent of $\widetilde{f}_\eta^\eta$.*

PROOF. This is a basic property of conditional expectation for jointly Gaussian variables, that is, it is easily verified that $\mathrm{Cov}(\widetilde{f}_z^\eta - z^T C_{X,\Sigma}(\eta), \widetilde{f}_\eta^\eta) = 0$ for all z. $\square$

4. Kac–Rice formulae for the global maximizer and its value. The characterization of the global maximizer from the last section, along with the Kac–Rice formula [1], allow us to express the joint distribution of

$$\eta^* = \operatorname*{argmax}_{\eta \in \mathcal{K}} f_\eta \quad \text{and} \quad f_{\eta^*} = \max_{\eta \in \mathcal{K}} f_\eta.$$

THEOREM 2 [Joint distribution of (η^*, f_{η^*})]. *Writing $\mathcal{K} = \bigcup_{j=0}^p \partial_j \mathcal{K}$ for a stratification of $\mathcal{K}$, for open sets $A \subseteq \mathbb{R}^p$, $O \subseteq \mathbb{R}$,*

$$\begin{aligned}
&\mathbb{P}\big(\eta^* \in A,\, f_{\eta^*} \in O\big) \\
&= \sum_{j=0}^p \int_{\partial_j \mathcal{K} \cap A} \mathbb{E}\big(\det\big(-\nabla^2 f_{|T_\eta \mathcal{K}}\big) \cdot 1_{\{\mathcal{V}_\eta^- \le \widetilde{f}_\eta^\eta \le \mathcal{V}_\eta^+, \mathcal{V}_\eta^0 \le 0,\, \widetilde{f}_\eta^\eta \in O\}} \big| \nabla f_{|T_\eta \mathcal{K}} = 0\big) \\
&\qquad \times \psi_{\nabla f_{|T_\eta \mathcal{K}}}(0) \mathcal{H}_j(d\eta),
\end{aligned} \tag{27}$$

where:

- *$\psi_{\nabla f_{|T_\eta \mathcal{K}}}$ is the density of the gradient in some basis for the tangent space $T_\eta \mathcal{K}$, orthonormal with respect to the standard inner product on $T_\eta \mathcal{K}$, that is, the standard Euclidean Riemannian metric on $\mathbb{R}^p$;*
- *the measure $\mathcal{H}_j$ is the Hausdorff measure induced by the above Riemannian metric on each $\partial_j \mathcal{K}$;*
- *the Hessian $\nabla^2 f_{|T_\eta \mathcal{K}}$ is evaluated in this orthonormal basis and, for $j = 0$, we take as convention the determinant of a 0×0 matrix to be 1 (in [1], this was denoted by $\nabla^2 f_{|\partial_j \mathcal{K}, \eta}$, to emphasize that it is the Hessian of the restriction of f to $\partial_j \mathcal{K}$).*

PROOF. This is the Kac–Rice formula, or the "meta-theorem" of Chapter 10 of [1] (see also [3, 4]), applied to the problem of counting the number of global maximizers in some set $A \subseteq \mathbb{R}^p$ having value in $O \subseteq \mathbb{R}$. That is,

$$\begin{aligned}
&\mathbb{P}\big(\eta^* \in A,\, f_{\eta^*} \in O\big) \\
&\quad = \mathbb{E}\big(\#\big\{\eta \in \mathcal{K} \cap A : \nabla f_{|T_\eta \mathcal{K}} = 0,\, \mathcal{V}_\eta^- \le \widetilde{f}_\eta^\eta \le \mathcal{V}_\eta^+,\, \mathcal{V}_\eta^0 \le 0,\, f_\eta \in O\big\}\big) \\
&\quad = \mathbb{E}\big(\#\big\{\eta \in \mathcal{K} \cap A : \nabla f_{|T_\eta \mathcal{K}} = 0,\, \mathcal{V}_\eta^- \le \widetilde{f}_\eta^\eta \le \mathcal{V}_\eta^+,\, \mathcal{V}_\eta^0 \le 0,\, \widetilde{f}_\eta^\eta \in O\big\}\big),
\end{aligned}$$

where the second equality follows from (21). Breaking down $\mathcal{K}$ into its separate strata, and then using the Kac–Rice formula, we obtain the result in (27). □

REMARK 2. As before, the conclusion of Theorem 2 does not actually depend on the convexity of $\mathcal{K}$. When $\mathcal{K}$ is only locally convex, the Kac–Rice formula [i.e., the right-hand side in (27)] counts the expected total number of global maximizers of f_η lying in some set $A \subseteq \mathbb{R}^p$, with the achieved maximum value in $O \subseteq \mathbb{R}$. For convex $\mathcal{K}$, our Morse condition on f_η implies an almost surely unique maximizer,

and hence the notation $\mathbb{P}(\eta^* \in A,\ f_{\eta^*} \in O)$ on the left-hand side of (27) makes sense as written. For locally convex $\mathcal{K}$, one simply needs to interpret the left-hand side as

$$\mathbb{P}\Big(\eta^* \in A \text{ for some } f_{\eta^*} = \max_{\eta \in \mathcal{K}} f_\eta, \max_{\eta \in \mathcal{K}} f_\eta \in O\Big).$$

REMARK 3. When f_η has constant variance, the distribution of the maximum value f_{η^*} can be approximated extremely well by the expected Euler characteristic [1] of the excursion set $f_\eta^{-1}(t, \infty) \cap \mathcal{K}$,

$$\sum_{j=0}^{p} \int_{\partial_j \mathcal{K}} \mathbb{E}\big(\det(-\nabla^2 f_{|T_\eta \mathcal{K}}) 1_{\{f_\eta > t, \nabla f_{|(T_\eta \mathcal{K})^\perp} \in N_\eta \mathcal{K}\}} | \nabla f_{|T_\eta \mathcal{K}} = 0\big) \psi_{\nabla f_{|T_\eta \mathcal{K}}}(0) \mathcal{H}_j(d\eta).$$

This approximation is exact when $\mathcal{K}$ is convex [14], since the Euler characteristic of the excursion set is equal to the indicator that it is not empty.

4.1. *Decomposition of the Hessian.* In looking at the formula (27), we note that the quantities $\widetilde{f}_\eta^\eta, \mathcal{V}_\eta^-, \mathcal{V}_\eta^+, \mathcal{V}_\eta^0$ inside the indicator are all independent, by construction, of $\nabla f_{|T_\eta K}$. It will be useful to decompose the Hessian term similarly. We write

$$-\nabla^2 f_{|T_\eta \mathcal{K}} = -H_\eta + G_\eta \cdot \widetilde{f}_\eta^\eta + R_\eta,$$

where

$$R_\eta = -\mathbb{E}_0\big(\nabla^2 f_{|T_\eta \mathcal{K}} | \nabla f_{|T_\eta \mathcal{K}}\big), \tag{28}$$

$$G_\eta \cdot \widetilde{f}_\eta^\eta = -\mathbb{E}_0\big(\nabla^2 f_{|T_\eta \mathcal{K}} | \widetilde{f}_\eta^\eta\big), \tag{29}$$

$$H_\eta = -\big(\nabla^2 f_{|T_\eta \mathcal{K}} - R_\eta\big) - G_\eta \cdot \widetilde{f}_\eta^\eta. \tag{30}$$

At a critical point of $f_{|\partial_j \mathcal{K}}$, notice that $R_\eta = 0$ (being a linear function of the gradient $\nabla f_{|T_\eta \mathcal{K}}$, which is zero at such a critical point). Furthermore, the pair of matrices $(G_\eta \widetilde{f}_\eta^\eta, H_\eta)$ is independent of $\nabla f_{|T_\eta \mathcal{K}}$. Hence, we can rewrite our key formulae for the distribution of the maximizer and its value.

