A Small-Sample Choice of the Tuning Parameter in Ridge Regression

Philip S. Boonstra, Bhramar Mukherjee, and Jeremy M. G. Taylor

Department of Biostatistics, University of Michigan, Ann Arbor 48109

Abstract

We propose new approaches for choosing the shrinkage parameter in ridge regression, a penalized likelihood method for regularizing linear regression coefficients, when the number of observations is small relative to the number of parameters. Existing methods may lead to extreme choices of this parameter, which will either not shrink the coefficients enough or shrink them by too much. Within this "small-n, large-p" context, we suggest a correction to the common generalized cross-validation (GCV) method that preserves the asymptotic optimality of the original GCV. We also introduce the notion of a "hyperpenalty", which shrinks the shrinkage parameter itself, and make a specific recommendation regarding the choice of hyperpenalty that empirically works well in a broad range of scenarios. A simple algorithm jointly estimates the shrinkage parameter and regression coefficients in the hyperpenalized likelihood. In a comprehensive simulation study of small-sample scenarios, our proposed approaches offer superior prediction over nine other existing methods.

Keywords: Akaike's information criterion, Cross-validation, Generalized cross-validation, Hyperpenalty, Marginal likelihood, Penalized likelihood

1 Introduction

Suppose we have data, $\{\boldsymbol{y}, \boldsymbol{x}\}$, comprising n observations of a continuous outcome Y and p covariates $\boldsymbol{X}$, with the covariate matrix $\boldsymbol{x}$ regarded as fixed. The quantity n is assumed to be approximately equal to or less than p. We relate Y and $\boldsymbol{X}$ by a linear model, $Y = \beta_0 + \boldsymbol{X}^\top\boldsymbol{\beta} + \sigma\varepsilon$, with $\varepsilon \sim N\{0, 1\}$. Up to an additive constant, the log-likelihood is

$$\ell(\boldsymbol{\beta}, \beta_0, \sigma^2) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \beta_0\mathbf{1}_n - \boldsymbol{x}\boldsymbol{\beta})^\top(\boldsymbol{y} - \beta_0\mathbf{1}_n - \boldsymbol{x}\boldsymbol{\beta}). \tag{1}$$

We center $\boldsymbol{y}$ and standardize $\boldsymbol{x}$ to have unit variance. As a consequence of this, although β_0 is estimated in the fitted model, our notation will implicitly reflect the assumption $\beta_0 = 0$.

We consider penalized estimation of $\boldsymbol{\beta}$, with our primary interest being prediction of future observations, rather than variable selection. Thus, we focus on L_2-penalization, i.e. ridge regression (Hoerl and Kennard, 1970), which, from a prediction perspective, can have favorable properties compared to other penalization methods (e.g. Frank and Friedman, 1993; Tibshirani, 1996; Fu, 1998; Zou and Hastie, 2005). Ridge regression may be viewed as a hierarchical linear model, similar to mixed effects modeling. Here the "random effects" are the elements of $\boldsymbol{\beta}$. An L_2-penalty on $\boldsymbol{\beta}$ implicitly assumes these are jointly and independently Normal with mean zero and variance σ^2/λ, because the penalty term matches the negative Normal log-density, up to a normalizing constant not depending on $\boldsymbol{\beta}$:

$$p_\lambda(\boldsymbol{\beta}, \sigma^2) = \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} - \frac{p}{2}\ln(\lambda) + \frac{p}{2}\ln(\sigma^2). \tag{2}$$

The scalar λ is the ridge parameter, controlling the shrinkage of $\boldsymbol{\beta}$ toward zero; larger values yield greater shrinkage. Given λ, the maximum penalized likelihood estimate of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}_\lambda = \arg\max_{\boldsymbol{\beta}|\lambda}\left\{\ell(\boldsymbol{\beta}, \sigma^2) - p_\lambda(\boldsymbol{\beta}, \sigma^2)\right\} = (\boldsymbol{x}^\top\boldsymbol{x} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{x}^\top\boldsymbol{y}. \tag{3}$$

When $n-1 \geq p$, a key result from Hoerl and Kennard (Theorem 4.3, 1970) is that $\lambda^* = \arg\min_{\lambda\geq 0}\mathrm{E}[(\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda)] > 0$, i.e. there exists $\lambda^* > 0$ for which the mean squared error (MSE) of $\boldsymbol{\beta}_\lambda$ decreases relative to $\lambda = 0$. If $\boldsymbol{x}^\top\boldsymbol{x}/n = \boldsymbol{I}_p$, then $\lambda^* = p\sigma^2/\boldsymbol{\beta}^\top\boldsymbol{\beta}$; however, there is no closed-form solution for λ^* in the general $\boldsymbol{x}^\top\boldsymbol{x}$ case. A strictly positive λ introduces bias in $\boldsymbol{\beta}_\lambda$ but decreases variance, making a bias-variance tradeoff. A choice of λ which is too small leads to overfitting the data, and one which is too large shrinks $\boldsymbol{\beta}$ by too much. To contrast these extremes, we will hereafter refer to this latter scenario as "underfitting." The existence of λ^* is relevant because prediction error, $\mathrm{E}[(\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda)^\top\boldsymbol{x}^\top\boldsymbol{x}(\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda)]$, is closely related to MSE and may correspondingly benefit from such a bias-variance tradeoff.

To approximate λ^*, one cannot simply maximize $\ell(\boldsymbol{\beta}, \sigma^2) - p_\lambda(\boldsymbol{\beta}, \sigma^2)$ jointly with respect to $\boldsymbol{\beta}$, σ^2 and λ, because the expression can be made arbitrarily large by plugging in $\boldsymbol{\beta} = 0$ and letting $\lambda \to \infty$. Typically, λ is selected by optimizing some other objective function. Our motivation for this paper is to investigate selection strategies for λ when n is "small", by which we informally mean $n < p$ or $n \approx p$, the complement being a more standard $n \gg p$ situation. This small-n situation increasingly occurs in modern genomic studies, whereas common approaches for selecting λ are often justified asymptotically in n.

Our contribution is two-fold. First, we present new ideas for choosing λ, including both a small-sample modification to a common existing approach and novel proposals. Our framework categorizes existing strategies into two classes, based on whether a goodness-of-fit criterion or a likelihood is optimized. Methods in either class may be susceptible to over- or underfitting; a third, new class extends the hierarchical perspective of ridge regression, the first level being $\ell(\boldsymbol{\beta}, \sigma^2)$ and the second $p_\lambda(\boldsymbol{\beta}, \sigma^2)$. Following ideas by Takada (1979), who showed that Stein's Positive Part Estimator corresponds to a posterior mode given a certain prior, and, more recently, Strawderman and Wells (2012), who place a hyperprior on the Lasso penalty parameter, we add a third level, defining a "hyperpenalty" on λ. This hyperpenalty induces shrinkage on λ itself, thereby protecting against extreme choices of λ. The second contribution follows naturally, namely, a comprehensive evaluation of all methods, both existing and newly proposed, in this small-n situation via simulation studies.

The remainder of this paper is organized as follows. We review current approaches for choosing λ (the first and second classes discussed above) in Sections 2 and 3 and propose a small-sample modification to one of these methods, generalized cross-validation (GCV, Craven and Wahba, 1979). In Section 4, we define a generic hyperpenalty function and explore a specific choice for the form of hyperpenalty in 4.1 and 4.2. Section 5 conducts a comprehensive simulation study. Our results suggest that the existing approaches for choosing λ can be improved upon in many small-n cases. Section 7 concludes with a discussion, in which we discuss useful extensions of the hyperpenalty framework.

