Biometrika (2012), **99**, 1, *pp*. 1–8

Printed in Great Britain

Advance Access publication on 31 July 2012

Effective degrees of freedom: a flawed metaphor

BY LUCAS JANSON, WILLIAM FITHIAN AND TREVOR J. HASTIE

Department of Statistics, Sequoia Hall, 390 Serra Mall,
Stanford University, Stanford, California 94305-4065, U.S.A.

ljanson@stanford.edu wfithian@stanford.edu hastie@stanford.edu

SUMMARY

To most applied statisticians, a fitting procedure's degrees of freedom is synonymous with its model complexity, or its capacity for overfitting to data. In particular, it is often used to parameterize the bias-variance tradeoff in model selection. We argue that, on the contrary, model complexity and degrees of freedom may correspond very poorly. We exhibit and theoretically explore various fitting procedures for which degrees of freedom is not monotonic in the model complexity parameter, and can exceed the total dimension of the ambient space even in very simple settings. We show that the degrees of freedom for any non-convex projection method can be unbounded.

Some key words: Model complexity; Number of parameters; Optimism.

1. INTRODUCTION

1·1. *A motivating example: best-subsets regression*

Consider observing data $y = \mu + \varepsilon$ with $\mu \in \mathbb{R}^n$ and independent errors $\varepsilon \sim N(0, \sigma^2 I_n)$, and producing an estimate of μ. A critical property of any estimator $\hat{\mu}$ is its so-called model complexity; informally, how flexibly it is able to conform to the observed response y. Commonly we set $\mu = X\beta$ for some $n \times p$ design matrix X.

For the ordinary-least-squares estimator $\hat{\mu}$, the most natural measure of complexity is the number p of fitted parameters, i.e., the degrees of freedom. For more general estimators, the effective degrees of freedom of Efron (1986), defined as $\sigma^{-2} \sum_{i=1}^{n} \operatorname{cov}\left(y_i, \hat{\mu}_i\right)$, has emerged as a popular and convenient measuring stick for comparing the complexity of very different fitting procedures. The name suggests that a method with p degrees of freedom has a complexity comparable to linear regression on p predictor variables, for which the effective degrees of freedom is p.

Now, suppose that instead of fitting a linear model using all p predictors, we fit the best-subsets regression of size k. That is, we find the best linear model that uses only a subset of k predictors, with $k < p$. What is the effective degrees of freedom, henceforth just degrees of freedom, of this method?

A simple, intuitive, and wrong argument predicts that the degrees of freedom, which depends on μ, is somewhere between k and p. We certainly expect it to be greater than k, since we use the data to select the best model of size k among all the possibilities. However, we have only p free parameters at our disposal, of which $p - k$ must be set to zero, so best-subsets regression with k parameters is still less complex than the saturated model with all p parameters and no constraints.

As convincing as this argument may seem, it is contradicted by a simple simulation with $n = 50$ and $p = 15$. Here $\mu = X\beta$, where the entries of X are independent $N(0, 1)$ variates