LEMMA 3. *For each fixed $\eta \in \mathcal{K}$, we have*

$$\widetilde{f}_\eta^\eta \sim N(\mu_\eta, \sigma_\eta^2),$$

independently of $(\mathcal{V}_\eta^-, \mathcal{V}_\eta^+, \mathcal{V}_\eta^0, H_\eta)$, with

$$\mu_\eta = \eta^T X (I - P_{\eta, X, \Sigma}) X \beta_0, \tag{31}$$

$$\sigma_\eta^2 = \eta^T X^T (I - P_{\eta, X, \Sigma}) \Sigma (I - P_{\eta, X, \Sigma}^T) X \eta, \tag{32}$$

and (recall) $P_{\eta, X, \Sigma} = \Sigma X V_\eta (V_\eta^T X^T \Sigma X V_\eta)^\dagger V_\eta^T X^T$, for an orthonormal basis V_η of $T_\eta \mathcal{K}$.

Moreover, the formula (27) *can be equivalently expressed as*

$$\begin{aligned}\mathbb{P}(\eta^* \in A,\, f_{\eta^*} \in O) &= \sum_{j=0}^{p} \int_{\partial_j \mathcal{K} \cap A} \mathbb{E}\big(\det(-H_\eta + G_\eta \widetilde{f}_\eta^{\eta}) \\ &\qquad \times 1_{\{\mathcal{V}_\eta^- \le \widetilde{f}_\eta^{\eta} \le \mathcal{V}_\eta^+,\, \mathcal{V}_\eta^0 \le 0,\, \widetilde{f}_\eta^{\eta} \in O\}}\big) \psi_{\nabla f_{|T_\eta \mathcal{K}}}(0) \mathcal{H}_j(d\eta) \qquad (33)\\ &= \sum_{j=0}^{p} \int_{\partial_j \mathcal{K} \cap A} \mathbb{E}\big(\mathbb{M}_{\Lambda_\eta, \mathcal{V}_\eta, \mu_\eta, \sigma_\eta^2}(1_O) 1_{\{\mathcal{V}_\eta^- \le \mathcal{V}_\eta^+,\, \mathcal{V}_\eta^0 \le 0\}}\big) \\ &\qquad \times \psi_{\nabla f_{|T_\eta \mathcal{K}}}(0) \det(G_\eta) \mathcal{H}_j(d\eta), \qquad (34)\end{aligned}$$

where

$$\mathbb{M}_{\Lambda, \mathcal{V}, \mu, \sigma^2}(h) = \int_{\mathcal{V}^-}^{\mathcal{V}^+} h(z) \det(\Lambda + zI) \frac{e^{-(z-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\, dz, \qquad (35)$$

and $\Lambda_\eta = G_\eta^{-1} H_\eta$.

REMARK 4. Until (34), we had not used the independence of $\widetilde{f}_\eta^{\eta}$ and $\mathcal{V}_\eta^-, \mathcal{V}_\eta^+, \mathcal{V}_\eta^0$ (Lemma 2). In (34), we do so, by first integrating over $\widetilde{f}_\eta^{\eta}$ (in the definition of $\mathbb{M}$), and then over $\mathcal{V}_\eta^-, \mathcal{V}_\eta^+, \mathcal{V}_\eta^0$.

4.2. *The conditional distribution.* The Kac–Rice formula can be generalized further. For a possibly random function h, with $h_{|\partial_j \mathcal{K}}$ continuous for each $j = 0, \ldots, p$, we see as a natural extension from (33),

$$\begin{aligned}\mathbb{E}(h(\eta^*)) = \sum_{j=0}^{p} \int_{\partial_j \mathcal{K}} \mathbb{E}\big(h(\eta) \det(-H_\eta + G_\eta \widetilde{f}_\eta^{\eta}) \\ \times 1_{\{\mathcal{V}_\eta^- \le \widetilde{f}_\eta^{\eta} \le \mathcal{V}_\eta^+,\, \mathcal{V}_\eta^0 \le 0\}}\big) \psi_{\nabla f_{|T_\eta \mathcal{K}}}(0) \mathcal{H}_j(d\eta).\end{aligned} \qquad (36)$$

This allows us to form a conditional distribution function of sorts. As defined in (35), $\mathbb{M}_{\Lambda, \mathcal{V}, \mu, \sigma^2}$ is not a probability measure, but it can be normalized to yield one:

$$\mathbb{Q}_{\Lambda, \mathcal{V}, \mu, \sigma^2}(g) = \frac{\mathbb{M}_{\Lambda, \mathcal{V}, \mu, \sigma^2}(g)}{\mathbb{M}_{\Lambda, \mathcal{V}, \mu, \sigma^2}(1)}. \qquad (37)$$

Working form (36), with $h(\eta) = g(f_\eta)$,

$$\begin{aligned}\mathbb{E}(g(f_{\eta^*})) &= \sum_{j=0}^{p} \int_{\partial_j \mathcal{K}} \mathbb{E}\big(\mathbb{M}_{\Lambda_\eta, \mathcal{V}_\eta, \mu_\eta, \sigma_\eta^2}(g) 1_{\{\mathcal{V}_\eta^- \le \mathcal{V}_\eta^+,\, \mathcal{V}_\eta^0 \le 0\}}\big) \\ &\qquad \times \psi_{\nabla f_{|T_\eta \mathcal{K}}}(0) \det(G_\eta) \mathcal{H}_j(d\eta) \qquad (38)\\ &= \sum_{j=0}^{p} \int_{\partial_j \mathcal{K}} \mathbb{E}\big(\mathbb{Q}_{\Lambda_\eta, \mathcal{V}_\eta, \mu_\eta, \sigma_\eta^2}(g) \mathbb{M}_{\Lambda_\eta, \mathcal{V}_\eta, \mu_\eta, \sigma_\eta^2}(1) 1_{\{\mathcal{V}_\eta^- \le \mathcal{V}_\eta^+,\, \mathcal{V}_\eta^0 \le 0\}}\big) \\ &\qquad \times \psi_{\nabla f_{|T_\eta \mathcal{K}}}(0) \det(G_\eta) \mathcal{H}_j(d\eta).\end{aligned}$$

In fact, the quantity $\mathbb{Q}_{\Lambda,\mathcal{V},\mu_\eta,\sigma^2_\eta}$ is a proper conditional distribution. That is, for suitable g,

$$\mathbb{E}\big(g(f_{\eta^*})|\eta^* = \eta, \Lambda_{\eta^*} = \Lambda, \mathcal{V}_{\eta^*} = \mathcal{V}\big) = \mathbb{Q}_{\Lambda,\mathcal{V},\mu_\eta,\sigma^2_\eta}(g). \tag{39}$$

See Section A.4 of the supplement [17] for details.