2 Goodness-of-fit-based methods for selection of λ

These methods define an objective function in terms of λ which is to be minimized. Commonly used is K-fold cross-validation, which partitions observations into K groups, $\kappa(1), \ldots, \kappa(K)$, and calculates $\boldsymbol{\beta}_\lambda$ K times using equation (3), each time leaving out group $\kappa(i)$, to get $\boldsymbol{\beta}_\lambda^{-\kappa(1)}, \boldsymbol{\beta}_\lambda^{-\kappa(2)}$, etc. For $\boldsymbol{\beta}_\lambda^{-\kappa(i)}$, cross-validated residuals are calculated on the observations in $\kappa(i)$, which did not contribute to estimating $\boldsymbol{\beta}$. The objective function estimates prediction error and is the sum of the squared cross-validated residuals:

$$\lambda_{\text{K-CV}} = \arg\min_\lambda \ln \sum_{i=1}^{K} (\boldsymbol{y}_{\kappa(i)} - \boldsymbol{x}_{\kappa(i)} \boldsymbol{\beta}_\lambda^{-\kappa(i)})^\top (\boldsymbol{y}_{\kappa(i)} - \boldsymbol{x}_{\kappa(i)} \boldsymbol{\beta}_\lambda^{-\kappa(i)}). \tag{4}$$

A suggested choice is $K = 5$ (Hastie et al., 2009). When $K = n$, some simplification (Golub et al., 1979) gives

$$\lambda_{n\text{-CV}} = \arg\min_\lambda \ln \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_\lambda)^2 / (1 - P_{\lambda[ii]} - 1/n)^2 \tag{5}$$

$$\text{with } \boldsymbol{P}_\lambda = \boldsymbol{x}(\boldsymbol{x}^\top \boldsymbol{x} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{x}. \tag{6}$$

$P_{\lambda[ii]}$ is the ith diagonal element of $\boldsymbol{P}_\lambda$ and measures the ith observation's influence in estimating $\boldsymbol{\beta}$. Further discussion of its interpretation is given in Section 2.1. From (5), observations for which $P_{\lambda[ii]}$ is large, ie influential observations, have greater weight. Re-centering $\boldsymbol{y}$ at each fold implies β_0 is re-estimated; this is reflected by the "$-1/n$" term in (5). This term does not appear in the derivations by Golub et al. (1979), which assume β_0 is known, but this difference in assumptions is important with regard to GCV, which is discussed next, and our proposed extension of GCV.

GCV multiplies each squared residual in (5) by $(1 - P_{\lambda[ii]} - 1/n)^2/(1 - \text{Trace}(\boldsymbol{P}_\lambda)/n - 1/n)^2$, thereby giving equal weight to all observations. Using the equality $\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}_\lambda = (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)\boldsymbol{y}$, further simplification yields

$$\lambda_{\text{GCV}} = \arg\min_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^2 \boldsymbol{y} - 2\ln(1 - \text{Trace}(\boldsymbol{P}_\lambda)/n - 1/n) \right\}. \tag{7}$$

Although derived using different principles, other methods reduce to a "model fit + penalty" or "model fit + model complexity" form similar to (7): Akaike's Information Criterion (AIC, Akaike, 1973) and the Bayesian Information Criterion (BIC, Schwarz, 1978). Respectively, each chooses λ as follows:

$$\lambda_{\text{AIC}} = \arg\min_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^2 \boldsymbol{y} + 2(\text{Trace}(\boldsymbol{P}_\lambda) + 2)/n \right\}, \tag{8}$$

$$\lambda_{\text{BIC}} = \arg\min_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^2 \boldsymbol{y} + \ln(n)(\text{Trace}(\boldsymbol{P}_\lambda) + 2)/n \right\}. \tag{9}$$

Asymptotically in n, GCV will choose the value of λ which minimizes the prediction criterion $\mathrm{E}\left[(\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda)^\top \boldsymbol{x}^\top \boldsymbol{x}(\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda)\right]$ (Golub et al., 1979; Li, 1986). Further, Golub et al. observe that GCV and AIC asymptotically coincide. BIC asymptotically selects the true underlying model from a set of nested candidate models (Sin and White, 1996; Hastie et al., 2009), so its justification for use in selecting λ, which is a shrinkage parameter, is weak. For all of these methods, optimality is based upon the assumption that $n \gg p$. When n is small, extreme overfitting is possible (Wahba and Wang, 1995; Efron, 2001), giving small bias/large variance estimates. A small-sample correction of AIC (AIC$_C$, Hurvich and Tsai, 1989; Hurvich et al., 1998) and a robust version of GCV (RGCV$_\gamma$, Lukas, 2006) exist:

$$\lambda_{\text{AIC}_C} = \arg\min_\lambda \big\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^2 \boldsymbol{y} + 2(\text{Trace}(\boldsymbol{P}_\lambda)+2)/(n-\text{Trace}(\boldsymbol{P}_\lambda)-3)) \big\}, \tag{10}$$

$$\begin{aligned}\lambda_{\text{RGCV}_\gamma} = \arg\min_\lambda \big\{ &\ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^2 \boldsymbol{y} - 2\ln(1-\text{Trace}(\boldsymbol{P}_\lambda)/n - 1/n) \\ &+ \ln(\gamma + (1-\gamma)\text{Trace}(\boldsymbol{P}_\lambda^2)/n) \big\}. \end{aligned} \tag{11}$$

For AIC$_C$, the modified penalty is the product of the original penalty, $2(\text{Trace}(\boldsymbol{P}_\lambda)+2)/n$, and $n/(n-\text{Trace}(\boldsymbol{P}_\lambda)-3)$. The authors do not consider the possibility of $n-\text{Trace}(\boldsymbol{P}_\lambda)-3<0$, which would inappropriately change the sign of the penalty, and we have found no discussion of this in the literature. In our implementation of AIC$_C$, we replace $n-\text{Trace}(\boldsymbol{P}_\lambda)-3$ with its positive part, $(n-\text{Trace}(\boldsymbol{P}_\lambda)-3)_+$, effectively making the criterion infinitely large in this case. As a rule of thumb, Burnham and Anderson (2002) suggest to use AIC$_C$ over AIC when $n < 40p$ (their threshold for small n) and thus also when $n \approx p$. RGCV$_\gamma$ subtracts another penalty from GCV based on a tuning parameter $\gamma \in (0,1]$, as in (11); we use $\gamma = 0.3$ based on Lukas' recommendation. Small choices of λ are more severely penalized, thereby offering protection against overfitting. To the best of our knowledge, the performance of AIC$_C$ or RGCV$_\gamma$ in the context of selecting λ in ridge regression has not been extensively studied.

2.1 Small-sample GCV

Trace($\boldsymbol{P}_\lambda$), with $\boldsymbol{P}_\lambda$ defined in (6), is the effective number of model parameters, excluding β_0 and σ^2. It decreases monotonically with $\lambda > 0$ and lies in the interval $(0, \min\{n-1, p\})$. The upper bound on Trace($\boldsymbol{P}_\lambda$) is not $\min\{n, p\}$ because the standardization of $\boldsymbol{x}$ reduces its rank by one when $n \leq p$. Although the parameters β_0 and σ^2 are counted (in the literal sense) in the model complexity terms of AIC and BIC, they have only an additive effect, being represented by the "$+\,2$" expressions in (8) and (9). For this reason, β_0 and σ^2 may be ignored in considering model complexity. However, from (7), GCV counts β_0, which is given by the "$-1/n$" term, but not σ^2; counting both *will* change the penalty, since the model complexity term is on the log-scale. This motivates our proposed small-sample correction to GCV, called GCV$_C$, which *does* count σ^2 as a parameter:

$$\lambda_{\text{GCV}_C} = \arg\min_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^2 \boldsymbol{y} - 2\ln((1 - \text{Trace}(\boldsymbol{P}_\lambda)/n - 2/n)_+) \right\}. \qquad (12)$$

As with AIC$_C$, $1 - \text{Trace}(\boldsymbol{P}_\lambda)/n - 2/n$ may be negative. In this case, subtracting the log of the positive part of $1 - \text{Trace}(\boldsymbol{P}_\lambda)/n - 2/n$ makes the objective function infinite. This is only a small-sample correction because the objective functions in (7) and (12) coincide as $n \to \infty$, and the asymptotic optimality of GCV transfers to GCV$_C$.

