COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 0, NO. 0, 1–15
https://doi.org/10.1080/03610918.2017.1315728

Taylor & Francis
Taylor & Francis Group

Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation

Severin Hauenstein [a], Simon N. Wood [b], and Carsten F. Dormann [a]

[a]Department of Biometry and Environmental System Analysis, University of Freiburg, Freiburg, Germany;
[b]School of Mathematics, University of Bristol, Bristol, United Kingdom

ABSTRACT

Generalized degrees of freedom (GDF), as defined by Ye (1998 JASA 93:120–131), represent the sensitivity of model fits to perturbations of the data. Such GDF can be computed for any statistical model, making it possible, in principle, to derive the effective number of parameters in machine-learning approaches and thus compute information-theoretical measures of fit. We compare GDF with cross-validation and find that the latter provides a less computer-intensive and more robust alternative. For Bernoulli-distributed data, GDF estimates were unstable and inconsistently sensitive to the number of data points perturbed simultaneously. Cross-validation, in contrast, performs well also for binary data, and for very different machine-learning approaches.

ARTICLE HISTORY
Received 14 April 2016
Accepted 31 March 2017

KEYWORDS
Boosted regression trees; Data perturbation; Model complexity; Random forest

MATHEMATICS SUBJECT CLASSIFICATION
Primary 62; Secondary 07

1. Introduction

In many scientific fields, statistical models have become a frequently used tool to approach research questions. Choosing the most appropriate model(s), that is, the model(s) best supported by the data, however, can be difficult. Especially in ecological, sociological, and psychological research, where data are often sparse while systems are complex, the evidence for one particular statistical model may not be conclusive. The existence of alternative models, which fit the data with comparable goodness, but yield considerably different predictions is ubiquitous (e.g., Draper, 1995; Madigan et al., 1995; Raftery, 1996).

In ecological and evolutionary research, statistical procedures are dominated by an "information-theoretical approach" (Burnham and Anderson, 1998, 2002), which essentially means that model fit is assessed by Akaike's Information Criterion (defined as AIC $= -2$ log-likelihood $+ 2$ number of parameters: Akaike, 1973). In this field, the AIC has become the paradigmatic standard for model selection of likelihood-based models as well as for determination of model weights in model averaging (e.g., Diniz-Filho et al., 2008; Hegyi and Garamszegi, 2011; Mundry, 2011). In recent years, and particularly in the field of species distribution analyses, non-parametric, likelihood-free approaches ("machine learning") have become more prevalent and in comparisons typically show better predictive ability (e.g., Elith et al., 2006; Elith and Leathwick, 2009; Olden et al., 2008; Recknagel, 2001). However, these approaches do not allow the computation of an AIC, because many of such "black-box" methods are neither likelihood-based, nor can one readily account for model

CONTACT Severin Hauenstein ✉ severin.hauenstein@biom.uni-freiburg.de Department of Biometry and Environmental System Analysis, University of Freiburg, Tennenbacherstr. 4, Freiburg, 79106 Germany.

complexity, as the number of parameters does not reflect the effective degrees of freedom (a phenomenon Elder (2003) calls the "paradox of ensembles"). Thus, at the moment ecologists are faced with the dilemma of either following the AIC-paradigm, which essentially limits their toolbox to generalized linear models (GLMs) and generalized additive models (GAMs), or use machine-learning tools and closing the AIC-door. From a statistical point of view, dropping an AIC-based approach to model selection and averaging is no loss, as an alternative approximation of the Kullback–Leibler distance is possible through cross-validation.

In this study, we explore a potential avenue to unite AIC and machine learning, based on the concept of generalized degrees of freedom (GDF). We explore the computation of GDF for Gaussian and Bernoulli-distributed data as plug-in estimates of the number of parameters in a model. We also compare such a GDF-AIC-based approach with cross-validation.

The article is organized as follows. We first review the generalized degrees of freedom concept and relate it to the degrees of freedom in linear models. Next, we briefly illustrate the relation between cross-validation and Kullback–Leibler divergence, as KL also underlies the derivation of the AIC. Through simulation and real data we then explore the computation of GDF for a variety of modeling algorithms and its stability. The article closes with a comparison of GDF-based AIC and cross-validation-derived deviance, and comments on computational efficiency of the different approaches and consequences for model averaging.

1.1. Generalized degrees of freedom

Generalized degrees of freedom (GDF), originally proposed by Ye (1998) and illustrated for machine learning by Elder (2003), can be used as a measure of model complexity through which we can make information theoretical approaches applicable to black-box algorithms.

In order to understand the conceptual properties of this method, we can make use of a somewhat simpler version of degrees of freedom (df). For a linear model m, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, df are computed as the trace of the hat (or projection) matrix $\mathbf{H}$, with elements h_{ij} (e.g., Hastie et al., 2009, p. 153):

$$\mathrm{df}_m = \mathrm{trace}(\mathbf{H}) = \mathrm{trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T). \tag{1}$$

For a linear model, this is the number of its independent parameters, that is, the rank of the model. To expand this concept to other, non-parametric methods, making it in fact independent of the actual fitted model, the definition above provides the basis for a *generalized* version of degrees of freedom. Thus, and according to Ye (1998),

$$\mathrm{GDF}_m = \mathrm{trace}\,(\mathbf{H}) = \sum_i h_{ii} = \sum_i \frac{\partial \hat{y}_i}{\partial y_i}, \tag{2}$$

where $\hat{y}_i$ is the fitted value. That is to say, a model is considered the more complex the more adaptively it fits the data, which is reflected in higher GDF. Or in the words of Ye (1998, p. 120): GDF quantify the "sensitivity of each fitted value to perturbations in the corresponding observed value." For additive-error models (i.e., $Y = f(X) + \varepsilon$, with $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$), we can use the specific definition that

$$\mathrm{GDF}_m = \frac{\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)}{\sigma_\varepsilon^2}, \tag{3}$$

where N denotes the number of observations (Hastie et al., 2009, p. 233). Beyond the additive-error model more generally, however, we have to approximate Eq. (2) in other ways. By fitting the model with perturbed y_i and assessing the response in $\hat{y}_i$, we can evaluate $\frac{\partial \hat{y}_i}{\partial y_i}$, according to Elder (2003), so that

$$\mathrm{GDF} \approx \sum_i \frac{\widehat{y'}_i - \hat{y}_i}{y'_i - y_i}, \tag{4}$$

where y'_i is perturbed in such a way that, for normally distributed data, $y'_i = y_i + \mathcal{N}(\mu = 0, \sigma^2)$ with σ^2 being small relatively to the variance of y.