4.3. *The Kac–Rice pivotal quantity.* Suppose that we are interested in testing the null hypothesis $H_0 : y \sim N(0, \Sigma)$. We might look at the observed value of the first knot $\lambda_1 = f_{\eta^*}$, and see if it was larger than we would expect under H_0. From the results of the last section,

$$\mathbb{P}(f_{\eta^*} > t) = \mathbb{E}\big(\mathbb{Q}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},\mu_{\eta^*},\sigma^2_{\eta^*}}(1_{(t,\infty)})\big),$$

and so the most natural strategy seems to be to plug our observed value of the first knot into the above formula. This, however, requires computing the above expectation, that is, the integral in (38).

In this section, we present an alternative approach that is effectively a conditional test, conditioning on the observed value of η^*, as well as Λ_{η^*} and $\mathcal{V}_{\eta^*}$. To motivate our test, it helps to take a step back and think about the measure $\mathbb{Q}_{\Lambda,\mathcal{V},\mu,\sigma^2}$ defined in (37). For fixed values of $\Lambda, \mathcal{V}, \mu, \sigma^2$, we can reexpress this (nonrandom) measure as

$$\mathbb{Q}_{\Lambda,\mathcal{V},\mu,\sigma^2}(g) = \int_{-\infty}^{\infty} g(t) \cdot q_{\Lambda,\mathcal{V},\mu,\sigma}(t)\, dt,$$

where $q_{\Lambda,\mathcal{V},\mu,\sigma}$ is a density function (supported on $[\mathcal{V}^-, \mathcal{V}^+]$). In other words, $\mathbb{Q}_{\Lambda,\mathcal{V},\mu,\sigma^2}(g)$ computes the expectation of g with respect to a density $q_{\Lambda,\mathcal{V},\mu,\sigma^2}$, so we can write $\mathbb{Q}_{\Lambda,\mathcal{V},\mu,\sigma^2}(g) = \mathbb{E}(g(W))$ where W is a random variable whose density is $q_{\Lambda,\mathcal{V},\mu,\sigma^2}$. Now consider the survival function

$$\mathbb{S}_{\Lambda,\mathcal{V},\mu,\sigma^2}(t) = \mathbb{Q}_{\Lambda,\mathcal{V},\mu,\sigma^2}(1_{(t,\infty)}) = \mathbb{P}(W > t).$$

A classic argument shows that $\mathbb{S}_{\Lambda,\mathcal{V},\mu,\sigma^2}(W) \sim \mathrm{Unif}(0, 1)$. Why is this useful? Well, according to Lemma 3 (or, Remark 1 following the lemma), the first knot $\lambda_1 = f_{\eta^*}$ almost takes the role of W above, except that there is a further level of randomness in η^*, and $\Lambda_{\eta^*}, \mathcal{V}_{\eta^*}$. That is, instead of the expectation of $g(f_{\eta^*})$ being given by $\mathbb{Q}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},\mu_{\eta^*},\sigma^2_{\eta^*}}(g)$, it is given by $\mathbb{E}(\mathbb{Q}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},\mu_{\eta^*},\sigma^2_{\eta^*}}(g))$. The key intuition is that the random variable

$$\mathbb{S}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},\mu_{\eta^*},\sigma^2_{\eta^*}}(f_{\eta^*}) = \mathbb{Q}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},\mu_{\eta^*},\sigma^2_{\eta^*}}(1_{(f_{\eta^*},\infty)}) \tag{40}$$

should still be uniformly distributed, since this is true conditional on $\eta^*, \Lambda_{\eta^*}, \mathcal{V}_{\eta^*}$ and unconditionally, the extra level of randomness in $\eta^*, \Lambda_{\eta^*}, \mathcal{V}_{\eta^*}$ just gets "averaged out" and does not change the distribution. Our next lemma formalizes this intuition and, therefore, provides a test for H_0 based on the (random) survival function in (40).

LEMMA 4 (Kac–Rice pivot). *The survival function of* $\mathbb{Q}_{\Lambda,\mathcal{V},\mu,\sigma^2}$, *with* $\Lambda = \Lambda_{\eta^*}$, $\mathcal{V} = \mathcal{V}_{\eta^*}$, $\mu = \mu_{\eta^*}$, $\sigma^2 = \sigma^2_{\eta^*}$, *and evaluated at* $t = f_{\eta^*}$, *satisfies*

$$\mathbb{S}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},\mu_{\eta^*},\sigma^2_{\eta^*}}(f_{\eta^*}) \sim \mathrm{Unif}(0,1). \tag{41}$$

PROOF. Fix some $h : \mathbb{R} \to \mathbb{R}$. A standard argument shows that (fixing $\Lambda, \mathcal{V}, \mu, \sigma^2$),

$$\mathbb{Q}_{\Lambda,\mathcal{V},\mu,\sigma^2}(h \circ \mathbb{S}_{\Lambda,\mathcal{V},\mu,\sigma^2}) = \int_0^1 h(t)\,dt.$$

Now we compute, applying (38) with g being a composition of functions,

$$\begin{aligned}
&\mathbb{E}\big(h\big(S_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},\mu_{\eta^*},\sigma^2_{\eta^*}}(f_{\eta^*})\big)\big)\\
&\quad= \sum_{j=0}^{p} \int_{\partial_j \mathcal{K}} \mathbb{E}\big(\mathbb{Q}_{\Lambda_\eta,\mathcal{V}_\eta,\mu_\eta,\sigma^2_\eta}(h \circ S_{\Lambda_\eta,\mathcal{V}_\eta,\mu_\eta,\sigma^2_\eta})\mathbb{M}_{\Lambda_\eta,\mathcal{V}_\eta,\mu_\eta,\sigma^2_\eta}(1)\\
&\qquad\times 1_{\{\mathcal{V}^-_\eta \le \mathcal{V}^+_\eta, \mathcal{V}^0_\eta \le 0\}}\big)\psi_{\nabla f_{|T_\eta\mathcal{K}}}(0)\det(G_\eta)\mathcal{H}_j(d\eta)\\
&\quad= \sum_{j=0}^{p} \int_{\partial_j \mathcal{K}} \mathbb{E}\bigg(\bigg[\int_0^1 h(t)\,dt\bigg]\mathbb{M}_{\Lambda_\eta,\mathcal{V}_\eta,\mu_\eta,\sigma^2_\eta}(1)1_{\{\mathcal{V}^-_\eta \le \mathcal{V}^+_\eta, \mathcal{V}^0_\eta \le 0\}}\bigg)\\
&\qquad\times \psi_{\nabla f_{|T_\eta\mathcal{K}}}(0)\det(G_\eta)\mathcal{H}_j(d\eta)\\
&\quad= \bigg[\int_0^1 h(t)\,dt\bigg] \sum_{j=0}^{p} \int_{\partial_j \mathcal{K}} \mathbb{E}\big(\mathbb{M}_{\Lambda_\eta,\mathcal{V}_\eta,\mu_\eta,\sigma^2_\eta}(1)1_{\{\mathcal{V}^-_\eta \le \mathcal{V}^+_\eta, \mathcal{V}^0_\eta \le 0\}}\big)\\
&\qquad\times \psi_{\nabla f_{|T_\eta\mathcal{K}}}(0)\det(G_\eta)\mathcal{H}_j(d\eta)\\
&\quad= \int_0^1 h(t)\,dt.
\end{aligned}$$