An explanation of why GCV$_C$ corrects the small-sample deficiency of GCV is as follows. If $n-1 = p$, the model-fit term in the objective function of (7), $\ln \boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^2\boldsymbol{y}$, tends to $-\infty$ as λ decreases. When $\lambda = 0$, the fitted values, $\boldsymbol{P}_\lambda \boldsymbol{y}$, will perfectly match the observations, $\boldsymbol{y}$, and the data are overfit. The penalty term, $-2\ln(1 - \text{Trace}(\boldsymbol{P}_\lambda)/n - 1/n)$, tends to ∞ as λ decreases, because Trace($\boldsymbol{P}_\lambda$) approaches $n-1$. The rates of convergence for the model-fit and penalty terms determine whether GCV chooses a too-small λ. If the model-fit term approaches $-\infty$ faster than the penalty approaches ∞, the objective function is minimized by setting λ as small as possible, which is $\lambda = 0$ when $n-1 = p$. Although this phenomenon is most striking in cases for which $n-1 = p$, as we will see in Section 5, this finding appears to hold when $n-1<p$. In this case, predictions will *nearly* match observations as λ decreases but remains numerically positive to allow for the matrix inversion in $\boldsymbol{P}_\lambda$, and the penalty term still approaches ∞ as λ decreases. Like GCV, the penalty function associated with GCV$_C$ also approaches ∞ as λ decreases. In contrast to GCV, however, the GCV$_C$ penalty equals ∞ when $\lambda = \tilde{\lambda} > 0$, where $\tilde{\lambda}$ is the solution to $1 - \text{Trace}(\boldsymbol{P}_\lambda)/n - 2/n = 0$, or, equivalently, $\text{Trace}(\boldsymbol{P}_\lambda) = n-2$. In other words, when fitting GCV$_C$, the effective number of remaining parameters, beyond σ^2 and β_0, will be less than $n-2$, and perfect fit of the observations to the predictions, i.e. $\lambda = 0$, cannot occur.

<u>Remark 1</u>: A reviewer observed that the GCV$_C$ penalty can be generalized according to $-2\ln((1-\text{Trace}(\boldsymbol{P}_\lambda)/n - c/n)_+)$ for $c \geq 1$; special cases of this include GCV ($c=1$) and GCV$_C$ ($c=2$). Extending the interpretation given above, this ensures that the effective number of remaining parameters, beyond σ^2 and β_0, will be less than $n-c$, rather than $n-2$. Preliminary results from allowing c to vary did not point to a uniformly better choice of $c>2$. Also, using $c=2$ is consistent with our original motivation for proposing GCV$_C$, namely properly counting the model parameters.

3 Likelihood-based methods for selection of λ

A second approach treats the ridge penalty in (2) as a negative log-density. One can consider a marginal likelihood, where λ is interpreted as the variance component of a mixed-effects model:

$$\begin{aligned} m(\lambda,\sigma^2) &= \ln \int_{\boldsymbol{\beta}} \exp\{\ell(\boldsymbol{\beta},\sigma^2) - p_\lambda(\boldsymbol{\beta},\sigma^2)\}\mathrm{d}\boldsymbol{\beta} \\ &= -\frac{1}{2\sigma^2}\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{P}_\lambda)\boldsymbol{y} - \frac{n}{2}\ln(\sigma^2) + \frac{1}{2}\ln\left|\boldsymbol{I}_n - \boldsymbol{P}_\lambda\right|. \end{aligned} \tag{13}$$

From this, $\boldsymbol{y}|\lambda, \sigma^2$ is multivariate Normal with mean $\mathbf{0}_n$ ($\boldsymbol{y}$ is centered) and covariance $\sigma^2(\boldsymbol{I}_n - \boldsymbol{P}_\lambda)^{-1}$. The maximum profile marginal likelihood (MPML) estimate, originally proposed for smoothing splines (Wecker and Ansley, 1983), profiles $m(\lambda, \sigma^2)$ over σ^2, replacing each instance with $\hat{\sigma}^2_\lambda = \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda) \boldsymbol{y}/n$, and maximizes the "concentrated" log-likelihood, $m(\lambda, \hat{\sigma}^2_\lambda)$:

$$\lambda_{\text{MPML}} = \arg\min_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda) \boldsymbol{y} - \frac{1}{n} \ln \left| \boldsymbol{I}_n - \boldsymbol{P}_\lambda \right| \right\}. \tag{14}$$

Closely related is the generalized/restricted MPML (GMPML, Harville, 1977; Wahba, 1985), which adjusts the penalty to account for estimation of regression parameters that are not marginalized. Here, only β_0 is not marginalized, so the adjustment is by one degree of freedom (see Supplement S1):

$$\lambda_{\text{GMPML}} = \arg\min_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{P}_\lambda) \boldsymbol{y} - \frac{1}{n-1} \ln \left| \boldsymbol{I}_n - \boldsymbol{P}_\lambda \right| \right\}. \tag{15}$$

In a smoothing-spline comparison of GMPML to GCV, Wahba (1985) found mixed results, with neither method offering uniformly better predictions. For scatterplot smoothers, Efron (2001) notes that GMPML may oversmooth, yielding large bias/small variance estimates.

<u>REMARK 2</u>: Rather than profiling over σ^2, one could jointly maximize $m(\lambda, \sigma^2)$ over λ and σ^2. We have not found this approach previously used as a selection criterion in ridge regression. Our initial investigation of this and its restricted likelihood counterpart gave results similar to MPML and GMPML, and so we do not consider it further.

An alternative to the marginal likelihood methods described above is to treat the objective function in (3) as an h-log-likelihood, or "h-loglihood", of the type proposed by Lee and Nelder (1996) for hierarchical generalized linear models. The link between penalized likelihoods, like ridge regression, and the h-loglihood was noted in the paper's ensuing discussion. To estimate σ^2 (the dispersion) and λ (the variance component), Lee and Nelder suggested an iterative profiling approach, yielding the maximum adjusted profile h-loglihood (MAPHL) estimate. In Supplement S2, we show one iteration proceeds as follows:

$$\sigma^{2(i)} \leftarrow \frac{\big(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\big) + \lambda^{(i-1)}\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}}{n-1} \tag{16}$$

$$\lambda^{(i)} \leftarrow \arg\min_\lambda \left\{\lambda\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}/\sigma^{2(i)} - \ln\left|\boldsymbol{I}_n - \boldsymbol{P}_\lambda\right|\right\} \tag{17}$$

$$\boldsymbol{\beta}^{(i)} \leftarrow \boldsymbol{\beta}_{\lambda^{(i)}} \tag{18}$$

and $\lambda_{\text{MAPHL}} = \lambda^{(\infty)}$.

Finally, Tran (2009) proposed the "Loss-rank" (LR) method for selecting λ. Its derivation, which we do not give, is likelihood-based, but the criterion resembles that of AIC in (8):

$$\lambda_{\text{LR}} = \arg\min_\lambda \left\{\ln \boldsymbol{y}^\top \left(\boldsymbol{I}_n - \boldsymbol{P}_\lambda\right)^2 \boldsymbol{y} - \frac{2}{n}\ln\left|\boldsymbol{I}_n - \boldsymbol{P}_\lambda\right|\right\}. \tag{19}$$

Tran also suggested a modified penalty term, which is dependent on $\boldsymbol{y}$, but this did not give appreciably different results from λ_{LR} in their study.

4 Maximization with Hyperpenalties

As noted previously, some existing methods may choose extreme values of λ, particularly when n is small, suggesting a need for a second level of shrinkage, that is, shrinkage of λ itself. We extend the hierarchical framework of (1) and (2) with a "hyperpenalty" on λ, $h(\lambda)$, which gives non-negligible support for λ over a finite range of values. The "hyperpenalized log-likelihood" is

$$hp\ell(\boldsymbol{\beta}, \lambda, \sigma^2) = \ell(\boldsymbol{\beta}, \sigma^2) - p_\lambda(\boldsymbol{\beta}, \sigma^2) - h(\lambda) - \ln(\sigma^2). \tag{20}$$

From the Bayesian perspective, when $h(\lambda)$ is in the form of a log-density, the hyperpenalty corresponds to a hyperprior on λ, and the hyperpenalized likelihood is the posterior (the expression $-\ln(\sigma^2)$ is the log-density of an improper prior on σ^2). In contrast to fully Bayesian methods, which characterize the entire posterior, we desire a single point estimate of $\boldsymbol{\beta}$, σ^2 and λ and focus on mode finding. Importantly, joint maximization with respect to $\boldsymbol{\beta}$, σ^2 and λ is now possible.