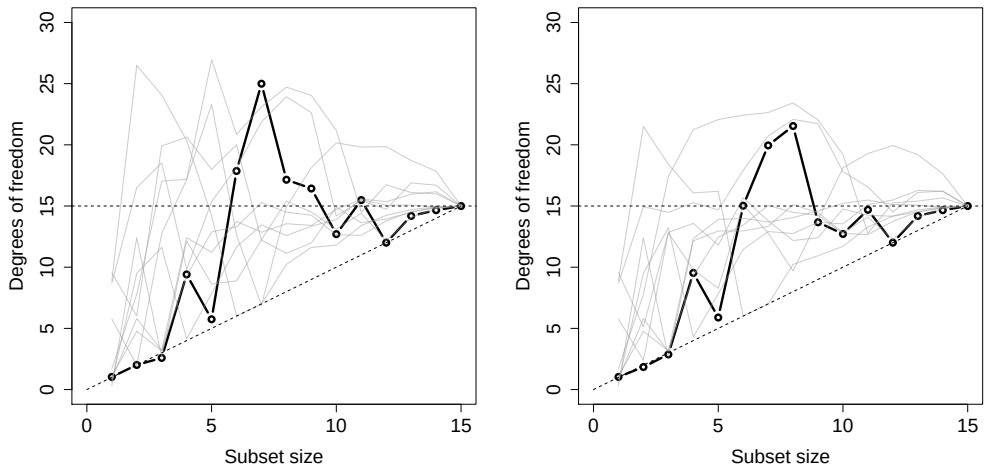

Fig. 1: Estimated degrees of freedom versus subset size for 10 realizations of μ, with one highlighted for clarity, in the simulation described in Section 1·1 using best-subsets regression on the left, and forward selection on the right. Simulation details, including code, are provided in the Supplementary Material. Standard errors for all points are below 0.5, and are not shown. Dashed lines show the constant ambient dimension p and the subset size k, for reference.

and the coefficients β_j are independent $N(0, 4)$ variates, and $\sigma^2 = 1$. Figure 1 shows that the degrees of freedom for both best-subsets regression and another method, forward selection, can exceed the ambient dimension p for values of $k < p$. The degrees of freedom of best-subsets regression exceeded p for some $k < p$ in 179 of 200 realizations of μ, or about 90% of the time. To understand why our intuition should lead us astray here, we must first review why effective degrees of freedom is defined as it is, and what classical concepts the definition is meant to generalize.

1·2. *Degrees of freedom in classical statistics*

The original meaning of degrees of freedom, the number of dimensions in which a random vector may vary, plays a central role in classical statistics. In ordinary linear regression with full-rank $n \times p$ predictor matrix X, the fitted response $\hat{\mu} = X\hat{\beta}$ is the orthogonal projection of y onto the p-dimensional column space of X, and the residual $r = y - \hat{\mu}$ is the projection onto its orthogonal complement, whose dimension is $n - p$. We say this linear model has p model degrees of freedom, with $n - p$ residual degrees of freedom.

If the error variance is σ^2, then r is constrained to have zero projection in p directions, and is free to vary, with variance σ^2, in the remaining $n - p$ orthogonal directions. In particular, if the model is correct, so that $E(y) = X\beta$, then the residual sum of squares, $\|r\|_2^2$, has distribution $\|r\|_2^2 \sim \sigma^2\chi^2_{n-p}$, leading to the unbiased variance estimate $\hat{\sigma}^2 = \|r\|_2^2/(n - p)$. More generally, t-tests and F-tests are based on comparing lengths of n-variate Gaussian random vectors after projecting onto appropriate linear subspaces.

In linear regression, the model degrees of freedom, henceforth just degrees of freedom, serves to quantify multiple related properties of the fitting procedure. The degrees of freedom coincides with the number of non-redundant free parameters in the model, and thus constitutes a natural

measure of model complexity or overfitting. In addition, the total variance of the fitted response $\hat{\mu}$ is exactly $\sigma^2 p$, which depends only on the number of linearly independent predictors and not on their size or correlation with each other.

The degrees of freedom also quantifies the optimism of the residual sum of squares as an estimate of out-of-sample prediction error. In linear regression, one can easily show that the residual sum of squares underestimates mean squared prediction error by $2\sigma^2 p$ on average. Mallows (1973) exploits this identity as a means of model selection, by computing the C_p statistic $\|r\|_2^2 + 2\hat{\sigma}^2 p$, an unbiased estimate of prediction error, for several models, and selecting the model with the smallest estimated test error. In this case the degrees of freedom of a model contributes a penalty to account for that model's complexity.