In order to adapt GDF to binary data, we need to reconsider this procedure. Since a perturbation cannot be achieved by adding small random noise to y_i, the only possibility is to replace 0's by 1's and vice versa. A perturb-all-data-at-once approach (Elder, 2003) as for Gaussian data is thus not feasible. We explore ways to perturb Bernoulli-distributed y below.

The equivalence of GDF and the number of parameters in the linear model encourages us to use GDF as a plug-in estimator of the number of parameters in the AIC computation.

1.2. Cross-validation and a measure of model complexity

Cross-validation is a standard approach to quantify predictive performance of a model, automatically accounting for model complexity (by yielding lower fits for too simple and too complex models, e.g., Hastie et al., 2009; Wood, 2006). Because each model is fitted repeatedly, cross-validation is computationally more expensive than the AIC, but the same problem arises for GDF: it requires many simulations to estimate it stably (see below).

We decided to use the log-likelihood as measure of fit for the cross-validation (Horne and Garton, 2006), with the following reasoning. Let f_θ denote the model density of our data, y, and let f_t denote the true density. Then the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) is

$$D_{\mathrm{KL}} = \int \left(\log f_t - \log f_{\hat{\theta}}\right) f_t \, dy. \tag{5}$$

AIC is an estimate of this value (in the process taking an expectation over the distribution of $\hat{\theta}$). Now $\int f_t \log f_t \, dy$ is independent of the model, and the same for all models under consideration (and hence gets dropped from the AIC). So the important part of the KL-divergence is

$$D_{\mathrm{KL}} \propto - \int f_t \log f_{\hat{\theta}} \, dy. \tag{6}$$

This model log-likelihood can be estimated by cross-validation, where the expectation is over the distribution of data not used in the estimation of $\hat{\theta}$, such that

$$\ell_{\mathrm{CV}} = - \sum_{i=1}^{K} \log f_{\hat{\theta}[-i]}(y_i), \tag{7}$$

where ℓ_{CV} is the sum over K folds of the log-likelihood of the test subset $y^{[i]}$, given the trained model $f_{\hat{\theta}[-i]}$, with parameter estimates $\theta^{[-i]}$ (Wood, 2015). To put this on the same scale as AIC, we multiply by -2 to obtain the cross-validated deviance.

With the assumption of AIC and (leave-one-out) cross-validation being asymptotically equivalent (Stone, 1977) and given the definition of AIC, we argue that it should be possible to extract a measurement from ℓ_{CV} that quantifies model complexity, specifically the estimated effective number of parameters, $\hat{p}$. Hence,

$$\begin{aligned} \text{AIC} &= -2\ell_m + 2\hat{p} \approx -2\ell_{\text{CV}} \\ \hat{p} &\approx \ell_m - \ell_{\text{CV}}. \end{aligned} \tag{8}$$

Embracing the small sample-size bias adjustment of AIC (Hurvich and Tsai, 1989; Sugiura, 1978), we get

$$\begin{aligned} \text{AICc} &= -2\ell_m + 2\hat{p} + \frac{2\hat{p}(\hat{p}+1)}{N-\hat{p}-1} \approx -2\ell_{\text{CV}} \\ \hat{p} &\approx \frac{(\ell_m - \ell_{\text{CV}})\,(N-1)}{\ell_m - \ell_{\text{CV}} + N} \end{aligned} \tag{9}$$

with $\hat{p}$ representing (estimated) model complexity, ℓ_m the maximum log-likelihood of the original (non-cross-validated) model, and N the number of data points.

Thus, we can compute both a cross-validation-based deviance that should be equivalent to the AIC, $-2\ell_{\text{CV}}$, as well as a cross-validation alternative to GDF, based on the degree of overfitting of the original model (Eqs. (8) and (9)).

This approach to computing model complexity is not completely unlike that of the DIC (Spiegelhalter et al., 2002), where p_D represents the effective number of parameters in the model and is computed as the mean of the marginal deviance in an MCMC-analysis minus the deviance of the mean estimates: $p_D = \overline{D(\theta)} - D(\bar{\theta}) = 2\ell(\bar{\theta}) - 2\overline{\ell(\theta)}$ (Wood, 2015). In Eq. (8), the likelihood estimate plays the role of $\ell(\bar{\theta})$, while the cross-validation estimates $\overline{\ell(\theta)}$.

2. Implementing and evaluating the GDF-approach for normally and Bernoulli-distributed data

We analyzed simulated and real datasets using five different statistical models. Then we computed their GDF, disturbing k data points at a time, with different intensity (only for normal data), and for different values of k.

2.1. Data: Simulated and real

First, we evaluated the GDF-approach on normally distributed data following Elder (2003), deliberately using a relatively small dataset: $N_{\text{norm}} = 250$. The response y was simulated as $y \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 x_4, 1)$, with $\beta_0 = -5$, $\beta_1 = 5$, $\beta_2 = -10$, $\beta_3 = 10$, $\beta_4 = 10$, and $x_i \sim \mathcal{U}(0, 1)$, $i = 1, 2, 3, 4$. (This simulation was repeated to control for possible influences of the data itself on the resulting GDF, but results were near-identical.)

We simulated binary data with $N_{\text{binom}} = 300$ (effective sample size $\text{ESS}_{\text{binom}} \approx 150$) and $y \sim \text{Bern}(\text{logit}^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 x4))$, with $\beta_0 = -6.66$, $\beta_1 = 5$, $\beta_2 = -10$, $\beta_3 = 10$, $\beta_4 = 10$, and $x_i \sim \mathcal{U}(0, 1)$, $i = 1, 2, 3, 4$.