For a general $h(\lambda)$, we find the joint mode of (20): $\{\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma}^2\} \leftarrow \arg\max_{\boldsymbol{\beta},\lambda,\sigma^2}\{hp\ell(\boldsymbol{\beta}, \lambda, \sigma^2)\}$. Alternatively, $\{\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma}^2\}$ may be calculated using conditional maximization steps:

$$\begin{aligned} \sigma^{2^{(i)}} &\leftarrow \arg\max_{\sigma^2}\left\{hp\ell(\boldsymbol{\beta}^{(i-1)}, \sigma^2, \lambda^{(i-1)})\right\} \\ &= \frac{\big(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\big) + \lambda^{(i-1)}\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}}{n+p+2} \end{aligned} \tag{21}$$

$$\lambda^{(i)} \leftarrow \arg\max_{\lambda}\left\{hp\ell(\boldsymbol{\beta}^{(i-1)}, \sigma^{2^{(i)}}, \lambda)\right\} \tag{22}$$

$$\boldsymbol{\beta}^{(i)} \leftarrow \arg\max_{\boldsymbol{\beta}}\left\{hp\ell(\boldsymbol{\beta}, \sigma^{2^{(i)}}, \lambda^{(i)})\right\} = \boldsymbol{\beta}_{\lambda^{(i)}} \tag{23}$$

with $\{\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma}^2\} = \{\boldsymbol{\beta}^{(\infty)}, \lambda^{(\infty)}, \sigma^{2^{(\infty)}}\}$. The only step that depends on the choice $h(\lambda)$ is (22); the other steps are available in closed form regardless of $h(\lambda)$.

Based on the expression for $p_\lambda(\boldsymbol{\beta}, \sigma^2)$ given in (2), if $\exp\{-h(\lambda)\} = o(\lambda^{-p/2})$, then, upon applying the maximization step in (22), $\lambda^{(i)}$ is guaranteed to be finite, regardless of the values of $\boldsymbol{\beta}^{(i-1)}$ and $\sigma^{2(i)}$. This relates to an earlier comment in the Introduction that one cannot simply maximize $\ell(\boldsymbol{\beta}, \sigma^2) - p_\lambda(\boldsymbol{\beta}, \sigma^2)$ alone. For example, using $h(\lambda) = C$, C constant, would yield the same result as maximizing $\ell(\boldsymbol{\beta}, \sigma^2) - p_\lambda(\boldsymbol{\beta}, \sigma^2)$, namely an infinite hyperpenalized log-likelihood. We will propose one such choice of $h(\lambda)$ that satisfies $\exp\{-h(\lambda)\} = o(\lambda^{-p/2})$ and is empirically observed to work well for the ridge penalty; different choices of $h(\lambda)$ may be better suited for other penalty functions.

4.1 Choice of hyperpenalty

Crucial to this approach is the determination of an appropriate hyperpenalty and accompanying hyperparameters. Our recommended hyperpenalty is based on the gamma distribution, namely $h(\lambda) = -(a-1)\ln(\lambda) + \lambda/b$. From the Bayesian perspective, this is natural because it is conjugate to the precision of the Normal distribution, which is one possible interpretation of λ (e.g. Tipping, 2001; Armagan and Zaretzki, 2010). From Supplement S3, the update for λ given in (22) becomes

$$\lambda^{(i)} = \frac{p + 2a - 2}{\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}/\sigma^{2(i)} + 2b}. \tag{24}$$

This additionally requires choosing values for a and b. We will do so by first choosing a desired prior mean for λ, given by a/b, and a value for $\lambda^{(i)}$ in (24), and then solving the two expressions for a and b. Necessary to this strategy is that the chosen value of $\lambda^{(i)}$ must result in a and b that are free of $\sigma^{2(i)}$ and $\boldsymbol{\beta}^{(i-1)}$. To choose a/b, recall the key result from Hoerl and Kennard (1970): when $\boldsymbol{x}^\top\boldsymbol{x}/n = \boldsymbol{I}_p$, $\lambda^* = \arg\min_{\lambda\geq 0}\mathrm{E}[(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)] = p\sigma^2/\boldsymbol{\beta}^\top\boldsymbol{\beta}$. While not of immediate practical use, since σ^2 and $\boldsymbol{\beta}$ are the parameters to be estimated, we note that $\sigma^2/\boldsymbol{\beta}^\top\boldsymbol{\beta} \approx (1/R^2 - 1)$, where $R^2 = \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}/(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta} + \sigma^2)$ is the coefficient of determination, and the approximation comes from substituting $\boldsymbol{x}^\top\boldsymbol{x}/n = \boldsymbol{I}_p$ for $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. In contrast to σ^2 or the individual elements of $\boldsymbol{\beta}$, there may be knowledge about R^2. Alternatively, we will propose a strategy (Section 4.2) to estimate R^2. Given an estimate or prior guess of $R^2 \in (0,1)$, say $\hat{R}^2$, we set $a/b = p(1/\hat{R}^2 - 1)$.

In addition to a sensible mean, it is important to have a and b be such that $\lambda^{(i)}$, which is the resulting update for λ, is not extreme. Let the update for λ given in (24) be $\lambda^{(i)} = (p-1)H^{(i)}$, where $H^{(i)}$ is the harmonic mean of $\sigma^{2(i)}/\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}$ and $(1/\hat{R}^2 - 1)$. Being a harmonic mean, the $(1/\hat{R}^2 - 1)$ term, which will typically be less than 10 for most analyses, moderates potentially large values of $\sigma^{2(i)}/\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}$, thereby preventing underfitting. Simultaneously, $\lambda^{(i)}$ increases linearly with p, which prevents overfitting in $n < p$ scenarios.

Solving these expressions, $a/b = p(1/\hat{R}^2 - 1)$ and $\lambda^{(i)} = (p-1)H^{(i)}$, yields $a = p/2$ and $b = (1/\hat{R}^2 - 1)^{-1}/2$. When the covariates are approximately uncorrelated and $\hat{R}^2$ is close to R^2, a/b will be close to λ^*. However, the uncertainty coming from the variance of λ makes this useful in the general $\boldsymbol{x}^\top\boldsymbol{x}$ case, for which no closed-form solution of λ^* exists. As we will see in the simulation study, this holds true even when $\hat{R}^2$ is far from R^2.

Finally, it is important that the hyperpenalty strategy not be inferior in a standard regression, when $n \gg p$. To establish this, we derive in Supplement S4 a large-n approximation for λ^* in the general $\boldsymbol{x}^\top\boldsymbol{x}$ case: $\lambda^* \approx \sigma^2\text{Tr}\ [(\boldsymbol{x}^\top\boldsymbol{x})^{-1}]/\boldsymbol{\beta}^\top(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{\beta}$. As $n \to \infty$, this converges in probability to $\sigma^2\text{Tr}\ [\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}]/\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\beta} < \infty$, from which we can see that λ^* asymptotes in n. From (3), because the crossproduct term, $\boldsymbol{x}^\top\boldsymbol{x}$, increases linearly in n, $\boldsymbol{\beta}_{\lambda^*}$, the ridge estimate of $\boldsymbol{\beta}$ at the optimal value of λ, approaches the standard ordinary least squares (OLS) estimate, $\boldsymbol{\beta}_{\lambda|\lambda=0}$. The $\boldsymbol{x}^\top\boldsymbol{x}$ expression grows linearly in n, but λ^* asymptotes in n, so the effect of λ^* is reduced. The implication is that any choice of λ induces the same effective shrinkage for large n. The parameters a and b do not depend on n, and so the effect of the hyperpenalty will decrease with n, as desired.