1·3. *Effective degrees of freedom*

For more general fitting procedures such as smoothing splines, generalized additive models, lasso, or ridge regression, the number of free parameters is often an inappropriate measure of model complexity. Typically these methods have a tuning parameter, but it is not clear a priori how to compare, e.g., a lasso fit with Lagrange parameter $\lambda = 3$ to a local regression fit with window width 0.5. When comparing different methods, or the same method with different tuning parameters, it can be quite useful to have some measure of complexity with a consistent meaning across a range of algorithms. To this end, various authors have proposed alternative more general definitions for the effective degrees of freedom of a method; see Buja et al. (1989) and references therein.

If the method is linear, that is, if $\hat{\mu} = Hy$ for some fixed hat matrix H, then the trace of H serves as a natural generalization. For linear regression H is a p-dimensional projection, so $\mathrm{tr}(H) = p$, coinciding with the original definition. Intuitively, when H is not a projection, $\mathrm{tr}(H)$ accumulates fractional degrees of freedom for directions of y that are shrunk, but not entirely eliminated, in computing $\hat{\mu}$.

For nonlinear methods, further generalization is necessary. The most popular definition, due to Efron (1986) and given in Equation (1), defines degrees of freedom in terms of the optimism of residual sum of squares as an estimate of test error, and applies to any fitting method.

Measuring or estimating optimism is a worthy goal in and of itself. But to justify our intuition that the degrees of freedom offers a consistent way to quantify model complexity, a bare requirement is that the degrees of freedom be monotone in model complexity when considering a fixed method. The term model complexity is itself rather metaphorical when describing arbitrary fitting algorithms, but has a concrete meaning for methods that minimize residual sum of squares subject to the fit $\hat{\mu}$ belonging to a closed constraint set $\mathcal{M}$, a model. Commonly, some tuning parameter γ indexes a nested set of models $\mathcal{M}_\gamma$, with $\gamma_1 \leq \gamma_2$ implying $\mathcal{M}_{\gamma_1} \subseteq \mathcal{M}_{\gamma_2} \subseteq \mathbb{R}^n$. Then the fitted vector for a tuning parameter γ is

$$\hat{\mu}^{(\gamma)} = \underset{z \in \mathcal{M}_\gamma}{\arg\min} \|y - z\|_2^2.$$

Examples include the lasso (Tibshirani, 1996) and ridge regression (Hoerl, 1962) in their constraint formulation, as well as best-subsets regression. The model $\mathcal{M}_k$ for best-subsets regression with k variables is a union of k-dimensional subspaces.

Because estimates from larger models conform more closely to the observed data, one naturally expects degrees of freedom to be monotone with respect to model inclusion. However, as we have already seen in Figure 2, monotonicity is far from guaranteed even in very simple examples, and can rise above p. Surprisingly, degrees of freedom need not be monotone even for methods projecting onto convex sets, including ridge regression and the lasso, although the

degrees of freedom cannot exceed the dimension of the convex set. The non-monotonicity of degrees of freedom for such convex methods was discovered independently by Kaufman & Rosset (2014), who give a thorough account. Among other results, they prove that the degrees of freedom of the projection onto a convex set must always be smaller than the dimension of that set. In contrast, we show that projection onto any closed non-convex set can have arbitrarily large degrees of freedom, regardless of the dimensions of $\mathcal{M}$ and y.

2. PRELIMINARIES

We consider fitting techniques with some tuning parameter, discrete or continuous, that can be used to vary a model from less to more constrained. In best-subsets regression, the tuning parameter k determines how many predictor variables are retained in the model. For a general fitting technique, we will use the notation $\hat{\mu}^{(k)}$ for the fitted response produced using tuning parameter k.