Our real-world data comprised a fairly small dataset ($N_{\text{Physeter}} = 261$, $\text{ESS}_{\text{Physeter}} = 115$), showing the occurrence of sperm whales (*Physeter macrocephalus*) around Antarctica (collected during cetacean IDCR-DESS SOWER surveys), and a larger global occurrence dataset ($N_{\text{Vulpes}} = 12722$, $\text{ESS}_{\text{Vulpes}} = 5401$) for the red fox (*Vulpes vulpes*, provided by the International Union for Conservation of Nature (IUCN)). We pre-selected predictors as described in

Dormann and Kaschner (2010) and Dormann et al. (2010), yielding the six and three (respectively) most important covariates. Since we are not interested in developing the best model, or comparing model performance, we did not strive to optimize the model settings for each dataset.

2.2. Implementing GDF for normally distributed data

A robust computation of GDF is a little more problematic than Eq. (4) suggests. Ye (1998) proposed to fit a linear regression to repeated perturbations of each data point, that is, to $(\widehat{y}_i' - \widehat{y})$ over $(y_i' - y_i)$ for each repeatedly perturbed data point i, calculating GDF as the sum of the slopes across all data points. As y_i is constant for all models, and $\widehat{y}_i$ is constant for the non-stochastic algorithms (GLM, GAM), the linear regression simplifies to $\widehat{y}_i'$ over y_i'. For internally stochastic models, we compute a mean $\overline{\widehat{y}_i'}$ as plugin for $\widehat{y}_i$, and therefore apply the procedure also to randomForests, Artificial Neural Networks (ANNs), and boosted regression trees (BRTs).

Elder (2003) presents this so-called "horizontal" method (the perturbed y' and fitted $\widehat{y}'$ are stored as different columns in two separate matrices, which are then regressed row-wise) as more robust compared to the "vertical" method (where for each perturbation a GDF is computed for each column in these matrices and later averaged across replicate perturbations). Convergence is poor when using the "vertical" method, so we restricted the computation of GDF to the "horizontal" method.

2.3. GDF for binary data

While normal data can be perturbed by adding normal noise (see the next section), binary data cannot. The only option is to invert their value, that is, $0 \rightarrow 1$ or $1 \rightarrow 0$. Clearly, we cannot perturb many data points simultaneously this way, as that will dilute any actual signal in the data. However, for large datasets it is computationally expensive to perturb all N data points individually, and repeatedly (to yield values for the horizontal method, see above). We thus varied the number of data points to invert, k, to evaluate whether we can raise k without biasing the GDF estimate.

2.4. How many data points to perturb simultaneously?

With the number of points to perturb, k, the number of replicates for each GDF-computation (n_{gdf}) and the amplitude of disturbance (for the normal data), we have three tuning parameters when computing GDF.

First, we calculated GDF for the simulated data and the sperm whale dataset with k taking increasing values (from 1 to near N, or to ESS for binary data). Thus, random subsets of size k of the response variable y were perturbed, yielding y'. After each particular perturbation, the model was re-fitted to y'. To gain insight about the variance of the computed GDF values, we repeated this calculation 100 times (due to very high computational effort, the number of re-runs had to be limited to 10 for both randomForest and BRT).

The perturbation for the normally distributed y were also drawn from a normal distribution $\mathcal{N}(0, 0.25 \cdot \sigma_{\text{simulation}})$. We evaluated the sensitivity to this parameter by setting it to 0.125 and 0.5.

2.5. Modeling approaches

We analyzed the data using generalized linear model (GLM), generalized additive model (GAM), randomForest, feed-forward Artificial Neural Network (ANN), and boosted regression trees (BRT). For the GLM, GDF should be identical to the trace of the Hessian (and the rank of the model), hence GLM serves as benchmark. For GAMs, different ways to compute the degrees of freedom of the model have been proposed (Wood, 2006), while for the other three methods the intrinsic stochasticity of the algorithm (or in the case of ANN of the initial weights) may yield models with different GDF each time.

All models were fit using R (R Core Team, 2014) and packages gbm (for BRT: Ridgeway et al., 2013, interaction depth = 3, shrinkage = 0.001, number of trees = 3,000, cv.folds = 5), mgcv (for GAM: Wood, 2006, thin plate regression splines), nnet (for ANN: Venables and Ripley, 2002, size = 7, decay = 0.03 for simulated and decay = 0.1 for real data, linout = TRUE for normal data), randomForest (Liaw and Wiener, 2002), and core package stats (for GLM, using also quadratic terms and first-order interactions). R-code for all simulation as well as data are available at https://github.com/biometry/GDF.

2.6. Computation of AIC and AIC-weights, from GDF and cross-validation

In addition to the number of parameters, the AIC-formula requires the likelihood of the data, given the model. As machine-learning algorithms may minimize a score function different from the likelihood, the result probably differs from a maximum likelihood estimate. To calculate the AIC, however, we have to assume that the distance minimized by non-likelihood methods is proportional to the likelihood, otherwise the AIC would not be a valid approximation of KL-divergence. No such assumption has to be made for cross-validation, as ℓ_{CV} here only serves as a measure of model performance. For the normal data, we compute the standard deviation of the model's residuals as plug-in estimate of σ.

For the binary data, we use the model fits as probability in the Bernoulli distribution to compute the likelihood of that model. We then calculated the AIC for all considered models based on their GDF value. Due to the small sample sizes, we used AICc (Hurvich and Tsai, 1989; Sugiura, 1978):

$$\mathrm{AICc} = -2\ell_m + 2\mathrm{GDF} + \frac{\mathrm{GDF}(\mathrm{GDF}+1)}{N - \mathrm{GDF} - 1}. \tag{10}$$

We used (10-fold) cross-validation, maintaining prevalence in the case of binary data, to compute the log-likelihood of the cross-validation, yielding ℓ_{CV}. To directly compare it to AICc, we multiplied ℓ_{CV} with -2. The cross-validation automatically penalizes for overfitting by making poorer predictions.