Remark 3: Choosing a "flat" hyperpenalty, namely $h(\lambda) = \ln(\lambda)$, is untenable when $p > 2$, because it does not satisfy $\exp\{-h(\lambda)\} = o(\lambda^{-p/2})$. Specifically, plugging $h(\lambda) = \ln(\lambda)$ into (20), the expression $hp\ell(\boldsymbol{\beta}, \lambda, \sigma^2)$ can be made arbitrarily large in λ by setting $\boldsymbol{\beta} = \mathbf{0}_p$.

4.2 Estimating R^2

For analyses in which one is unable or unwilling to make a prior guess of R^2 to use as $\hat{R}^2$, here we describe a strategy to estimate R^2 from the data at hand. We note first that $R^2 = \text{Cor}(Y, \boldsymbol{X}^\top\boldsymbol{\beta})^2$, where 'Cor' denotes the Pearson correlation. It is known that the empirical prediction error from using a vector of fitted values, $\boldsymbol{x}\hat{\boldsymbol{\beta}}$, corresponding to the observed outcomes $\boldsymbol{y}$ will be optimistic when $\hat{\boldsymbol{\beta}}$ depends on $\boldsymbol{y}$. This means that the empirical R^2, $\bar{R}^2 = \hat{\text{Cor}}(\boldsymbol{y}, \boldsymbol{x}\hat{\boldsymbol{\beta}})^2$, will also be optimistic, i.e. upwardly biased, for R^2. In contrast, Efron (1983) showed how an estimate of prediction error using the bootstrap will be pessimistic. Applied to our context, given B bootstrapped datasets, the bootstrap-estimate of R^2, $\hat{R}^2_{\text{boot}} = (1/B)\sum_{b=1}^{B} \hat{\text{Cor}}(\boldsymbol{y}_{*(-b)}, \boldsymbol{x}_{*(-b)}\hat{\boldsymbol{\beta}}^{*(b)})^2$, where the $*(b)$ and $*(-b)$ notation indicate that the training and test datasets do not overlap, will be biased downward from R^2. Efron (1983) suggested that a particular linear combination of the optimistic and pessimistic prediction error estimates would provide an approximately unbiased estimate of prediction error. Analogizing this to estimating R^2, the linear combination is given by $0.632\hat{R}^2_{\text{boot}} + 0.368\bar{R}^2$. The weight is based on a bootstrapped dataset containing, on average, about $e^{-1}n \approx 0.632n$ unique observations from the original dataset. When $n > p$, this "632-estimate" of R^2, using, say, OLS to estimate $\boldsymbol{\beta}$ in each bootstrap, would provide a reasonable estimate of R^2 for our purposes. However, in the $n < p$ scenarios we are specifically interested in, OLS is not an option, and other methods, e.g. ridge regression, must be used to estimate $\boldsymbol{\beta}$ in each dataset. In addition, when p is large, the bootstrap may add a non-trivial computational component, if determining $\hat{\boldsymbol{\beta}}^{*(b)}$ is computationally expensive. So as to minimize any added burden due to estimating R^2, which is only a preprocessing step before applying the hyperpenalty approach, we propose to modify the 632-estimate by replacing the bootstrap with 5-CV:

$$\hat{R}^2_{632} = 0.632 \times (1/5)\sum_{i=1}^{5} \hat{\text{Cor}}(\boldsymbol{y}_{\kappa(i)}, \boldsymbol{x}_{\kappa(i)}\boldsymbol{\beta}_{\lambda_{\text{5-CV}}})^2 + 0.368 \times \hat{\text{Cor}}(\boldsymbol{y}, \boldsymbol{x}\boldsymbol{\beta}_{\lambda_{\text{5-CV}}})^2, \tag{25}$$

where we use the same $\kappa(i)$ notation defined in Section 2. To summarize: first calculate $\lambda_{\text{5-CV}}$. Then, calculate a weighted average of the cross-validated empirical correlation and the standard empirical correlation over all the data. Let HYP($\hat{R}^2_{632}$) denote the data-dependent hyperpenalty approach using $h(\lambda) = -(a-1)\ln(\lambda) + \lambda/b$, with $a = p/2$ and $b = (1/\hat{R}^2_{632} - 1)^{-1}/2$.

5 Simulation Study

Our simulation study is designed to mimic the current reality of the "-omics" era in which many covariates are analyzed but few contribute substantial effect sizes. The relevant quantities are described as follows:

Covariates ($\boldsymbol{x}$). One simulated dataset consists of training and validation data generated from the same model. The dimension of the training data is $n \times p$, with $n \in \{25, 100, 250, 4000\}$ and $p \in \{100, 4000\}$. The $n \times p$ matrix $\boldsymbol{x}$ is drawn from $N_p\{\mathbf{0}_p, \boldsymbol{\Sigma}_{\boldsymbol{X}}\}$. For the validation data, a $2000 \times p$ matrix $\boldsymbol{x}_{\text{new}}$ is sampled from this same distribution. We construct $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ according to "approximately uncorrelated" and "positively correlated" scenarios. The construction of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ is described in Supplement S5. Briefly: we begin with a block-wise compound symmetric matrix with 10 blocks. Within blocks, the correlation is $\rho = 0$ (approximately uncorrelated) or $\rho = 0.4$ (positively correlated), and between blocks, there is zero correlation. We then stochastically perturb the matrix in such a way to generate $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, using algorithms by Hardin et al. (2013), as to maintain positive-definiteness but mask the underlying structure. This perturbation occurs at each simulation iterate, and, as we outline in the Supplement, it is less extreme for the approximately uncorrelated case, so that the resulting $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ is close to $\boldsymbol{I}_p$.

Parameters ($\boldsymbol{\beta}$,σ^2). To better account for many plausible configurations of the coefficients, which would be difficult using a single, fixed choice of $\boldsymbol{\beta}$, we specify a generating distribution, drawing $\boldsymbol{\beta}$ once per simulation iterate and making it common to all observations. We draw $\boldsymbol{\beta}$ from a mixture density. Let $Z_i \in \{1, 2, 3\}$, $i = 1, \ldots, p$, be a random variable with

$\Pr(Z_i = 1) = \Pr(Z_i = 2) = 0.005$. Then, construct $\boldsymbol{\beta}$ as follows:

$$\alpha_i | Z_i \overset{ind}{\sim} \begin{cases} t_3\{\sigma^2 = 1/3\}, & Z_i = 1 \\ \text{Exp}\{1\}, & Z_i = 2 \\ N\{0, \sigma^2 = 10^{-6}\}, & Z_i = 3 \end{cases} \tag{26}$$

$$\boldsymbol{\beta} = \boldsymbol{\alpha} \times AR_1(\pi), \tag{27}$$

where $AR_1(\pi)$ is a $p \times p$, first-order auto-regressive matrix with correlation coefficient $\pi \in \{0, 0.3\}$. The sampling density for $\boldsymbol{\alpha}$ is a mixture of scaled t_3, Exponential, and Normal distributions, and 99% of the coefficients have small effect sizes, coming from the Normal component. The vector $\boldsymbol{\alpha}$ is scaled to obtain $\boldsymbol{\beta}$, which encourages neighboring coefficients to have similar effects, depending on π. So that some meaningful signal is present in every dataset, we ensure $\#\{i : Z_i \neq 3\} \geq 3$, regardless of p. Given $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, and $R^2 \in \{0.05, 0.2, 0.4, 0.6, 0.8, 0.95\}$, we calculate $\sigma^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta}(1/R^2 - 1)$.

Outcomes ($\boldsymbol{y}|\boldsymbol{x}$) The outcomes from the training and validation data are $\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta}, \sigma^2$ and $\boldsymbol{y}_{\text{new}}|\boldsymbol{x}_{\text{new}}, \boldsymbol{\beta}, \sigma^2$, respectively. For all 192 combinations of π, p, n, uncorrelated/correlated $\boldsymbol{x}$, and R^2, 10,000 (when $p = 100$) or 500 ($p = 4000$) training and validation datasets are sampled.