□

REMARK 5. In particular, under H_0, we have $\mu_{\eta^*} = 0$, and Lemma 4 shows that

$$\mathbb{S}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},0,\sigma^2_{\eta^*}}(f_{\eta^*}) \sim \mathrm{Unif}(0,1).$$

This proves our main result, Theorem 1, noting that the statistic in (17) is just $\mathbb{S}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},0,\sigma^2_{\eta^*}}(f_{\eta^*})$ written out a little more explicitly.

REMARK 6. We have used the survival function of f_{η^*}, conditional on η^* being the global maximizer, as well the additional local information $(\Lambda_{\eta^*}, \mathcal{V}_{\eta^*})$. Conditioning on this extra information makes the test very simple to compute, at least in the lasso, group lasso and nuclear norm cases. If we were to marginalize over these quantities, we would have a more powerful test. In general, it seems

difficult to analytically marginalize over these quantities, but perhaps Monte Carlo schemes would be feasible. Further, Lemma 4 holds for any μ_{η^*}, implying that this marginalization over $(\Lambda_{\eta^*}, \mathcal{V}_{\eta^*})$ would need access to the unknown μ_{η^*}. Under the global null, $H_0 : \mu_{\eta^*} = 0$, this would not an issue, though it would cause a problem for the selective inference work described in Section 6.3.

5. Practicalities and examples. We now summarize the steps required to apply the Kac–Rice test in an instance of the regularized regression problem (9). The test procedure is to compute the statistic

$$\mathbb{S}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},0,\sigma^2_{\eta^*}}(\lambda_1) = \frac{\int_{\lambda_1}^{\mathcal{V}^+_{\eta^*}} \det(\Lambda_{\eta^*} + zI)\phi_{\sigma^2_{\eta^*}}(z)\,dz}{\int_{\mathcal{V}^-_{\eta^*}}^{\mathcal{V}^+_{\eta^*}} \det(\Lambda_{\eta^*} + zI)\phi_{\sigma^2_{\eta^*}}(z)\,dz}, \tag{42}$$

and compare this against Unif(0, 1). Recalling that $\Lambda_\eta = H_\eta^{-1} G_\eta$, this leaves us with essentially 6 quantities to be computed—$\lambda_1, \mathcal{V}^+_{\eta^*}, \mathcal{V}^-_{\eta^*}, G_{\eta^*}, H_{\eta^*}, \sigma^2_{\eta^*}$ from (11), (24), (23), (29), (30), (32), respectively, and the above integral to be calculated.

If we know the dual seminorm $\mathcal{Q}$ of the penalty $\mathcal{P}$ in closed form, then the first knot λ_1 can be found explicitly from (11); otherwise, it can be found numerically by solving the (convex) optimization problem

$$\lambda_1 = \max_{\eta \in \mathbb{R}^p} \eta^T X^T (I - P_{XC^\perp}) y \quad \text{subject to} \quad \mathcal{P}(\eta) \le 1.$$

The remaining quantities, $\mathcal{V}^+_{\eta^*}, \mathcal{V}^-_{\eta^*}, G_{\eta^*}, H_{\eta^*}, \sigma^2_{\eta^*}$, all depend on η^* and on the tangent space $T_{\eta^*}\mathcal{K}$. Again, depending on $\mathcal{Q}$, the maximizer η^* can either be found in closed form, or numerically by solving the above optimization problem. Once we know the projection operator onto the tangent space $T_{\eta^*}\mathcal{K}$, (32) is an explicit expression for $\sigma^2_{\eta^*}$; furthermore, $\mathcal{V}^-_{\eta^*}, \mathcal{V}^+_{\eta^*}$ are given by two more tractable (convex) optimization problems (which in some cases admit closed form solutions); see Section A.6 of the supplement [17].

The quantities G_{η^*}, H_{η^*} are different; however, even once we know η^* and the tangent space $T_{\eta^*}\mathcal{K}$, finding G_{η^*}, H_{η^*} involves computing the Hessian $\nabla^2 f_{|T_{\eta^*}\mathcal{K}}$, which requires a geometric understanding of the curvature of f around $T_{\eta^*}\mathcal{K}$. That is, G_{η^*}, H_{η^*} cannot be calculated numerically (say, via an optimization procedure, as with $\lambda_1, \mathcal{V}^-_{\eta^*}, \mathcal{V}^+_{\eta^*}$), and demand a more problem-specific, mathematical focus. For this reason, computation of G_{η^*}, H_{η^*} can end up being an involved process (depending on the problem). In the examples that follow, we do not give derivation details for the Hessian $\nabla^2 f_{|T_{\eta^*}\mathcal{K}}$, but refer the reader to [1] for the appropriate background material.

To see a high-level summary of the quantities involved, see Algorithm 1.

We now revisit the lasso example, and then consider the group lasso and nuclear norm penalties, the latter yielding applications to principal components and matrix

Algorithm 1 Computing the Kac–Rice pivot

1: Solve for λ_1 and η^* {note: we use concise notation throughout, see Section 2.4}
2: Form an orthonormal basis V_{η^*} of the tangent space $T_{\eta^*}\mathcal{K}$
3: Compute the projection $P_{\eta^*,X,\Sigma}$ in (19)
4: Evaluate the conditional variance $\sigma^2_{\eta^*}$ and $C_{X,\Sigma}(\eta^*)$ from (32) and (22)
5: **if** $\nabla^2 f_{|T_{\eta^*}\mathcal{K}} = 0$ **then**
6: Let $\Lambda_{\eta^*} = 0$
7: **else if** $\nabla^2 f_{|T_{\eta^*}\mathcal{K}} \neq 0$ **then**
8: Let $\Lambda_{\eta^*} = G^{-1}_{\eta^*} H_{\eta^*}$ from (29) and (30)
9: **end if**
10: Solve the optimization problems (23) and (24), yielding $\mathcal{V}^-_{\eta^*}$, $\mathcal{V}^+_{\eta^*}$
11: Evaluate the integrals in (42) to obtain $\mathbb{S} = \mathbb{S}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},0,\sigma^2_{\eta^*}}(\lambda_1)$

completion. We remark that in the lasso and group lasso cases, the matrix $\Lambda_{\eta^*} = G^{-1}_{\eta^*} H_{\eta^*}$ is zero, simplifying the computations. In contrast, it is nonzero for the nuclear norm case.