As mentioned in the introduction, a general formula for degrees of freedom can be motivated by the following relationship between expected prediction error, which we will represent by the symbol EPE, and residual sum of squares for ordinary least squares (Mallows, 1973):

$$\text{EPE} = E(\|r\|_2^2) + 2\sigma^2 p.$$

Analogously, once a fitting technique and tuning parameter k are chosen for fixed data y, the degrees of freedom functional, which we will denote by DF, is defined by:

$$E\left\{\sum_{i=1}^{n}(y_i^* - \hat{\mu}_i^{(k)})^2\right\} = E\left\{\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{(k)})^2\right\} + 2\sigma^2 \text{DF}(\mu, \sigma^2, k), \tag{1}$$

where σ^2 is the variance of the ε_i, assumed finite, and y_i^* is a new independent copy of y_i with mean μ_i. Thus degrees of freedom is defined as a measure of the optimism of residual sum of squares. This definition in turn leads to a simple closed form expression for degrees of freedom under very general conditions, as shown by the following theorem.

THEOREM 1 (EFRON (1986)). *For $i \in \{1, \ldots, n\}$, let $y_i = \mu_i + \varepsilon_i$, where the μ_i are non-random and the ε_i have mean zero and finite variance. Let $\hat{\mu}_i,\ i \in \{1, \ldots, n\}$ denote estimates of μ_i from some fitting technique based on a fixed realization of the y_i, and let $y_i^*,\ i \in \{1, \ldots, n\}$ be independent of and identically distributed as the y_i. Then*

$$E\left\{\sum_{i=1}^{n}(y_i^* - \hat{\mu}_i)^2\right\} - E\left\{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\right\} = 2\sum_{i=1}^{n}\text{cov}(y_i, \hat{\mu}_i)$$

Remark 1. For independent and identically distributed errors with finite variance σ^2, Theorem 1 implies that,

$$\text{DF}(\mu, \sigma^2, k) = \frac{1}{\sigma^2}\,\text{tr}\left\{\,\text{cov}(y, \hat{\mu}^{(k)})\right\} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{cov}\left(y_i, \hat{\mu}_i^{(k)}\right). \tag{2}$$

When using a linear fitting method with hat matrix H, Equation (2) reduces to $\text{DF}(\mu, \sigma^2, k) = \text{tr}(H)$.

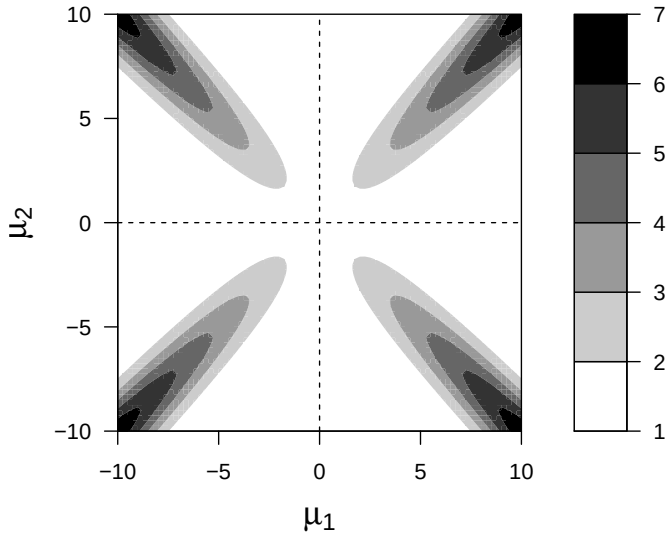

Fig. 2: Heatmap of the degrees of freedom for 1-best-subset regression fit to data from the model $y \sim N(\mu, I_2\sigma^2)$, as a function of the true mean vector $\mu \in \mathbb{R}^2$.

3. UNBOUNDED DEGREES OF FREEDOM FOR NON-CONVEX MODELS

Before stating our main result, we present a very simple example illustrating how large the degrees of freedom can be. Consider estimating a no-intercept linear regression model with design matrix $X = I_2$ and response $y \sim N(\mu, I_2)$, with $\mu \in \mathbb{R}^2$. Suppose further that, in order to obtain a more parsimonious model than the full bivariate regression, we instead estimate the best fitting of the two univariate models, in other words, best-subsets regression with model size $k = 1$.

Figure 2 shows the effective degrees of freedom for this model, plotted as a function of μ. As before, the degrees of freedom can exceed the ambient dimension of 2. However, the plot shows the degrees of freedom growing steadily as μ moves diagonally away from the origin, raising the question of how large it can get.