We compare the 11 methods listed in Table 1. n-CV is left out, being approximated by GCV and computationally expensive. AIC is replaced with its small-sample correction, AIC$_C$. The new methods considered are GCV$_C$ and the hyperpenalty-approach, HYP($\hat{R}^2_{632}$). We also present HYP(R^2), which is the hyperpenalty method using the true, unknown value of R^2. The difference in prediction error between HYP($\hat{R}^2_{632}$) and HYP(R^2) quantifies any possible gain from improving upon our strategy for estimating R^2. All methods differ only in λ, which determines the estimate of $\boldsymbol{\beta}$ via (3). The criterion by which we evaluate methods on

Table 1: Annotation of all methods from the simulation study and the corresponding Equations and References. The '⇒' indicates a new method.

Abbrev.	Name	Eqn.	Reference
5-CV	Five-fold cross-validation	(4)	Hastie et al. (Section 7.10, 2009)
BIC	Bayesian Information Criterion	(9)	Schwarz (1978)
AIC_C	Corrected Akaike's Information Criterion	(10)	Hurvich et al. (1998)
GCV	Generalized cross-validation	(7)	Craven and Wahba (1979)
⇒GCV_C	Corrected generalized cross-validation	(12)	Section 2.1
RGCV_γ	Robust generalized cross-validation	(11)	Lukas (2006)
MPML	Maximum profile marginal likelihood	(14)	Wecker and Ansley (1983)
GMPML	Generalized maximum profile marginal likelihood	(15)	Harville (1977); Wahba (1985)
MAPHL	Maximum adjusted profile h-likelihood	(16)-(18)	Lee and Nelder (1996)
LR	Loss-rank	(19)	Tran (2009)
⇒$\text{HYP}(\hat{R}^2_{632})$	Hyperpenalty, $\hat{R}^2$ based on "632" estimator	(21)–(25)	Section 4, Efron (1983)

the validation data is relative MSPE, rMSPE(λ):

$$\text{rMSPE}(\lambda) = 1000 \times (\text{MSPE}(\lambda)/\text{MSPE}(\lambda_{\text{opt}}) - 1), \tag{28}$$

where $\text{MSPE}(\lambda) \propto (\boldsymbol{y}_{\text{new}} - \boldsymbol{x}_{\text{new}}\boldsymbol{\beta}_\lambda)^\top(\boldsymbol{y}_{\text{new}} - \boldsymbol{x}_{\text{new}}\boldsymbol{\beta}_\lambda)$ and $\lambda_{\text{opt}} = \arg\min_\lambda \text{MSPE}(\lambda)$. Thus, rMSPE measures the percentage increase above the smallest possible MSPE, and rMSPE $= 0$ is ideal. Equivalently, rMSPE measures the inefficiency of each method. We used an iterative grid search to calculate λ_{opt} as well as λ for all methods except MAPHL and $\text{HYP}(\hat{R}^2_{632})$, for which explicit maximization steps are available.

We primarily focus on the subset of simulations for which $\pi = 0.3$ and $R^2 \in \{0.2, 0.4, 0.8\}$. Tables of the remaining rMSPEs are given in Supplement S6. Table 2 gives rMSPE; values in boldface are the column-wise minima, excluding $\text{HYP}(R^2)$, and those with an asterisk are less than twice each column-wise minimum. Figure 1 compares the rMSPE of $\text{HYP}(\hat{R}^2_{632})$ to the median rMSPE of remaining methods, excluding GCV_C, as an overall performance comparison. Figure 2 compares the rMSPE of GCV to GCV_C. Finally, Figure 3 gives histograms of $\ln(\lambda/\lambda_{\text{opt}})$ for each of the methods from one scenario.

From Table 2, the new methods, $\text{HYP}(\hat{R}^2_{632})$ and GCV_C, achieve the stated goal of being useful in $n \approx p$ or $n < p$ situations, as shown by the frequency of being in boldface or annotated with an asterisks. This is most evident in the smaller R^2 scenarios: either $\text{HYP}(\hat{R}^2_{632})$ or GCV_C

Table 2: Average rMSPE, defined in (28), for the 11 methods in Table 1 and HYP(R^2), which is the hyperpenalty approach using the true value of R^2. Values in **bold** are the column-wise minima, excluding HYP(R^2), and those with an '*' are less than twice the column-wise minimum. All settings here use $\pi = 0.3$ (27). The '⇒' indicates a new method.

***p* = 100, Approximately Uncorrelated**

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	64.1	23.4	*8.9	***0.8**	72.1	*25.9	*8.8	***0.8**	*73.8	*34.1	*10.4	***0.7**
BIC	369.7	$> 10^4$	103.9	31.6	249.4	$> 10^4$	253.5	16.5	*62.8	$> 10^4$	264.0	3.4
AIC$_C$	*22.0	18.5	*8.5	***0.8**	69.8	40.7	14.6	***0.8**	281.6	214.4	33.1	***0.7**
GCV	79.3	242.8	*7.8	***0.8**	82.5	267.2	*8.2	***0.8**	*63.4	434.8	*9.1	***0.7**
⇒GCV$_C$	*23.4	28.4	*7.8	***0.8**	37.2	*27.5	*8.2	***0.8**	*84.2	49.1	*9.1	***0.7**
RGCV$_{0.3}$	***19.1**	51.1	46.4	13.7	63.7	139.1	115.5	*1.0	276.5	579.5	109.5	***0.7**
MPML	311.9	18.5	***6.8**	***0.8**	234.3	*19.5	***6.7**	***0.8**	*63.7	*25.5	***7.6**	***0.7**
GMPML	57.4	18.0	***6.8**	***0.8**	68.5	***18.9**	***6.7**	***0.8**	*65.4	***24.0**	***7.6**	***0.7**
MAPHL	343.2	21.4	***6.8**	***0.8**	230.6	*22.8	*6.8	***0.8**	***56.2**	*33.0	***7.6**	***0.7**
LR	180.4	17.4	19.2	11.4	164.1	40.7	46.8	16.0	*73.9	176.3	182.0	21.6
⇒HYP($\hat{R}^2_{632}$)	*23.7	***8.2**	*8.0	***0.8**	***16.5**	*19.3	*12.0	***0.8**	*59.2	*46.6	*8.3	***0.7**
HYP(R^2)	*8.2*	*7.1*	*5.3*	*0.7*	*11.1*	*7.7*	*4.3*	*0.8*	*12.8*	*15.2*	*7.5*	*0.7*

***p* = 100, Positively Correlated**

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	96.9	*24.6	*7.3	***0.8**	103.7	*22.1	*7.2	*0.9	*95.1	*22.5	*7.8	***0.8**
BIC	600.5	$> 10^4$	62.5	15.0	447.8	$> 10^4$	73.4	24.4	170.9	$> 10^4$	138.1	17.7
AIC$_C$	*43.9	*15.1	*5.4	***0.8**	96.8	*18.2	*5.7	***0.8**	355.9	47.0	*13.6	***0.8**
GCV	121.7	109.8	*6.9	***0.8**	119.7	97.1	*6.6	***0.8**	*87.0	132.0	*7.4	***0.8**
⇒GCV$_C$	*57.7	27.9	*6.8	***0.8**	*63.5	*20.7	*6.5	***0.8**	***70.8**	*25.0	***7.3**	***0.8**
RGCV$_{0.3}$	*47.8	49.9	29.2	5.0	122.9	74.0	29.3	11.5	377.4	88.8	66.4	1.9
MPML	248.8	*15.4	*5.1	*1.0	268.8	*13.9	*6.6	*1.0	166.8	*27.4	*11.3	***0.8**
GMPML	*67.6	*15.3	***5.0**	*1.0	*79.5	***13.4**	*6.4	*1.0	*79.9	*25.3	*11.0	***0.8**
MAPHL	258.1	*16.5	*5.5	*1.1	218.6	*16.1	*7.4	*1.0	*118.2	*34.2	*12.1	***0.8**
LR	*52.0	*25.2	16.8	4.3	*82.2	40.6	24.5	7.6	171.7	81.6	58.6	15.5
⇒HYP($\hat{R}^2_{632}$)	***35.8**	***13.9**	*6.9	*1.4	***41.1**	*16.6	***5.5**	1.8	*87.1	***20.1**	18.1	*1.2
HYP(R^2)	*17.2*	*10.9*	*5.1*	*1.4*	*26.0*	*10.3*	*3.0*	*1.8*	*31.8*	*15.7*	*20.2*	*1.2*