Also, it is important to point out that in all three problem cases, we have $\operatorname{span}(C) = \mathbb{R}^p$, so the notational shortcut that we applied in Sections 3 and 4 has no effect (see Section 2.4), and we can use the formulae from these sections as written.

5.1. *Example*: *The lasso* (*revisited*). For the lasso problem (3), we have $\mathcal{P}(\beta) = \|\beta\|_1$ and $C = \{u : \|u\|_\infty \leq 1\}$, so $\mathcal{Q}(\beta) = \|\beta\|_\infty$ and $\mathcal{K} = C^\circ = \{v : \|v\|_1 \leq 1\}$. Our Morse assumption on the process $f_\eta = \eta^T X^T y$ over $\mathcal{K}$ (which amounts to an assumption on the predictor matrix X) implies that there is a unique index j^* such that

$$\lambda_1 = |X^T_{j^*} y| = \|X^T y\|_\infty = \max_{\|\eta\|_1 \leq 1} \eta^T X^T y.$$

Then in this notation $\eta^* = \operatorname{sign}(X^T_{j^*} y) \cdot e_{j^*}$ (where e_{j^*} is the j^*th standard basis vector), and the normal cone to $\mathcal{K}$ at η^* is

$$N_{\eta^*}\mathcal{K} = \{v \in \mathbb{R}^p : \operatorname{sign}(v_{j^*}) = \operatorname{sign}(X^T_{j^*} y), |v_j| \leq |v_{j^*}| \text{ for all } j \neq j^*\}.$$

Because this is a full-dimensional set, the tangent space to $\mathcal{K}$ at η^* is $T_{\eta^*}\mathcal{K} = (N_{\eta^*}\mathcal{K})^\perp = \{0\}$. This greatly simplifies our survival function test statistic (40) since all matrices in consideration here are 0×0 and, therefore, have determinant 1, giving

$$\mathbb{S}_{\Lambda_{\eta^*},\mathcal{V}_{\eta^*},0,\sigma^2_{\eta^*}} = \frac{\Phi(\mathcal{V}^-_{\eta^*}/\sigma_{\eta^*}) - \Phi(\lambda_1/\sigma_{\eta^*})}{\Phi(\mathcal{V}^-_{\eta^*}/\sigma_{\eta^*}) - \Phi(\mathcal{V}^+_{\eta^*}/\sigma_{\eta^*})}.$$

The lower and upper limits $\mathcal{V}^+_{\eta^*}, \mathcal{V}^-_{\eta^*}$ are easily computed by solving two linear fractional programs; see Section A.6 of the supplement [17]. The variance $\sigma^2_{\eta^*}$ of f_{η^*} is given by (32), and again simplifies because $T_{\eta^*}\mathcal{K}$ is zero dimensional, becoming

$$\sigma^2_{\eta^*} = (\eta^*)^T X^T \Sigma X \eta^* = X^T_{j^*} \Sigma X_{j^*}.$$

Plugging in this value gives the test statistic as in (6) in Section 1.1. The reader can return to this section for examples and discussion in the lasso case.

5.2. *Example*: *The group lasso*. The group lasso [23] can be viewed as an extension of the lasso for grouped (rather than individual) variable selection. Given a predefined collection $\mathcal{G}$ of groups, with $\bigcup_{g\in\mathcal{G}} g = \{1, \ldots, p\}$, the group lasso penalty is defined as

$$\mathcal{P}(\beta) = \sum_{g=1}^{G} w_g \|\beta_g\|_2,$$

where $\beta_g \in \mathbb{R}^{|g|}$ denotes the subset of components of $\beta \in \mathbb{R}^p$ corresponding to g, and $w_g > 0$ for all $g \in \mathcal{G}$. We note that

$$C = \{u \in \mathbb{R}^p : \|u_g\|_2 \le w_g, g \in \mathcal{G}\},$$

so the dual of the penalty is

$$\mathcal{Q}(\beta) = \max_{g\in\mathcal{G}} w_g^{-1} \|\beta_g\|_2,$$

and

$$\mathcal{K} = C^\circ = \left\{v \in \mathbb{R}^p : \sum_{g\in\mathcal{G}} w_g \|v_g\|_2 \le 1\right\}.$$

Under the Morse assumption on $f_\eta = \eta^T X^T y$ over $\mathcal{K}$ (again, this corresponds to an assumption about the design matrix X), there is a unique group g^* such that

$$\lambda_1 = w_{g^*}^{-1} \|X^T_{g^*} y\|_2 = \max_{g\in\mathcal{G}} w_g^{-1} \|X^T_g y\|_2 = \max_{\sum_{g\in\mathcal{G}} w_g \|\eta_g\|_2 \le 1} \eta^T X^T y,$$

where we write $X_g \in \mathbb{R}^{n\times|g|}$ to denote the matrix whose columns are a subset of those of X, corresponding to g. Then the maximizer η^* is given by

$$\eta^*_g = \begin{cases} \dfrac{X^T_g y}{w_g \|X^T_g y\|_2}, & \text{if } g = g^*, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for all } g \in \mathcal{G},$$

and the normal cone $N_{\eta^*}\mathcal{K}$ is seen to be

$$N_{\eta^*}\mathcal{K} = \{v \in \mathbb{R}^p : v_{g^*} = cX^T_{g^*} y, \|v_g\|_2/w_g \le c\|X^T_{g^*} y\|_2/w_{g^*} \\ \text{for all } g \ne g^*, c \ge 0\}.$$

Hence, the tangent space $T_{\eta^*}\mathcal{K} = (N_{\eta^*}\mathcal{K})^\perp$ is

$$T_{\eta^*}\mathcal{K} = \{u \in \mathbb{R}^p : u_{g^*}^T X_{g^*}^T y = 0, u_g = 0 \text{ for all } g \neq g^*\},$$

which has dimension $r^* - 1$, with $r^* = \operatorname{rank}(X_{g^*})$. An orthonormal basis V_{η^*} for this tangent space is given by padding an orthonormal basis for $(\operatorname{span}(X_{g^*}^T y))^\perp$ with zeros appropriately. From this, we can compute the projection operator

$$P_{\eta^*,X,\Sigma} = \Sigma X V_{\eta^*} \big(V_{\eta^*}^T X^T \Sigma X V_{\eta^*}\big)^\dagger V_{\eta^*}^T X^T,$$

and the variance of f_{η^*} as

$$\sigma_{\eta^*}^2 = \frac{1}{w_{g^*}^2 \|X_{g^*}^T y\|_2^2} y^T X_{g^*} X_{g^*}^T (I - P_{\eta^*,X,\Sigma}) \Sigma X_{g^*} X_{g^*}^T y.$$