For $\mu = (a, a)^T$ and a large, y falls in the positive quadrant with high probability and the best univariate model chooses the larger of the two response variables. Figure 3(a) illustrates the fit for several realizations of y. For $i \in \{1, 2\}$, $\hat{\mu}_i^{(1)}$ is either 0 or approximately a depending on small changes in y. As a result, the variance of y is far smaller than that of $\hat{\mu}^{(1)}$. Since the correlation between y_i and $\hat{\mu}_i^{(1)}$ is around 0.5, $\sum_{i=1}^n \mathrm{cov}(y_i, \hat{\mu}_i^{(1)})$ is also much larger than the variance of the y_i, and the large degrees of freedom can be inferred from Equation (2).

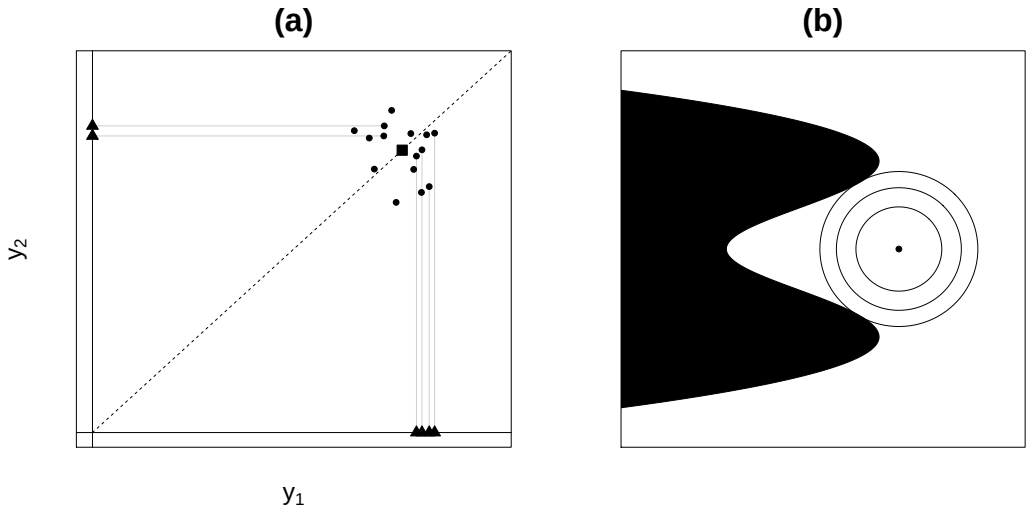

Fig. 3: (a) Sketch of the example described in Section 3. The square is μ and the solid black lines are the coordinate axes. Some realizations of y are shown as circles, along with a few of their best-subset projections $\hat{\mu}^{(1)}$, shown as triangles. The dashed line divides the points y with respect to which axis they are closer to. (b) Sketch of a regression problem with $n = 2$ and a non-convex constraint set. The filled area is the constraint set, the point the true mean vector μ, and the circles are the contours of the least squares objective function.