For model averaging, we computed model weights w_m for each model m, once for the GDF-based AICc and for the cross-validation log-likelihood, using the equation for Akaike-weights (Burnham and Anderson, 2002, p. 75; Turkheimer et al., 2003):

$$w_m^{\mathrm{AICc}} = \frac{e^{-\frac{1}{2}\Delta_m^{\mathrm{AICc}}}}{\sum_{r=1}^{M} e^{-\frac{1}{2}\Delta_r^{\mathrm{AICc}}}}, \tag{11}$$

where $\Delta_m^{\mathrm{AICc}} = \mathrm{AICc}_m - \mathrm{AICc}_{\min}$, taking the smallest AICc, that is, the AICc of the best of the candidate models as $\mathrm{AICc}_{\min}$; M is the number of models to be averaged over.

The same idea can be applied to cross-validated log-likelihoods, so that

$$w_m^{\mathrm{CV}} = \frac{e^{\Delta_m^{\mathrm{CV}}}}{\sum_{r=1}^{M} e^{\Delta_r^{\mathrm{CV}}}}, \tag{12}$$

where $\Delta_m^{\mathrm{CV}} = \ell_{\mathrm{CV}_{\max}} - \ell_{\mathrm{CV}_m}$, with $\ell_{\mathrm{CV}_{\max}}$ being the largest cross-validated log-likelihood in the model set.

3. Results

3.1. GDF configuration analysis

For normally distributed data, increasing the number of points perturbed simultaneously typically slightly increased the variance of the generalized degrees of freedom calculated for the model (Fig. 1, left column, GLM, GAM, but not for randomForest and ANN). For GLM and GAM, GDF computations yielded exactly the same value as the model's rank (indicated by the dashed horizontal line).

For simulated Bernoulli data (Fig. 1, central column), we also observed an effect of the number of points perturbed on the actual GDF value, which decreased for GLM, GAM, and ANN, but increased for BRTs with the number of data points perturbed. Several data points needed to be perturbed (> 20) to yield an approximately correct estimate. More worryingly, GDF depended nonlinearly on the number of data points perturbed, with values varying by a factor of 2 for GAM and ANN. For GAM, the sensitivity occurs because the smoothing parameter selection is sensitive to the quite severe information loss as more and more data are perturbed. GLM and BRT yielded more consistent but still systematically varying GDF estimates.

The same pattern was observed for the sperm whale data (Fig. 1, right column). For the GLM, there was still some bias observable, also in the sperm whale-case study's GLM. We attribute it to the fact that by perturbing the binary data we also alter the prevalence.

The two replicate simulations yielded consistent estimates, except in the case of the normal GAM and normal BRT. This suggests that both methods "fitted into the noise" specific to the dataset, while the other methods did not. We did not observe this phenomenon with the binary data.

For randomForest, GDF estimates center around 0, meaning that perturbations of the data did not affect the model predictions. This is to some extent explicable by the stochastic nature of randomForest, giving different predictions when fitted to the same data. We interpret the value of 0 as the perturbations creating less variability than the intrinsic stochasticity of this approach.

This is not a general feature of stochastic approaches, as in neural networks and boosted regression trees the intrinsic stochasticity seems to be much less influential, and both approaches yielded relatively consistent GDF values. The actual GDF estimate of course depends on the settings of the methods and should not be interpreted as representative.

To compute the GDF for normally distribute data, we have to choose the strength of perturbation. The GDF value is robust to this choice, unless many data points are disturbed (Fig. 2). Only for BRT does increasing the strength of perturbation lead to a consistent, but small, decrease in GDF estimates, suggesting again that BRTs fit into the noise.