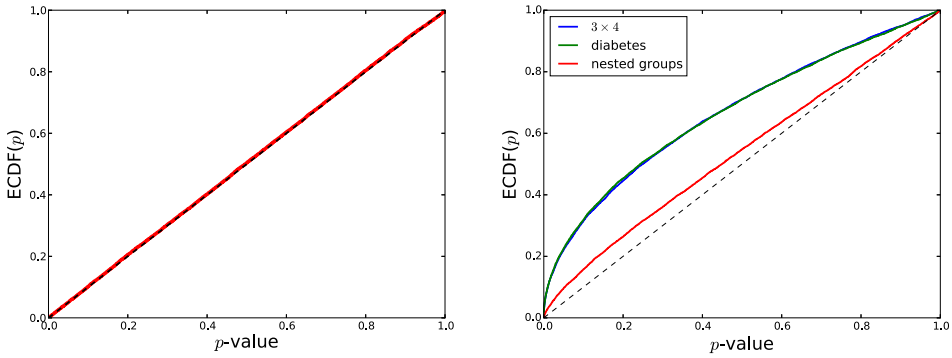
The quantities $\mathcal{V}_{\eta^*}^-, \mathcal{V}_{\eta^*}^+$ can be readily computed by solving two convex programs, see Section A.6 of the supplementary document [17]. Finally, we have $H_{\eta^*} = 0$ in the group lasso case, as the special form of curvature matrix of a sphere implies that $G_{\eta^*} \widetilde{f}_{\eta^*}^{\eta^*} = -\nabla^2 f_{|T_{\eta^*}\mathcal{K}}$ in (29). This makes $\Lambda_{\eta^*} = G_{\eta^*}^{-1} H_{\eta^*} = 0$, and the test statistic (42) for the group lasso problem becomes

$$\frac{\int_{\lambda_1}^{\mathcal{V}_{\eta^*}^+} z^{r^*-1} \phi_{\sigma_{\eta^*}^2}(z)\, dz}{\int_{\mathcal{V}_{\eta^*}^-}^{\mathcal{V}_{\eta^*}^+} z^{r^*-1} \phi_{\sigma_{\eta^*}^2}(z)\, dz} = \frac{\mathbb{P}(\chi_{r^*} \leq \mathcal{V}_{\eta^*}^+/\sigma_{\eta^*}) - \mathbb{P}(\chi_{r^*} \leq \lambda_1/\sigma_{\eta^*})}{\mathbb{P}(\chi_{r^*} \leq \mathcal{V}_{\eta^*}^+/\sigma_{\eta^*}) - \mathbb{P}(\chi_{r^*} \leq \mathcal{V}_{\eta^*}^-/\sigma_{\eta^*})}. \tag{43}$$

In the above, χ_{r^*} denotes a chi distributed random variable with r^* degrees of freedom, and the equality follows from the fact that the missing multiplicative factor in the χ_{r^*} density [namely, $2^{1-r^*/2}/\Gamma(r^*/2)$] is common to the numerator and denominator, and hence cancels.

Figure 2(a) shows the empirical distribution function of a sample of 20,000 p-values from problem instances sampled randomly from a variety of different group lasso setups (all under the global null model $\beta_0 = 0$):

- small case: X is 3×4, a fixed matrix slightly perturbed by Gaussian noise; there are 2 groups of size 2, one with weight $\sqrt{2}$, the other with weight 0.1;
- fat case: X is $100 \times 10{,}000$ with features drawn from the compound symmetric Gaussian distribution having correlation 0.5; here are 1000 groups each of size 10, each having weight $\sqrt{10}$;
- tall case: X is $10{,}000 \times 100$ with features drawn from the compound symmetric Gaussian distribution having correlation 0.5; there are 1000 groups each of size 10, each having weight $\sqrt{10}$;
- square case: X is 100×100 with features drawn from the compound symmetric Gaussian distribution having correlation 0.5; there are 10 groups each of size 10, each having weight $\sqrt{10}$;



(a) Kac-Rice test (b) Covariance test

FIG. 2. *The left panel shows the empirical distribution function of a sample of* 20,000 *p-values* (43) *computed from various group lasso setups, which agrees very closely with the uniform distribution. The right panel shows the empirical distribution functions in three different group lasso setups, each over* 10,000 *samples, when p-values are instead computed using an* Exp(1) *approximation for the covariance test. This approximation ends up being anti-conservative whereas the Kac–Rice test is exact.*

- diabetes case 1: X is 442×10, the diabetes data set from Efron et al. [7]; there are 4 (arbitrarily created) groups: one of size 4, one of size 2, one of size 3 and one of size 1, with varying weights;
- diabetes case 2: X is 442×10, the diabetes data set from Efron et al. [7]; there are now 10 groups of size 1 with i.i.d. random weights drawn from $1 + 0.2\,\mathrm{Unif}(0, 1)$ (generated once for the entire simulation);
- nested case 1: X is 100×10, with two nested groups (the column space for one group of size 2 is contained in that of the other group of size 8) with the weights favoring inclusion of the larger group first;
- nested case 2: X is 100×10, with two nested groups (the column space for one group of size 2 is contained in that of the other group of size 8) with the weights favoring inclusion of the smaller group first;
- nested case 3: X is 100×12, with two sets of two nested groups (in each set, the column space for one group of size 2 is contained in that of the other group of size 4) with the weights chosen according to group size;
- nested case 4: X is 100×120, with twenty sets of two nested groups (in each set, the column space for one group of size 2 is contained in that of the other group of size 4) with the weights chosen according to group size.

As we can see from the plot, the p-values are extremely close to uniform.

In comparison, arguments similar to those given in Lockhart et al. [10] for the lasso case would suggest that for the group lasso, under the null hypothesis,

$$\frac{\lambda_1(\lambda_1 - \mathcal{V}^-_{\eta^*})}{\sigma^2_{\eta^*}} \xrightarrow{d} \mathrm{Exp}(1) \qquad \text{as } n, p \to \infty,$$

under some conditions (one of these being that $\mathcal{V}^-_{\eta^*}$ diverges to ∞ fast enough). Figure 2(b) shows the empirical distribution function of 10,000 samples from three of the above scenarios, demonstrating that, while asymptotically reasonable, the Exp(1) approximation for the covariance test in the group lasso case can be quite anti-conservative in finite samples.

5.3. *Example*: *Nuclear norm*. In this setting, we treat the coefficients in (9) as a matrix, instead of a vector, denoted by $B \in \mathbb{R}^{n\times p}$. We consider a nuclear norm penalty on B,

$$\mathcal{P}(B) = \|B\|_* = \operatorname{tr}(D),$$

where D is the diagonal matrix of singular values in the singular value decomposition $B = UDV^T$. Here, the dual seminorm is

$$\mathcal{Q}(B) = \|B\|_{\mathrm{op}} = \max(D),$$

the operator norm (or spectral norm) of B, that is, its maximum singular value. Therefore, we have

$$C = \{A : \|A\|_{\mathrm{op}} \le 1\},$$
$$\mathcal{K} = C^\circ = \{W : \|W\|_* \le 1\}.$$

Examples of problems of the form (9) with nuclear norm penalty $\mathcal{P}(B) = \|B\|_*$ can be roughly categorized according to the choice of linear operator $X = X(B)$. For example:

- *principal components analysis*: if $X : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ is the identity map, then λ_1 is the largest singular value of $y \in \mathbb{R}^{n\times p}$, and moreover, $\mathcal{V}^-_{\eta^*}$ is the second largest singular value of y;
- *matrix completion*: if $X : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ zeros out all of the entries of its argument outside some index set $\mathcal{O} \subseteq \{1,\dots,n\} \times \{1,\dots,p\}$, and leaves the entries in $\mathcal{O}$ untouched, then problem (9) is a noisy version of the matrix completion problem [5, 12];
- *reduced rank regression*: if $X : \mathbb{R}^{n\times p} \to \mathbb{R}^{m\times p}$ performs matrix multiplication, $X(B) = XB$, and $y \in \mathbb{R}^{m\times p}$, then problem (9) is often referred to as reduced rank regression [13].