We see below that the degrees of freedom actually diverges as $a \to \infty$,

$$\begin{aligned}
\frac{1}{a}\mathrm{DF}\{(a,a)^T,1,1\} &= \frac{1}{2}E\left[\frac{1}{a}\sum_{i=1}^{2}\left\{(y_i^*-\hat{\mu}_i^{(1)})^2-(y_i-\hat{\mu}_i^{(1)})^2\right\}\right] \\
&= \frac{1}{2}E\left[\frac{1}{a}\sum_{i=1}^{2}\left\{(y_i^*-\hat{\mu}_i^{(1)})^2-(y_i-\hat{\mu}_i^{(1)})^2\right\}\mathbb{I}_{y\in Q_1}\right] \\
&\quad + \frac{1}{2}E\left[\frac{1}{a}\sum_{i=1}^{2}\left\{(y_i^*-\hat{\mu}_i^{(1)})^2-(y_i-\hat{\mu}_i^{(1)})^2\right\}\mathbb{I}_{y\notin Q_1}\right] \\
&= \frac{1}{2}E\left(\frac{1}{a}\Big[a^2+2a\varepsilon_1^*+\{\varepsilon_1^*\}^2+\left\{\varepsilon_2^*-\min(\varepsilon_1,\varepsilon_2)\right\}^2\right. \\
&\qquad \left.-a^2-2a\min\{\varepsilon_1,\varepsilon_2\}-\left\{\min(\varepsilon_1,\varepsilon_2)\right\}^2\Big]\mathbb{I}_{y\in Q_1}\right)+o(1) \\
&\to \frac{1}{2}E\left\{2\varepsilon_1^*-2\min(\varepsilon_1,\varepsilon_2)\right\} = E\left\{\max(\varepsilon_1,\varepsilon_2)\right\}, a\to\infty,
\end{aligned}$$

where Q_1 is the first quadrant of $\mathbb{R}^2$, $\mathbb{I}_S$ is the indicator function on the set S, and $\varepsilon_1^*, \varepsilon_2^*$ are noise realizations independent of one another and of y. The $o(1)$ term comes from the the fact that $a^{-1}\sum_{i=1}^{2}\{(y_i^*-\hat{\mu}_i^{(1)})^2-(y_i-\hat{\mu}_i^{(1)})^2\}$ is $O_p(a)$ while $\mathrm{pr}(y\notin Q_1)$ shrinks exponentially fast in a, as it is a Gaussian tail probability. The convergence in the last line follows by the dominated convergence theorem applied to the first term in the preceding line. For large a,

$E\{\max(\varepsilon_1, \varepsilon_2)\} \approx 0.56$, giving $\text{DF}\{(a,a)^T, 1, 1\} \approx 0.56a$. Equivalently, the degrees of freedom would also diverge if a were held fixed and $\sigma^2 \to 0$.

The phenomenon we have just illustrated is not an idiosyncratic pathology of best-subsets regression, but in fact can occur whenever we project onto a non-convex model.

THEOREM 2. *For a fitting technique that minimizes squared error subject to a non-convex, closed constraint $\hat{\mu}^{(k)} \in \mathcal{M}_k \subset \mathbb{R}^n$, consider the model,*

$$y = \mu + \sigma\varepsilon, \quad \varepsilon_i \sim F \quad (i = 1, \ldots, n),$$

where the ε_i are independent and F is a mean-zero distribution with finite variance supported on an open neighborhood of 0. Then there exists some μ^* such that $\text{DF}(\mu^*, \sigma^2, k) \to \infty$ as $\sigma^2 \to 0$.

Proof of Theorem 2. An intuitive proof sketch follows below, while a rigorous proof is deferred to the Supplementary Material. □

Best-subsets regression has a non-convex constraint set for $k < p$, and our toy example gave some insight for why the degrees of freedom can be much greater than the ambient dimension. We now give intuition for how the theorem generalizes to any non-convex constraint set.

Place the true mean at a point with non-unique projection onto the constraint set; see Figure 3(b). The salient feature of the figure is that the spherical contour spans the gap of the divot where it meets the constraint set. A point with non-unique projection must exist by the Motzkin–Bunt Theorem (Motzkin, 1935; Kritikos, 1938). The constraint set for $\hat{\mu} = X\hat{\beta}$ is just an affine transformation of the constraint set for $\hat{\beta}$, and thus a non-convex $\hat{\beta}$-constraint is equivalent to a non-convex $\hat{\mu}$ constraint. Then the fit depends sensitively on the noise process, even when the noise is very small, since y is projected onto multiple well-separated sections of the constraint set. Thus as the magnitude of the noise, σ, goes to zero, the variance of $\hat{\mu}$ remains roughly constant. Equation (2) then tells us that degrees of freedom can be made arbitrarily large, as it will be roughly proportional to σ^{-1}. By inserting that rate into Equation (1), we also see that, for $\sigma^2 \to 0$, the difference between the expected prediction error and the residual sum of squares converges to zero, with both just approaching the expected squared bias of the model.

