Biometrika (2012), **99**, 1, *pp*. 1–4

Printed in Great Britain

Advance Access publication on 31 July 2012

Supplementary material for 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

1. PROOF OF THEOREM 2

Proof. The proof relies heavily on the fact that for every non-convex set $\mathcal{M}$ in Euclidean space there is at least one point whose projection onto $\mathcal{M}$ is not unique. This fact was proved independently in the 1934 Rijks-Universiteit PhD thesis by L. N. H. Bunt, Motzkin (1935), and Kritikos (1938). A schematic for this proof in two dimensions is provided in Figure 1. Let μ be

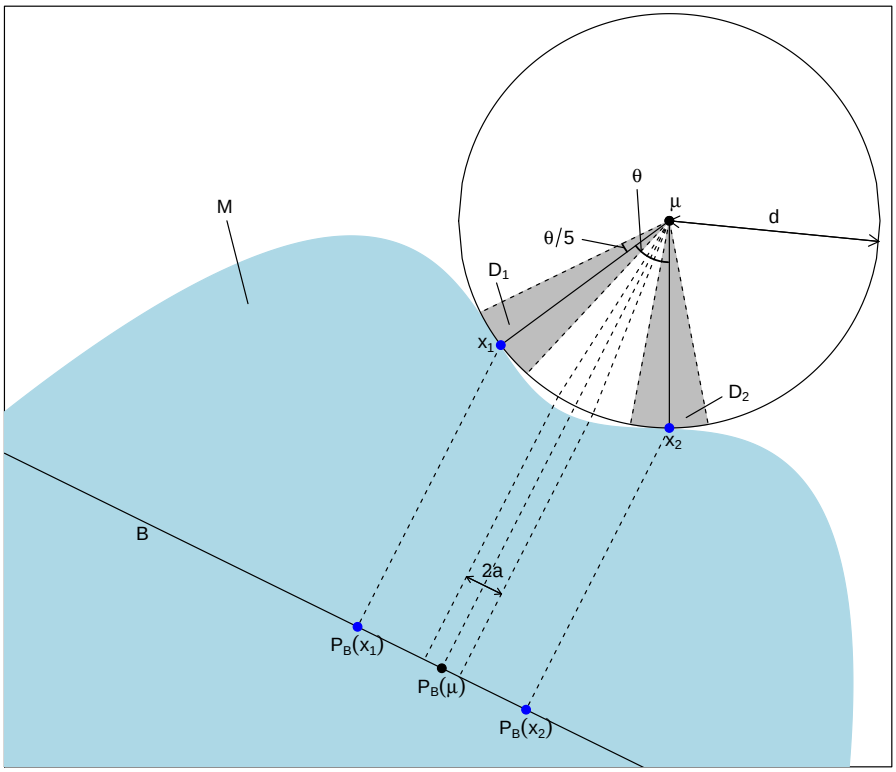

Fig. 1. Schematic for the proof of Theorem 2, in two dimensions.

a point with non-unique projection onto the non-convex set $\mathcal{M}$ and let x_1 and x_2 be two distinct projections of μ onto $\mathcal{M}$. Let $d = ||\mu - x_1||_2 = ||\mu - x_2||_2$ be the Euclidean distance between

μ and $\mathcal{M}$, and

$$\theta = \cos^{-1}\left\{\frac{(x_1-\mu)(x_2-\mu)}{|x_1-\mu||x_2-\mu|}\right\}$$

be the angle between x_1 and x_2, taken as vectors from μ. Define the set

$$\mathcal{D}_1 = \left\{v \in \mathbb{R}^n : \cos^{-1}\left\{\frac{(x_1-\mu)(v-\mu)}{|x_1-\mu||v-\mu|}\right\} < \frac{\theta}{5},\ ||v-\mu||_2 < d\right\},$$