***p* = 4000, Approximately Uncorrelated**

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	42.8	20.0	*6.9	***1.0**	47.5	*20.3	*7.5	***1.1**	40.5	*13.0	*6.2	*1.2
BIC	101.1	116.1	93.2	41.1	50.7	68.8	61.6	72.4	***6.9**	*13.3	15.3	370.5
AIC$_C$	34.4	21.4	10.5	2.1	102.1	50.6	22.5	8.0	323.1	163.5	78.0	86.1
GCV	43.3	18.4	*6.5	***1.0**	46.8	*18.7	*7.7	***1.1**	36.6	***12.4**	*6.5	*1.2
⇒GCV$_C$	*17.9	13.9	*5.5	***1.0**	42.4	***14.4**	***5.1**	***1.1**	110.3	*17.0	***5.4**	*1.2
RGCV$_{0.3}$	29.2	36.6	35.2	5.5	91.9	95.9	72.3	19.8	307.6	279.8	137.4	211.6
MPML	100.9	60.5	*5.9	***1.0**	50.7	62.4	11.3	*1.2	***6.9**	*13.3	15.3	***1.1**
GMPML	40.2	17.0	*5.7	***1.0**	47.0	*18.3	*5.6	*1.2	33.8	***12.4**	*7.0	***1.1**
MAPHL	100.8	115.9	93.0	6.7	50.5	68.6	61.4	7.3	***6.9**	*13.2	15.2	6.4
LR	100.7	13.7	12.3	6.3	50.7	*27.3	21.5	14.0	***6.9**	55.9	48.4	73.2
⇒HYP($\hat{R}^2_{632}$)	***9.4**	***6.5**	***4.9**	*1.2	***16.1**	*15.3	*6.7	2.4	74.5	34.6	11.1	7.1
HYP(R^2)	*16.7*	*12.9*	*6.9*	*1.0*	*29.6*	*12.8*	*4.8*	*1.5*	*20.9*	*6.2*	*2.3*	*12.3*

***p* = 4000, Positively Correlated**

Method/$\{n/R^2\}$	25/0.2	100/0.2	250/0.2	4000/0.2	25/0.4	100/0.4	250/0.4	4000/0.4	25/0.8	100/0.8	250/0.8	4000/0.8
5-CV	74.5	*17.5	*7.8	*0.8	79.3	*19.2	9.6	***0.8**	*71.0	*14.4	*7.4	***1.0**
BIC	333.4	98.4	85.7	10.8	253.1	74.4	73.8	15.7	*96.6	29.3	52.3	42.1
AIC$_C$	*46.0	*12.8	***4.8**	***0.6**	*74.6	*15.1	*6.3	***0.8**	235.8	49.8	16.9	5.3
GCV	81.8	*19.0	*8.6	*0.8	*74.9	*19.9	9.6	***0.8**	*71.2	*14.2	*7.4	***1.0**
⇒GCV$_C$	*49.7	*14.9	*6.1	*0.8	*48.4	*13.8	*6.4	***0.8**	*61.3	*10.9	***4.3**	***1.0**
RGCV$_{0.3}$	*52.7	33.0	18.5	*0.9	100.0	47.7	25.9	*1.4	248.4	77.6	24.3	12.7
MPML	248.2	*14.9	*5.3	*0.9	237.9	*19.2	*5.0	*1.4	*96.6	28.9	22.2	2.2
GMPML	*65.7	*14.3	*5.2	*0.9	*67.3	*14.4	***4.7**	*1.4	*71.7	*10.6	*6.8	2.2
MAPHL	332.5	98.1	85.0	7.3	252.3	74.2	73.1	8.1	*96.1	29.1	51.7	9.0
LR	82.7	*18.3	10.9	2.3	90.7	*26.1	17.4	3.1	126.3	64.0	31.2	8.6
⇒HYP($\hat{R}^2_{632}$)	***33.8**	***11.5**	*5.1	*0.7	***38.0**	***13.3**	*5.0	2.1	***56.8**	***10.3**	*4.9	32.5
HYP(R^2)	*17.0*	*5.8*	*3.3*	*0.7*	*14.5*	*7.2*	*3.6*	*2.0*	*31.0*	*7.7*	*3.9*	*35.4*

is frequently the best performing method when $R^2 = 0.2$ or $R^2 = 0.4$. However, these are not uniformly best across all scenarios considered, i.e. the "4000/0.8" column in the bottom sub-table or "25/0.8" column in the second-from-bottom sub-table. Some of this deficiency of $\textsc{hyp}(\hat{R}^2_{632})$ may be due to $\hat{R}^2_{632}$ being far from R^2. Figure S3 in Supplement S6 plots the empirical MSE of $\hat{R}^2_{632}$ and demonstrates that our strategy very accurately estimates R^2 except for the $n = 25$ scenarios. However, even $\textsc{hyp}(R^2)$, which uses the true value of R^2, has large rMSPE in these settings, which suggests that the optimal choice of λ is not well-approximated by the $p(1/R^2 - 1)$ expression discussed in Section 4.1. Also competitive are $\textsc{gmpml}$ and 5-$\textsc{cv}$; $\textsc{gcv}$ performs well except in the $n = 100$ scenarios, in which $n \approx p$. The remaining methods, $\textsc{bic}$, $\textsc{aic}_C$, $\textsc{rgcv}_{0.3}$, $\textsc{mpml}$, $\textsc{maphl}$, and $\textsc{lr}$ have large rMSPE in some scenarios.

To further explore this, Figure 1 plots the ratio of the rMSPE corresponding to $\textsc{hyp}(\hat{R}^2_{632})$ to the median rMSPE from the existing methods, to represent the performance of a typical method. When this ratio is less than one, $\textsc{hyp}(\hat{R}^2_{632})$ has smaller rMSPE. When $p = 100$, the $y = 1$ line is sometimes crossed when $\log_{10}(n) = 2.5$, or $n = 250$. When $p = 4000$, the $y = 1$ line is exceeded when $n = 4000$ or, regardless of n, when $R^2 = 0.8$ and the covariates are approximately uncorrelated. Based on the discussion in 4.1, $\textsc{hyp}(\hat{R}^2_{632})$ will become equivalent to the asymptotically optimal methods as $n \to \infty$.

As evidenced in the table, $\textsc{gcv}_C$ has markedly smaller rMSPE than $\textsc{gcv}$ when n is small. The two values of rMSPE coincide as n increases. We argued in Section 2.1 that the $\textsc{gcv}$ penalty has the potential to elicit undesirable behavior when $p = n - 1$, namely choosing $\lambda = 0$, or as close as possible thereto. Figure 2 gives the ramifications of this in terms of prediction error, plotting locally-smoothed values of rMSPE of $\textsc{gcv}$ and $\textsc{gcv}_C$ over many values of n for the $p = 100$ scenarios; $n - 1$ is less than, equal to, and greater than p. In all eight panels, which correspond to different R^2 and uncorrelated/correlated combinations, there is a peak in the $\textsc{gcv}$ curve beginning near $n - 1 = p$. $\textsc{gcv}_C$ effectively eliminates this