The first knot in the solution path is given by $\lambda_1 = \|X^T(y)\|_{\mathrm{op}}$, with X^T denoting the adjoint of the linear operator X. Assuming that $X^T(y)$ has singular value decomposition $X^T(y) = UDV^T$ with $D = \operatorname{diag}(d_1, d_2, \dots)$ for $d_1 \ge d_2 \ge \cdots$, and that the process $f_\eta = \langle \eta, X^T(y)\rangle$ is Morse over $\eta \in \mathcal{K}$, there is a unique $\eta^* \in \mathcal{K}$ achieving the value λ_1,

$$\eta^* = U_1 V_1^T,$$

where U_1, V_1 are the first columns of U, V, respectively. The normal cone $N_{\eta^*}\mathcal{K}$ is

$$N_{\eta^*}\mathcal{K} = \{cU_1V_1^T + c\widetilde{U}\widetilde{D}\widetilde{V}^T : \widetilde{U}^T\widetilde{U} = I, \widetilde{V}^T\widetilde{V} = I, \widetilde{D} = \operatorname{diag}(\widetilde{d}_1, \widetilde{d}_2, \ldots), U_1^T\widetilde{U} = 0, V_1^T\widetilde{V} = 0, \max(\widetilde{D}) \le 1, c \ge 0\},$$

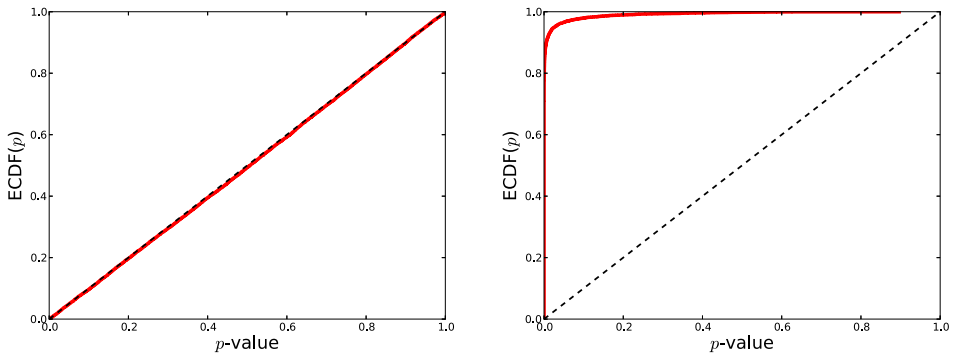
and so the tangent space $T_{\eta^*}\mathcal{K} = (N_{\eta^*}\mathcal{K})^\perp$ is

$$T_{\eta^*}\mathcal{K} = \operatorname{span}(\{U_1V_j^T, j = 2, \ldots, p\} \cup \{U_jV_1^T, j = 2, \ldots, n\}).$$

From this tangent space, the marginal variance $\sigma^2_{\eta^*}$ in (32) can be easily computed. This leaves $\mathcal{V}^-_{\eta^*}, \mathcal{V}^+_{\eta^*}, G_{\eta^*}, H_{\eta^*}$ to be addressed. As always, the quantities $\mathcal{V}^-_{\eta^*}, \mathcal{V}^+_{\eta^*}$ can be determined numerically, as the optimal values of two convex programs; see Section A.6 of the supplementary document [17], though these problems admit closed form solutions. We direct the reader to Section A.5 of the supplement [17] for details of the computations.

In Figure 3, we plot the empirical distribution function of a sample of 20,000 p-values computed over problem instances that have been randomly sampled from the following scenarios, all employing the nuclear norm penalty (and all under the null model $B_0 = 0$, with B_0 being the underlying coefficient matrix):

- *principal components analysis*: y is $2 \times 2, 3 \times 4, 50 \times 50, 100 \times 20, 30 \times 1000, 30 \times 5, 1000 \times 1000$;
- *matrix completion*: y is 10×5 with 50% of its entries observed at random, 100×30 with 20% of its entries observed at random, 10×5 with a nonrandom pattern of observed entries, 20×10 with a nonrandom pattern of observed entries, 200×10 with 10% of its entries observed at random;



(a) Our test (b) Covariance test

FIG. 3. *The left panel shows the empirical distribution function of a sample of* 20,000 *p-values computed over a variety of problem setups that utilize the nuclear norm penalty. The right panel shows the distribution of* 20,000 *covariance test p-values when an* Exp(1) *approximation is used, which shows the exponential approximation to be clearly inappropriate for the nuclear norm setting.*

- *reduced rank regression*: X is 100×10 whose entries are drawn from the compound symmetric Gaussian distribution with correlation 0.5, and y is 100×5.

As is evident in the plot, the agreement with uniform is excellent [as above, each particular scenario above produces Unif(0, 1) p-values regardless of how X was chosen, modulo the Morse assumption].

Again, along the lines of the covariance test, we consider approximation of $\lambda_1(\lambda_1 - \mathcal{V}^-_{\eta^*})/\sigma^2_{\eta^*}$ by an Exp(1) distribution under the null hypothesis. Figure 3(b) shows that this approximation is quite far off, certainly much more so than in the other examples. Preliminary calculations confirm mathematically that the Exp(1) distribution is not the right limiting distribution here; we will pursue this in future work.

6. Discussion. We derived an exact (nonasymptotic) p-value for testing a global null hypothesis in a general regularized regression setup. Our test is based on a geometric characterization of regularized regression estimators and the Kac–Rice formula, and has a close relationship to the covariance test for the lasso problem [10]. In fact, the Exp(1) limiting null distribution of the covariance test can be derived from the formulae given here. These two tests give similar results for lasso problems, but our new test has exact (not asymptotic) error control under the null, with less assumptions on the predictor matrix X.

Another strength of our approach is that it provably extends well beyond the lasso problem; in this paper, we examine tests for both the group lasso and nuclear norm regularization problems. Still, the test can be applied outside of these cases too, and is limited only by the difficulty in evaluating the p-value in practice (which relies on geometric quantities to be computed).

We recall that the covariance test for the lasso can be applied at any knot along the solution path, to test if the coefficients of the predictors not yet in the current model are all zero. In other words, it can be used to test more refined null hypotheses, and not only the global null. Our paper only addresses global testing, but since it was written, there has been much follow-up work in developing tests for the null hypotheses visited further down the regularization path, in various problem settings of interest. We discuss these advances later, and first discuss the important issues of non-Gaussian errors, and power.