and $\mathcal{D}_2$ analogously for x_2. Let $\mathcal{B}$ be a one-dimensional affine subspace that is both parallel to the line connecting x_1 and x_2, and contained in the hyperplane defined by μ, x_1, and x_2. Denoting the projection operator onto $\mathcal{B}$ by $P_{\mathcal{B}}$, let $z = ||P_{\mathcal{B}}y - P_{\mathcal{B}}\mu||_2/\sigma$, and $\tilde{y} = ||P_{\mathcal{B}}\hat{\mu}^{(k)} - P_{\mathcal{B}}\mu||_2$. Let $a = d\cos\{(\pi-\theta/5)/2\}$. We now have,

$$\begin{aligned}
\operatorname{tr}\{\operatorname{cov}(y,\hat{\mu}^{(k)})\} &\geq \operatorname{cov}(P_{\mathcal{B}}y, P_{\mathcal{B}}\hat{\mu}^{(k)}) \\
&= E(\sigma z\tilde{y}) \\
&= E(\sigma z\tilde{y}1_{y\in\mathcal{D}_1\cup\mathcal{D}_2}) + E(\sigma z\tilde{y}1_{y\notin\mathcal{D}_1\cup\mathcal{D}_2}) \\
&\geq \sigma E(z\tilde{y}1_{y\in\mathcal{D}_1}) + \sigma E(z\tilde{y}1_{y\in\mathcal{D}_2}) \\
&\geq a\sigma\big\{E(z1_{y\in\mathcal{D}_2}) - E(z1_{y\in\mathcal{D}_1})\big\}, \\
&= 2a\sigma E(z1_{y\in\mathcal{D}_2}).
\end{aligned}$$

The first inequality follows from the translation and rotation invariance of the trace of a covariance matrix, and from the positivity of the diagonal entries of the covariance matrix for the case of projection fitting methods. For the second inequality, $E(\sigma z\tilde{y}1_{y\notin\mathcal{D}_1\cup\mathcal{D}_2}) \geq 0$, again because of the positivity of the degrees of freedom of projection methods, applied to the same model with a noise process that has support on $\mathcal{D}_1$ and $\mathcal{D}_2$ removed. The third inequality follows from considering the projections of $\mathcal{D}_1$ and $\mathcal{D}_2$ onto $\mathcal{M}$ and then onto $\mathcal{B}$, and noting that the two double projections must be separated by at least a distance of $2a$.

Defining

$$\mathcal{F}_1 = \left\{v \in \mathbb{R}^n : \cos^{-1}\left\{\frac{(x_1-\mu)(v-\mu)}{|x_1-\mu||v-\mu|}\right\} < \frac{\theta}{5}\right\}$$

and $\mathcal{F}_2$ analogously for x_2, $\operatorname{pr}(y \in \mathcal{F}_1 \setminus \mathcal{D}_1) = \operatorname{pr}(y \in \mathcal{F}_2 \setminus \mathcal{D}_2) \to 0$ as $\sigma^2 \to 0$. Thus,

$$\operatorname{tr}\{\operatorname{cov}(y,\hat{\mu}^{(k)})\} \geq 2a\sigma\left\{E(z1_{y\in\mathcal{F}_2}) + o(\sigma)\right\}.$$

Neither z nor the event $y \in \mathcal{F}_2$ depend on σ, so define the constant $b = 2aE(z1_{y\in\mathcal{F}_2}) > 0$ which is independent of σ. Thus we have shown that,

$$\begin{aligned}
\textsc{df}(\mu^*,\sigma^2,k) &= \frac{1}{\sigma^2}\operatorname{tr}\{\operatorname{cov}(y,\hat{\mu}^{(k)})\} \\
&\geq \frac{b+o(\sigma)}{\sigma} \\
&\to \infty
\end{aligned}$$

as $\sigma^2 \to 0$. □

2. DETAILED EXPLANATION OF EXAMPLES

In Sections 3 and 1·1, degrees of freedom is estimated by computing an unbiased estimator of degrees of freedom for each simulated noise realization. This unbiased estimator for degrees