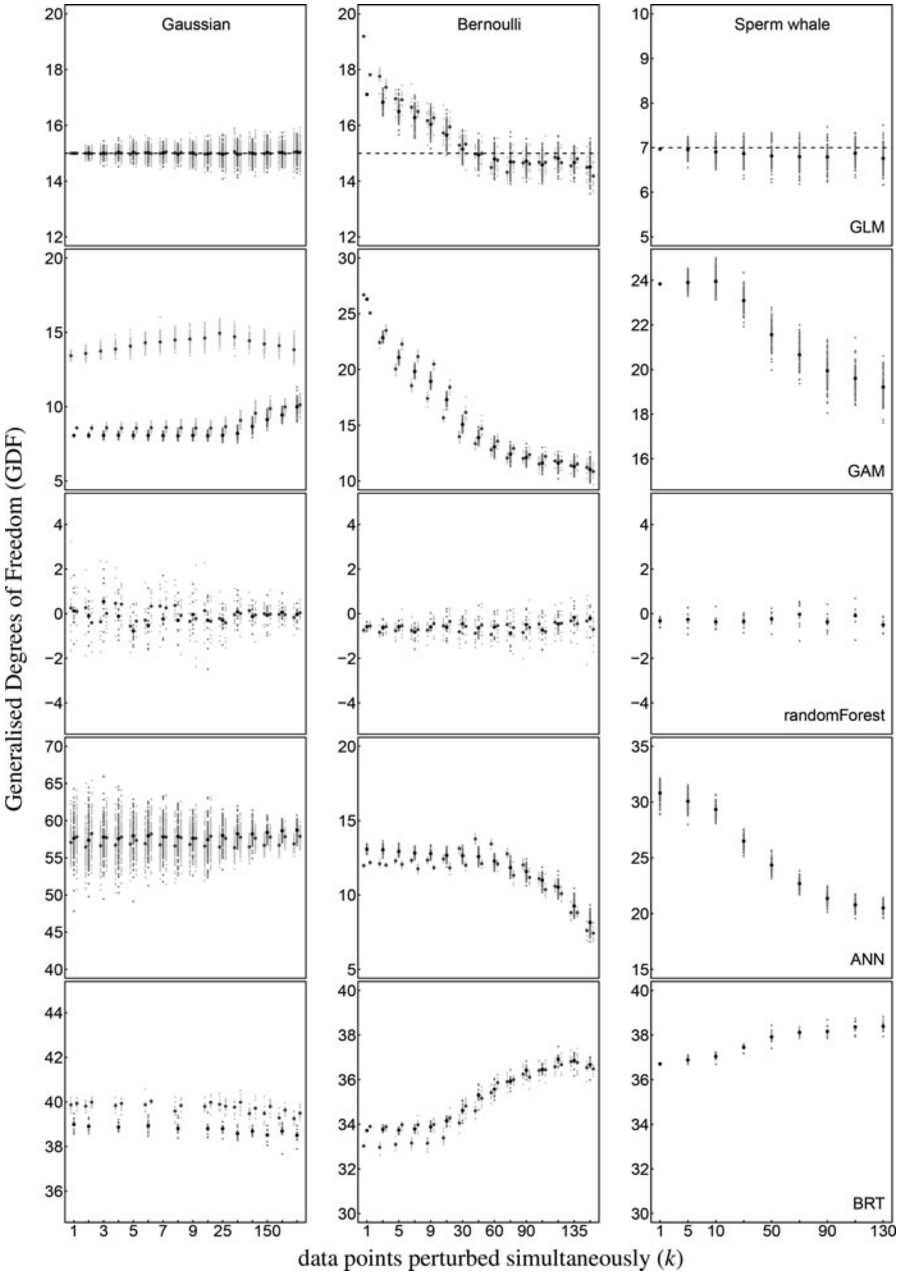

Figure 1. Models' GDF as a function of the numbers of data points perturbed (k) (ranked abscissa) for the simulated Gaussian (left), Bernoulli (center), and sperm whale data (right column) and the five model types (rows). The filled black ● and dark gray dots ● represent the mean GDF for the two replications at each level of k (in light-gray open dots ○ and black open dots ○). The dashed line in the first row - - - - represents the number of model parameters of the GLM.

Estimates of GDF for randomForest and ANN (but not BRT) respond with decreasing variance to increasing the strength of perturbation. Increasing the intensity of perturbation seems to overrule their internal stochasticity.

It seems clear from the results presented so far that no single best perturbation strength and optimal proportion of data to perturb exists for all methods. For the following analyses, we

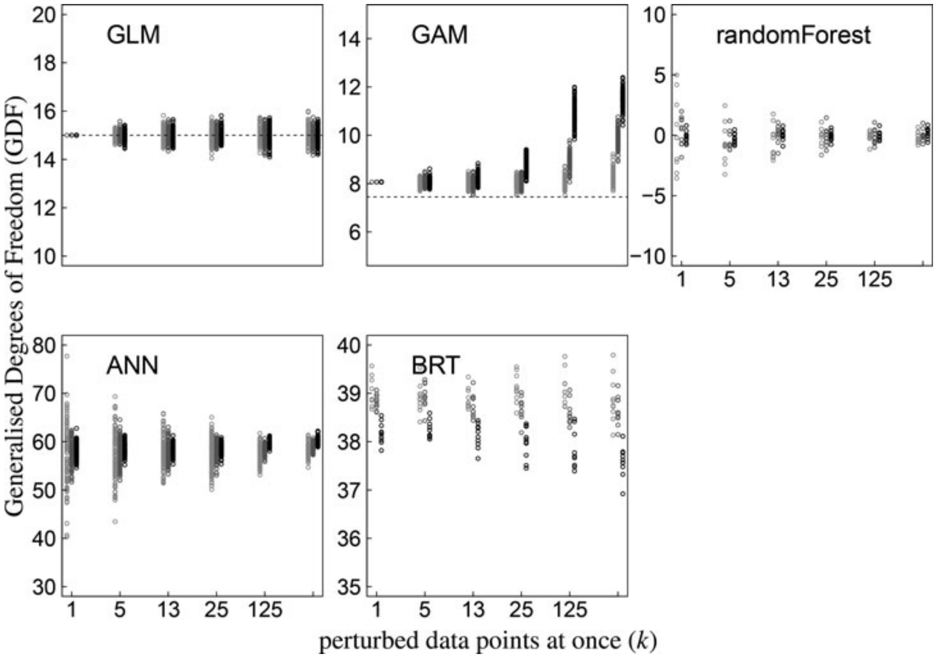

Figure 2. Estimated GDF along an increasing numbers of perturbed data points (k) for the simulated Gaussian data. For each k, there are 100 GDF replications computed with $\sigma = 0.125\sigma_y$ (○), $\sigma = 0.25\sigma_y$ (○), and $\sigma = 0.5\sigma_y$ (○), hence lower to higher perturbation magnitude. The dashed line - - - - represents the model's rank (GLM) and the sum of estimated degrees of freedom (GAM). *N.B.*, the *x*-axis is rank-transformed.

use, for normal data, perturbation $= 0.25\sigma_y$ and $k = N$ (except for GAM, where $k = 0.2N$); and for Bernoulli data $k = 0.5N$ (except for BRT and ANN, where $k = 0.04N$).

3.2. Efficiency of GDF and cross-validation computations

Both GDF and cross-validation require multiple analyses. For the GDF, we need to run several perturbations, and possibly replicate this many times. For cross-validation, we may also want to repeatedly perform the 10-fold cross-validation itself to yield stable estimates. The mean GDF and CV-log-likelihood (ℓ_{CV}) over 1–1,000 replicates are depicted in Fig. 3. With 100 runs, both estimates have stabilized, but 100 runs for GDF represent 25,000 model evaluations (due to the $k = 50$ perturbations and 50 internal replicates for a dataset of $N = 250$), while for 10-fold cross-validation these represent only 1,000 model evaluations, making it 25 times less costly.

3.3. GDF-based AIC versus cross-validation

We analyzed four datasets with the above settings for GDF and cross-validation, two simulated (Gaussian and Bernoulli) and two real-world datasets (sperm whale and red fox geographic distributions).