6.1. *Non-Gaussian errors*. Throughout, our calculations have rather explicitly used the fact that $X^T y$ is Gaussian distributed. One could potentially appeal to the central limit theorem if the components of $y - X\beta_0$ are i.i.d. from some error distribution (treating X as fixed), though the calculations in this work focus on the extreme values, so the accuracy of the central limit theorem in the tails may be in doubt. We do not address these theoretical issues here, but instead consider simulations under heavy-tailed, skewed noise. In particular, we drew errors according to a t-distribution with 5 degrees of freedom plus an independent centered Exp(1)

distribution. Figure A.8 in the Appendix [17] shows that the lasso and group lasso p-values are relatively well behaved, while the nuclear norm p-values seem to break down.

6.2. *Power.* An important issue raised by the editors and referees of this paper concerns the power of our proposed tests. As correctly noted by the editors and referees, the Kac–Rice test is not optimal in terms of power, in a number of practical scenarios. For example, consider the Kac–Rice test in the lasso case, that is, with the ℓ_1 regularizer $\mathcal{P}(\beta) = \|\beta\|_1$, and the associated the global testing problem $H_0 : \beta_0 = 0$. When the design X is orthogonal, it is known that the most powerful test against sparse alternatives is the max test, which rejects based on $\lambda_1 = \|X^T y\|_\infty$; for example, see Arias-Castro, Candès and Plan [2]. The Kac–Rice test, on the other hand, rejects based on a monotone function of $\lambda_1 - \lambda_2$, the first two largest absolute entries of $X^T y$; see (7). Both theoretical and empirical calculations show that a test based on the gap $\lambda_1 - \lambda_2$ is not as powerful as a test based solely on λ_1, though the former still can have substantial power.

More examples along these lines can be drawn up, showing that the Kac–Rice test, while still providing nontrivial power in a wide variety of specific applications, falls short when compared to the most powerful test against a particular class of alternatives. This is not surprising, and we feel, should not be alarming either. If one has a particular global testing setup in mind, with a particular set of alternatives in mind, then of course one should use the most powerful test when available. The strength of the Kac–Rice framework developed in this paper does not lie in its power for global testing problems such as those considered in the lasso, group lasso or nuclear norm cases discussed in Sections 5.1, 5.2 and 5.3, respectively. Its real strengths are twofold:

- it provides a single unified framework under which we can derive exact tests for global hypotheses, shown to be applicable in a wide range of settings, and in principle at least, applicable in settings that do not currently possess well-established global testing theory;
- it leads to exact tests beyond the global null, of selective hypotheses along the regularization path, which as far as we can tell was not generically possible before this work.

The second point above is further discussed below.

6.3. *Related work on selective inference.* The global Kac–Rice framework proposed here has spawned new work on selective hypothesis testing, in each of the major example areas considered (lasso, group lasso and nuclear norm problems). Selective inference is a term used to describe hypothesis testing when the hypotheses themselves are generated or *selected* based on observable data. It is a new field with a small but growing literature, and we believe, has an important place in the modern statistical toolkit full of adaptive regression and classification procedures.

We refer the reader to [9] for a fairly comprehensive theoretical development of this topic.

After the current paper was written, our Kac–Rice framework has led to selective tests—that is, tests for hypothesis beyond the global null, at later steps in the regularization path—for least angle regression and the lasso [16], grouped stepwise regression [11] and principal components analysis [6]. To the best of our knowledge, these papers are among the first available rigorous inferential tests for the randomly selected hypotheses visited in these complex procedures. In a sense, each of the aforementioned works were consequences of the Kac–Rice framework, obtained by extending the arguments laid out in Sections 5.1, 5.2 and 5.3, respectively. We are not aware of other global testing frameworks that can also accommodate such extensions, in general.

6.4. *Gaussian random field theory.* Finally, the construction of the process $\widetilde{f}_\eta^\eta$ is a contribution to the theory of smooth Gaussian random fields as described in [1]. Our construction allows earlier proofs that apply only to (centered) Gaussian random fields of constant variance (i.e., marginally stationary) to smooth random fields with arbitrary variance. Even in the marginally stationary case, the conditional distribution $\mathbb{Q}$ defined in (38) provides a new tool for exact selective inference at critical points of such random fields. We leave this, and many other topics, for future work.

Acknowledgments. We are indebted to Robert Tibshirani, without whose help and suggestions this work would not have been possible. Two referees and an Associate Editor provided feedback on preliminary versions of this work that has helped the authors improve this paper.

SUPPLEMENTARY MATERIAL

Supplement to "Inference in adaptive regression via the Kac–Rice formula" (DOI: 10.1214/15-AOS1386SUPP; .pdf). We provide some proofs, as well as supplementary details on computing the Kac–Rice test.

REFERENCES

[1] ADLER, R. J. and TAYLOR, J. E. (2007). *Random Fields and Geometry*. Springer, New York. MR2319516

[2] ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

[3] AZAÏS, J.-M. and WSCHEBOR, M. (2008). A general expression for the distribution of the maximum of a Gaussian field and the approximation of the tail. *Stochastic Process. Appl.* **118** 1190–1218. MR2428714

[4] BRILLINGER, D. R. (1972). On the number of solutions of systems of random equations. *Ann. Math. Statist.* **43** 534–540. MR0300329

[5] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240
[6] CHOI, Y., TAYLOR, J. and TIBSHIRANI, R. (2014). Selecting the number of principal components: Estimation of the true rank of a noisy matrix. Preprint. Available at arXiv:1410.8260.
[7] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166
[8] FEDERER, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078
[9] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.
[10] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970
[11] LOFTUS, J. and TAYLOR, J. (2014). A significance test for forward stepwise model selection. Preprint. Available at arXiv:1405.3920.
[12] MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857
[13] MUKHERJEE, A., CHEN, K., WANG, N. and ZHU, J. (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika* **102** 457–477. MR3371016
[14] TAKEMURA, A. and KURIKI, S. (2002). On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. *Ann. Appl. Probab.* **12** 768–796. MR1910648
[15] TAYLOR, J. (2013). The geometry of least squares in the 21st century. *Bernoulli* **19** 1449–1464. MR3102558
[16] TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). Exact post-selection inference for sequential regression procedures. Preprint. Available at arXiv:1401.3889.
[17] TAYLOR, J., LOFTUS, J. and TIBSHIRANI, R. J. (2015). Supplement to "Inference in adaptive regression via the Kac–Rice formula." DOI:10.1214/15-AOS1386SUPP.
[18] TAYLOR, J., TAKEMURA, A. and ADLER, R. J. (2005). Validity of the expected Euler characteristic heuristic. *Ann. Probab.* **33** 1362–1396. MR2150192
[19] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[20] TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641
[21] TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375
[22] TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205
[23] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

J. E. TAYLOR
J. R. LOFTUS
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: jtaylo@stanford.edu
joftius@stanford.edu

R. J. TIBSHIRANI
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
BAKER HALL
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: ryantibs@cmu.edu