Theorem 2 and its proof have practical ramifications for the use of degrees of freedom in model selection criteria. Akaike information criterion and Bayes information criterion use degrees of freedom to mitigate overfitting of a model. Since theoretical results for these criteria, such as conditions for model consistency, rely solely on the definition of degrees of freedom which we have also used, those results still hold in the settings we consider here, but only so long as the correct degrees of freedom is used. Our result sheds light on the dangers of using a naïve estimate of degrees of freedom based on model size. For instance, using k as the degrees of freedom for best-subsets regression with k can be arbitrarily far from the truth, resulting in values of model selection criteria that are also arbitrarily wrong.

Our proof also clarifies that the degrees of freedom are most likely to be large when the true mean is nearly equidistant from two or more well-separated parts of the model constraint set. For methods like best-subsets or forward selection that explicitly seek a parsimonious model fit, this could occur if there are several different parsimonious models that describe μ almost equally well. Thus, if we use a parsimony-seeking model because we know the truth to be parsimonious, we would not expect the degrees of freedom to misbehave very often. By contrast, if we force the fit to be parsimonious even when the truth is not, the degrees of freedom may well misbehave, as it does in our simulation examples.

4. DISCUSSION

The common intuition that effective degrees of freedom serves as a consistent and interpretable measure of model complexity merits some skepticism. Our results, combined with those of Kaufman & Rosset (2014), demonstrate that for many widely-used convex and non-convex fitting techniques, the degrees of freedom can be non-monotone with respect to model nesting. Furthermore, in the non-convex case, the degrees of freedom can exceed the dimension of the model space by an arbitrarily large amount, and may do so in run-of-the-mill datasets.

In light of the above, the term degrees of freedom seems misleading, as it is suggestive of a quantity corresponding to model size or complexity. It is also misleading to consider degrees of freedom as a measure of overfitting, or how flexibly the model conforms to the data, since a model is always at least as flexible as a submodel. By definition, the effective degrees of freedom of Efron (1983) measures optimism of in-sample error as an estimate of out-of-sample error, but we should not be too quick to carry over our intuition from linear models.

ACKNOWLEDGEMENT

We thank the editor, associate editor, and two reviewers for helpful comments that signficantly improved the paper. Lucas Janson was partially supported by the National Institutes of Health. William Fithian was partially supported by the National Science Foundation and the Gerald J. Lieberman Fellowship. Trevor Hastie was partially supported by the National Institutes of Health and the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proof of Theorem 2 and further explanation, including code, for the examples in Sections 1·1 and 3.

REFERENCES

BUJA, A., HASTIE, T. J. & TIBSHIRANI, R. J. (1989). Linear smoothers and additive models. *The Annals of Statistics* **17**, 453–510.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

EFRON, B. (1986). How Biased Is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association* **81**, 461–470.

HOERL, A. E. (1962). Application of Ridge Analysis to Regression Problems. *Chemical Engineering Progress* **58**, 54–59.

KAUFMAN, S. & ROSSET, S. (2014). When Does More Regularization Imply Fewer Degrees of Freedom? Sufficient Conditions and Counter Examples from Lasso and Ridge Regression. *Biometrika (to appear)* .

KRITIKOS, M. N. (1938). Sur quelques propriétés des ensembles convexes. *Bulletin Mathématique de la Société Roumaine des Sciences* **40**, 87–92.

MALLOWS, C. L. (1973). Some Comments on C_p. *Technometrics* **15**, 661–675.

MOTZKIN, T. (1935). Sur quelques propriétés caractéristiques des ensembles convexes. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur* **21**, 562–567.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**, 267–288.

[*Received April* 2012. *Revised September* 2012]