Both generalized degrees of freedom (GDF_m) and cross-validation log-likelihood-differences $(\Delta\ell_m^{\text{CV}})$ measure model complexity in an asymptotically equivalent way (Section 1.2). For finite datasets both approaches yield identical rankings of model complexity but are rather different in absolute value (Eq. (2)). Particularly the red fox-case study yields

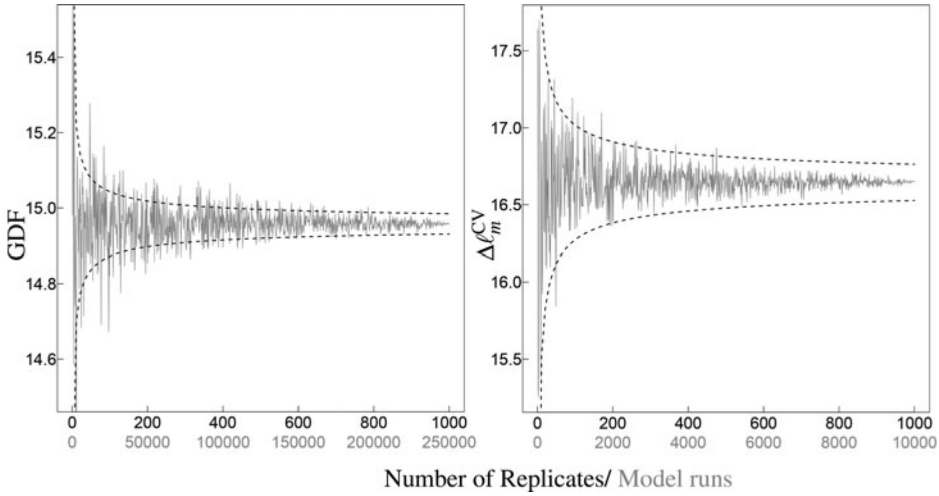

Figure 3. Development of mean generalized degrees of freedom (GDF, left) and cross-validation log-likelihood-differences ($\Delta\ell_m^{CV}$, right) over 1,000 replicates (250,000 and 10,000 model runs, respectively). GDF computation with $k = 50$ data points perturbed per model run and 50 internal replicates for the Gaussian simulation data ($N = 250$), i.e., 250 model evaluations per replicate. $\Delta\ell_m^{CV}$ derives from 10-fold cross-validation, i.e., 10 model evaluations per replicate. The dashed lines display the mean GDF and $\Delta\ell_m^{CV}$, respectively, of all replicates ± one standard error.

very low GDF estimates for GLM, GAM, and randomForest, while their cross-validation model complexity is much higher.

Model complexity is only one term in the AIC-formula (Eq. (8)) and for large datasets the log-likelihood will dominate. To put the differences between the two measures of model complexity into perspective, we computed AIC_{GDF} for all datasets and compared it to the equivalent cross-validation deviance ($-2\ell_{\text{CV}}$).

Across the entire range of datasets analyzed both approaches yielded extremely similar results (Fig. 4, right). Within each dataset, however, the pattern is more idiosyncratic, revealing a high sensitivity to low sample size ($N < 1{,}000$, i.e., all datasets except the red fox).

3.4. GDF, $\Delta\ell_m^{CV}$, and model weights

For all four datasets, one modeling approach always substantially outperformed the others, making model averaging an academic exercise. We compared model weights (according to Eqs. (11) and (12)) purely to quantify the differences in AIC_{GDF} and cross-validation deviance for model averaging. Only for the Bernoulli simulation was the difference noticeable (Table 1). Here GLM and randomForest shared the model weight when quantified based on cross-validation, while for the GDF approach GLM took all the weight.

4. Discussion

So far, Ye's (1998) generalized degrees of freedom concept did not attract much attention in the statistical literature, even though it builds on established principles and applies to machine learning, where model complexity is unknown. Shen and Huang (2006) have explored the perturbation approach to GDF in the context of adaptive model selection for linear models and later extended it to linear mixed effect models (Zhang et al., 2012). They also extended it to the exponential family (including the Bernoulli distribution) and even to classification and

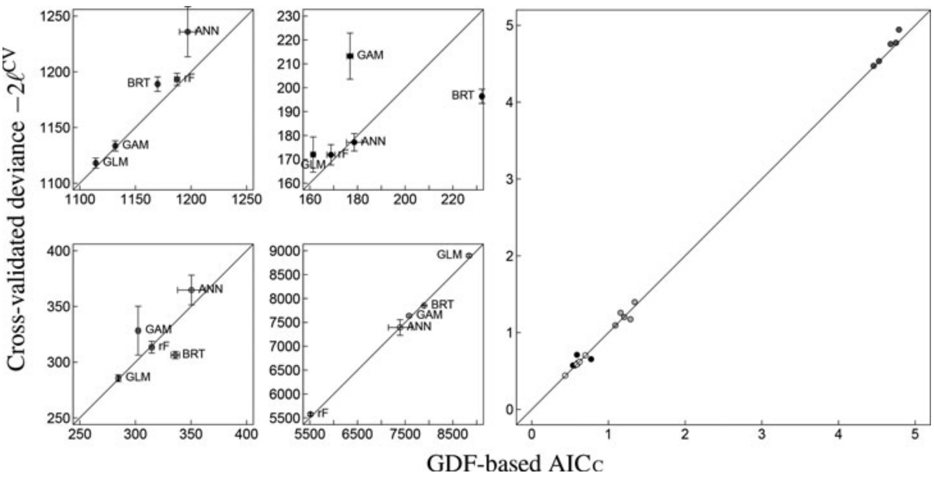

Figure 4. Cross-validated deviance versus GDF-based AICc. Error bars display ± one standard error (sometimes too small to be visible). The identity line ($y = x$) indicates the equivalence of both measures. The levels of gray represent the four datasets: Gaussian simulated (top-left, •) and Bernoulli simulated (top-center, •); sperm whale (bottom-left, •) and red fox (bottom-center, ○) data. In the right panel per datum GDF and $-2\ell^{CV}$, respectively, i.e., divided by the number of data points.

regression trees (Shen et al., 2004). Our study differs in that the models considered are more diverse and internally include weighted averaging, which apparently posed a challenge to the GDF-algorithm.

4.1. GDF for normally distributed data

For normally distributed data, our explorations demonstrate a low sensitivity to the intensity of perturbation used to compute GDF. Furthermore, across the five modeling approaches employed here, GDF estimates are stable and constant for different numbers of data points perturbed simultaneously.