of freedom can be obtained from Equation (2) by exploiting the linearity of the expectation and trace operators,

$$\begin{aligned} \mathrm{DF}(\mu, \sigma^2, k) &= \frac{1}{\sigma^2} \operatorname{tr}\{\operatorname{cov}(y, \hat{\mu}^{(k)})\} \\ &= \frac{1}{\sigma^2} E\left[\{y - \mu\}^T \{\hat{\mu}^{(k)} - E(\hat{\mu}^{(k)})\}\right] \\ &= \frac{1}{\sigma^2} E\left(\varepsilon^T \hat{\mu}^{(k)}\right), \end{aligned} \tag{1}$$

where the last inequality follows because $E(\varepsilon) = 0$. Note that for the full ordinary least squares regression, we have $\mathrm{DF}(\mu, \sigma^2, p) = p$, so that it is equally true that

$$\mathrm{DF}(\mu, \sigma^2, k) = \frac{1}{\sigma^2} E\left(\varepsilon^T \hat{\mu}^{(k)} - \varepsilon^T \hat{\mu}^{(p)} + p\sigma^2\right)$$

Writing $\delta_k = \sigma^{-2} \varepsilon^T \left(\hat{\mu}^{(k)} - \hat{\mu}^{(p)}\right) + p$, we can estimate the true degrees of freedom by averaging δ_k over many simulations. Since this estimate of degrees of freedom is an average of independent and identically distributed random variables, its standard deviation can be estimated by the emipirical standard deviation of the δ_k divided by the square root of the number of simulations.

The following is the code for the best subsets regression simulation in Section 1·1.

```
set.seed(1)
library(leaps)
library(gplots)
n = 50; p = 15; By = 20000
dfmat = matrix(0, By, p)

x = matrix(rnorm(n * p), n, p)
beta <- rnorm(p)*2
mu = x %*% beta
for(j in 1:By){
    if(j %% 100 == 0) cat(j,"\n")
    y = rnorm(n) + mu
    temp = regsubsets(x, y, nbest = 1, nvmax = p, intercept = FALSE)
    for(i in 1:p){
        jcoef = coef(temp, id = i)
        xnames = names(jcoef)
        which = match(xnames, letters[1:p])
        if(i == 1){
            yhat = matrix(x[ , which], n, 1) %*% jcoef
        } else{
            yhat = x[ , which] %*% jcoef
        }
        dfmat[j, i] = sum((y - mu) * yhat)
    }
    dfmat[j, ] <- p + dfmat[j,] - dfmat[j,p]
}
df = apply(dfmat, 2, mean)
error = sqrt(apply(dfmat, 2, var) / By)
```

The code for the forward selection simulation in Section 1·1 is almost identical.

```
set.seed(1)
library(leaps)
library(gplots)
n = 50; p = 15; By = 20000
dfmat = matrix(0, By, p)

x = matrix(rnorm(n * p), n, p)
beta <- rnorm(p)*2
mu = x %*% beta
for(j in 1:By){
    if(j %% 100 == 0) cat(j,"\n")
    y = rnorm(n) + mu
    temp = regsubsets(x, y, nbest = 1, nvmax = p, method="forward", intercept = FALSE)
    for(i in 1:p){
        jcoef = coef(temp, id = i)
        xnames = names(jcoef)
```

```
            which = match(xnames, letters[1:p])
            if(i == 1){
                yhat = matrix(x[ , which], n, 1) %*% jcoef
            } else{
                yhat = x[ , which] %*% jcoef
            }
            dfmat[j, i] = sum((y - mu) * yhat)
        }
        dfmat[j, ] <- p + dfmat[j,] - dfmat[j,p]
    }
    df = apply(dfmat, 2, mean)
    error = sqrt(apply(dfmat, 2, var) / By)
```

REFERENCES

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

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

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