GDF estimates were consistent with the rank of GLM models and in line with the estimated degrees of freedom reported by the GAM. For neural networks and boosted regression trees,

Table 1. Comparison of Akaike weights w_{AIC} and cross-validation weights w_{CV} (see Eq. (12)) for Gaussian and Bernoulli simulation data and real-life sperm whale and red fox abundance data. The "surprisingly" good performance of GLMs in the sperm whale case study remains unexplained, but is fully reproducible.

Model	w_{AIC}	w_{CV}	w_{AIC}	w_{CV}
	Gaussian simulation		Bernoulli simulation	
GLM	0.9998	0.9996	0.9762	0.4682
GAM	$1.548 \cdot 10^{-4}$	$4.390 \cdot 10^{-4}$	$4.310 \cdot 10^{-4}$	$5.185 \cdot 10^{-10}$
rF	$1.407 \cdot 10^{-16}$	$4.772 \cdot 10^{-17}$	$2.319 \cdot 10^{-2}$	0.4960
ANN	$1.158 \cdot 10^{-18}$	$2.632 \cdot 10^{-26}$	$1.759 \cdot 10^{-4}$	$3.578 \cdot 10^{-2}$
BRT	$8.647 \cdot 10^{-13}$	$4.001 \cdot 10^{-16}$	$4.131 \cdot 10^{-16}$	$2.381 \cdot 10^{-6}$
	Sperm whale		Red fox	
GLM	0.9999	0.9997	0	0
GAM	$1.373 \cdot 10^{-4}$	$5.387 \cdot 10^{-10}$	0	0
rF	$3.079 \cdot 10^{-7}$	$8.966 \cdot 10^{-7}$	1	1
ANN	$5.369 \cdot 10^{-15}$	$6.558 \cdot 10^{-18}$	0	0
BRT	$7.473 \cdot 10^{-12}$	$2.996 \cdot 10^{-5}$	0	0

Table 2. Measures of model complexity (mean ± standard deviation): Generalized degrees of freedom (GDF) and $\Delta\ell_m^{CV}$ derived from cross-validation (see Section 1.2) for Gaussian and Bernoulli simulation data and real-world sperm whale and red fox distribution data. The true ranks for the GLMs are 15, 15, 7, and 8, respectively.

Model	GDF	$\Delta\ell_m^{CV}$	GDF	$\Delta\ell_m^{CV}$
	Gaussian simulation		Bernoulli simulation	
GLM	15.1 ± 0.30	16.7 ± 1.97	14.5 ± 0.41	19.4 ± 3.18
GAM	10.0 ± 0.41	10.7 ± 2.09	11.0 ± 0.58	27.0 ± 3.82
randomForest	− 0.0 ± 0.33	2.8 ± 2.67	− 0.2 ± 0.41	1.3 ± 2.10
ANN	58.7 ± 0.82	69.4 ± 5.76	12.6 ± 0.25	12.1 ± 1.67
BRT	38.5 ± 0.28	45.1 ± 2.17	34.2 ± 0.17	20.3 ± 1.29
	Sperm whale		Red fox	
GLM	6.8 ± 0.24	7.2 ± 1.40	7.9 ± 0.38	37.4 ± 13.02
GAM	19.2 ± 0.59	29.8 ± 7.63	9.0 ± 0.21	37.1 ± 10.37
randomForest	− 0.5 ± 0.26	− 1.1 ± 2.66	1.1 ± 5.83	31.7 ± 13.22
ANN	29.3 ± 0.49	35.0 ± 4.96	50.3 ± 0.66	64.2 ± 80.69
BRT	37.0 ± 0.17	26.0 ± 1.29	59.0 ± 0.12	54.6 ± 2.30

GDF estimates appear plausible, but cannot be compared with any self-reported values. Compared to the cross-validation method, GDF values are typically, but not consistently, lower by 10–30% (Table 2).

For randomForest, GDF estimates were essentially centered on zero. It seems strange to find that an algorithm that uses hundreds of classification and regression trees internally to actually have no (or even negative) degrees of freedom. We expected a low value due to the averaging of submodels (called the “paradox of ensembles” by Elder, 2003), but not such a complete insensitivity to perturbations. (Within this study, SH and CFD independently reprogrammed our GDF function to make sure that this was not due to a programming error.) As Eq. (4) shows, the perturbation of individual data points is compared to the change in the model expectation for this data point, and then summed over all data points. To yield a GDF of 0, the change in expectation (numerator) must be much smaller than the perturbation itself (denominator). This is possible when expectations are variable due to the stochastic nature of the algorithm. It seems that randomForest is much more variable than the other stochastic approaches of boosting and neural networks.

4.2. Bernoulli GDF

Changing the value of Bernoulli data from 0 into 1 (or vice versa) is a stronger perturbation than adding a small amount of normal noise to Gaussian data. As our exploration has shown, the GDF for such Bernoulli data are indeed much less well-behaved than for the normal data. Not only is the estimated GDF dependent on the number of data points perturbed, also is this dependence different for each modeling approach we used. This makes GDF computation impractical for Bernoulli data. As a consequence, we did not attempt to extend GDF in this way to other distributions, as in our perception only a general, distribution-, and model-independent algorithm is desirable.

4.3. GDF versus model complexity from cross-validation

Cross-validation is typically used to get a non-optimistic assessment of model fit (e.g., Hawkins, 2004). As we have shown, it can also be used to compute a measure of model complexity similar (in principle) to GDF (Eq. (8) and Table 2). Both express model complexity as the effective number of parameters fitted in the model. GDF and cross-validation-based model complexity estimator $\Delta\ell_m^{\mathrm{CV}}$ are largely similar, but may also differ substantially (Fig. 4, red fox case study). Since the "correct" value for this estimator is unknown, we cannot tell which approach actually works better. Given our inability to choose the optimal number of data points to perturb (except for GLM), we prefer $\Delta\ell_m^{\mathrm{CV}}$, which does not make any such assumption.

4.4. Remaining problems

To make the GDF approach more generally applicable, a new approach has to be found. The original idea of Ye (1998) is appealing, but not readily transferable in the way we had hoped.

Another problem, even for Gaussian data where this approach seems to be performing fine, is the high computational burden. GDF estimation requires tens of thousands of model evaluations, giving it very limited appeal, except for small datasets and fast modeling approaches. Cross-validation, as alternative, is at least an order of magnitude faster, but still requires around 1,000 evaluations. If the aim is to compute model weights for model averaging, no *precise* estimation of model complexity is needed and even the results of a single K-fold cross-validation based on Eq. (8) can be used. It was beyond the scope of this study to develop an efficient cross-validation-based approach to compute degrees of freedom, but we clearly see this as a more promising way forward than GDF.

4.5. Alternatives to AIC

The selection of the most appropriate statistical model is most commonly based on Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951), a measure representing the distance between the true and an approximating model. Thus, we assume that model m, for which the distance to the true model is minimal, is the KL-best model. Yet, since KL-discrepancy is not observable, even if a true model existed, many statisticians have attempted to find a metric approximation (e.g., Burnham and Anderson, 2002; Burnham et al., 1994). Akaike (1973), who proposed this measure as the basis for model selection in the first place, developed the AIC to get around the discussed problem.

The point of the cross-validated log-likelihood is that we do away with the approximation that yields the degrees of freedom term in the AIC, instead estimating the model-dependent part of the KL-divergence directly. This approach is disadvantageous if AIC can be computed from a single model fit. But if the effective degrees of freedom (EDF) terms for the AIC would require repeated model fits then there is no reason to use the AIC-approximation to the KL-divergence, rather than a more direct estimator. If leave-one-out cross-validation is too expensive, then we can leave out several data points, at the cost of some Monte Carlo variability (resulting from the fact that averaging over all possible left-out sets is generally impossible).

Choosing the number of cross-validation folds, K, is a challenge and a frequently debated issue. Many studies refer to Kohavi (1995), who proposed $K = 10$, even if the choice is not limited by computational feasibility.

5. Conclusion

We have shown that the idea of using GDF to extend information-theoretical measures of model fit (such as AIC) to non-likelihood models is burdened with large computational costs and yields variable results for different modeling approaches. Cross-validation is more variable than GDF, but a more direct way to compute measures of model complexity, fit, and weights (in a model averaging context). As cross-validation may, but need not, employ the likelihood fit to the hold-out, it appears more plausible for models that do not make likelihood assumptions. Thus, we recommend repeated (>100 times) K-fold cross-validation to estimate any of the statistics under consideration.

Funding

This work was partly performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

ORCID

Severin Hauenstein http://orcid.org/0000-0001-8003-0944
Simon N. Wood http://orcid.org/0000-0002-2034-7453
Carsten F. Dormann http://orcid.org/0000-0002-9835-1794

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., Csaki, F. (eds) 2nd *Int. Symp. Inf. Theory.* Akademiai Kiado, Budapest, pp 267–281.

Burnham, K. P., Anderson, D. R. (1998). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Berlin: Springer.

Burnham, K. P., Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. Berlin: Springer.

Burnham, K. P., Anderson, D. R., White, G. C. (1994). Evaluation of the Kullback-Leibler discrepancy for model selection in open population capture-recapture models. *Biometrical Journal* 36(3): 299–315.

Diniz-Filho, J. A. F., Rangel, T. F. L. V. B., Bini, L. M. (2008). Model selection and information theory in geographical ecology. *Global Ecology and Biogeography* 17:479–488.

Dormann, C., Kaschner, K. (2010). Where's the Sperm Whale? A Species Distribution Example Analysis. Available at: http://www.mced-ecology.org/?page_id=355.

Dormann, C. F., Gruber, B., Winter, M., Herrmann, D. (2010). Evolution of climate niches in European mammals? *Biology Letters* 6:229–232.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B* 57:45–97.

Elder, J. F. (2003). The generalization paradox of ensembles. *Journal of Computational and Graphical Statistics* 12(4):853–864.

Elith, J., Graham, C. H., Anderson, R. P., DudÃ-k, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., SoberÃ³n, J., Williams, S., Wisz, M. S., Zimmermann, N. E., Araujo, M. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29(2):129–151.

Elith, J., Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40(1):677–697.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Berlin: Springer.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences* 44(1):1–12.

Hegyi, G., Garamszegi, L. Z. (2011). Using information theory as a substitute for stepwise regression in ecology and behavior. *Behavioral Ecology and Sociobiology* 65(1):69–76.

Horne, J. S., Garton, E. O. (2006). Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in kernel home-range analysis. *Journal of Wildlife Management* 70:641–648.

Hurvich, C. M., Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* 76(2):297–307.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Mellish, C. S. (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, pp. 1137–1143. Available at: http://robotics.stanford.edu/~ronnyk.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 1:79–86.

Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. *R News* 2(3):18–22.

Madigan, D., York, J., Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique* 63(2):215–232.

Mundry, R. (2011). Issues in information theory-based statistical inference – commentary from a frequentist's perspective. *Behavioral Ecology and Sociobiology* 65(1):57–68.

Olden, J. D., Lawler, J. J., Poff, N. L. (2008). Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology* 83(2):171–193.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* 83(2):251–266.

Recknagel, F. (2001). Applications of machine learning to ecological modelling. *Ecological Modelling* 146(1–3):303–310.

Ridgeway, G., et al. (2013). *gbm: Generalized Boosted Regression Models, R Package Version 2.1*:CRAN.

Shen, X., Huang, H.-C. (2006). Optimal model assessment, selection, and combination. *Journal of the American Statistical Association* 101(474):554–568.

Shen, X., Huang, H.-C., Ye, J. (2004). Adaptive model selection and assessment for exponential family distributions. *Technometrics* 46(3):306–317.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* 64:583–639.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* 39(1):44–47.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods* 7(1):13–26.

Team, R. C. (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Turkheimer, F. E., Hinz, R., Cunningham, V. J. (2003). On the undecidability among kinetic models: From model selection to model averaging. *Journal of Cerebral Blood Flow and Metabolism* 23(4):490–498.

Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S*. 4th ed. New York: Springer.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. New York: Chapman and Hall/CRC.

Wood, S. N. (2015). *Core Statistics*. Cambridge: Cambridge University Press.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* 93(441):120–131.

Zhang, B., Shen, X., Mumford, S. L. (2012). Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. *Computational Statistics and Data Analysis* 56(3):574–586.