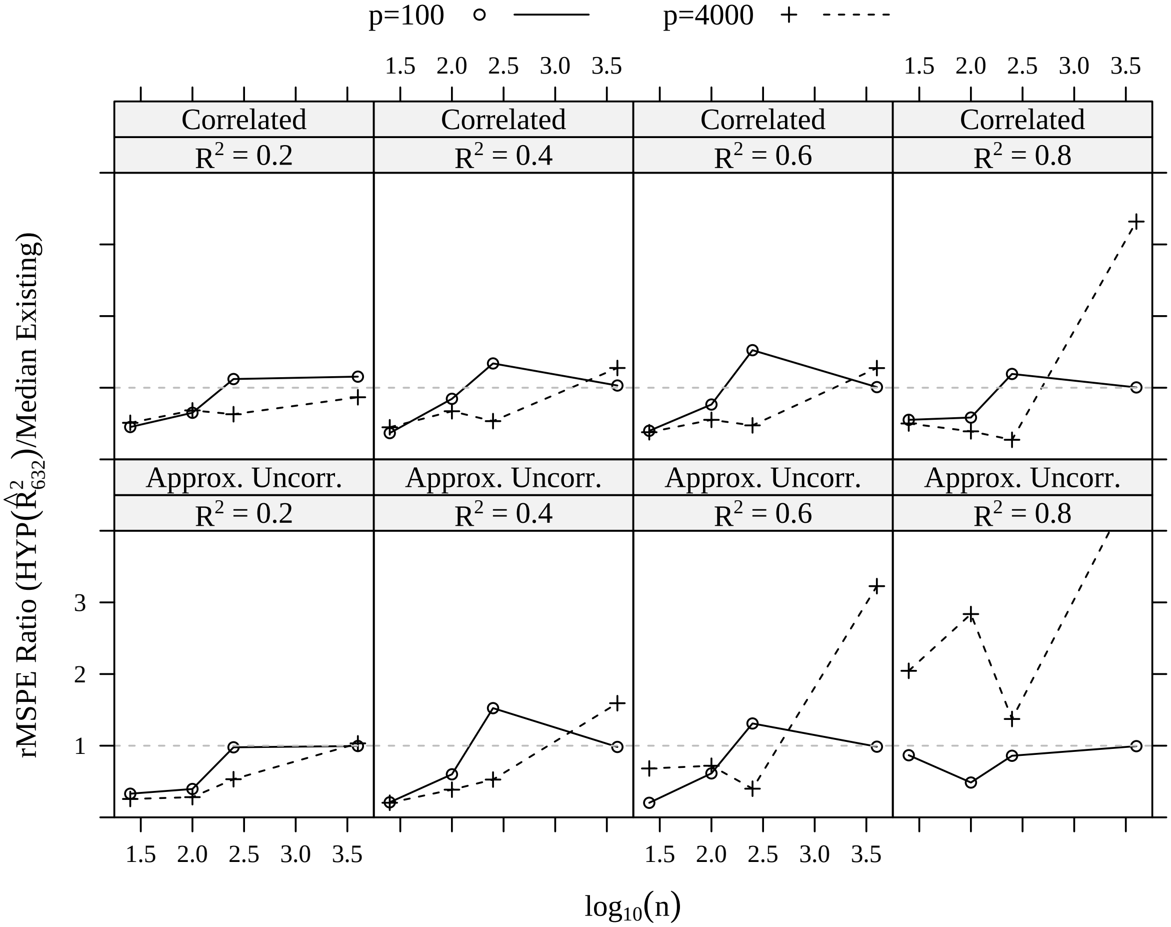

Figure 1: Ratio of the rMSPE of HYP($\hat{R}^2_{632}$) to the median rMSPE of remaining methods, excluding GCV$_C$, over $\log_{10}(n)$. Values less than one indicate that HYP($\hat{R}^2_{632}$) has smaller prediction error.

behavior and has almost uniformly smaller rMSPE when $n-1 \leq p$ and nearly equal rMSPE when $n-1 > p$.

Finally, we compare the choice of λ for each method. Figure 3 plots histograms of $\ln(\lambda/\lambda_{\text{opt}})$ for the eleven methods from one simulation setting in Table 2: $p = 4000$, $n = 100$, $\pi = 0.3$, $R^2 = 0.2$, and correlated covariates. When $\ln(\lambda/\lambda_{\text{opt}}) = 0$, the method has selected the optimal λ. In this small-n scenario, all of the existing methods, at times, choose a very small or large λ. In contrast, the shrinkage from hyperpenalization is evident: the histogram for HYP($\hat{R}^2_{632}$) has a considerably smaller range. It has the overall smallest rMSPE in this

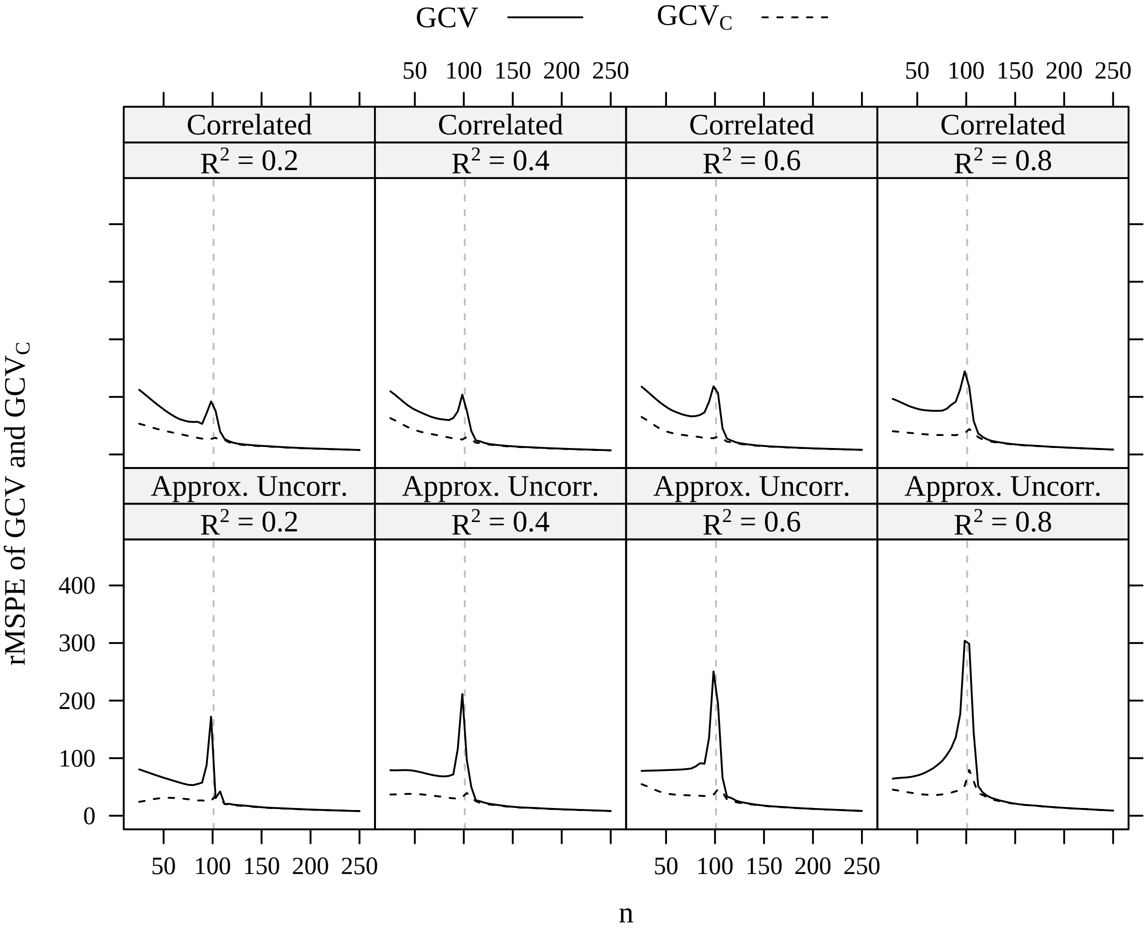

Figure 2: Locally-smoothed values of rMSPE of GCV and GCV$_C$ over n in scenarios for which $p = 100$. The curve corresponding to GCV changes in behavior near the point $n - 1 = p$, given by the vertical dashed line.

scenario (Table 2). Finally, GCV has $\ln(\lambda/\lambda_{\text{opt}})$ as small as -12, and GCV$_C$ has $\ln(\lambda/\lambda_{\text{opt}})$ slightly less than -2, providing additional evidence that GCV$_C$ prevents overfitting.

6 Bardet-Biedl Data Analysis

To evaluate these methods in a real dataset, we consider the rat gene-expression data first reported in Scheetz et al. (2006). Tissue from 120 12-week old rats was analyzed using microarrays (Affymetrix GeneChip Rat Genome 230 2.0 Array), normalized, and log-transformed.

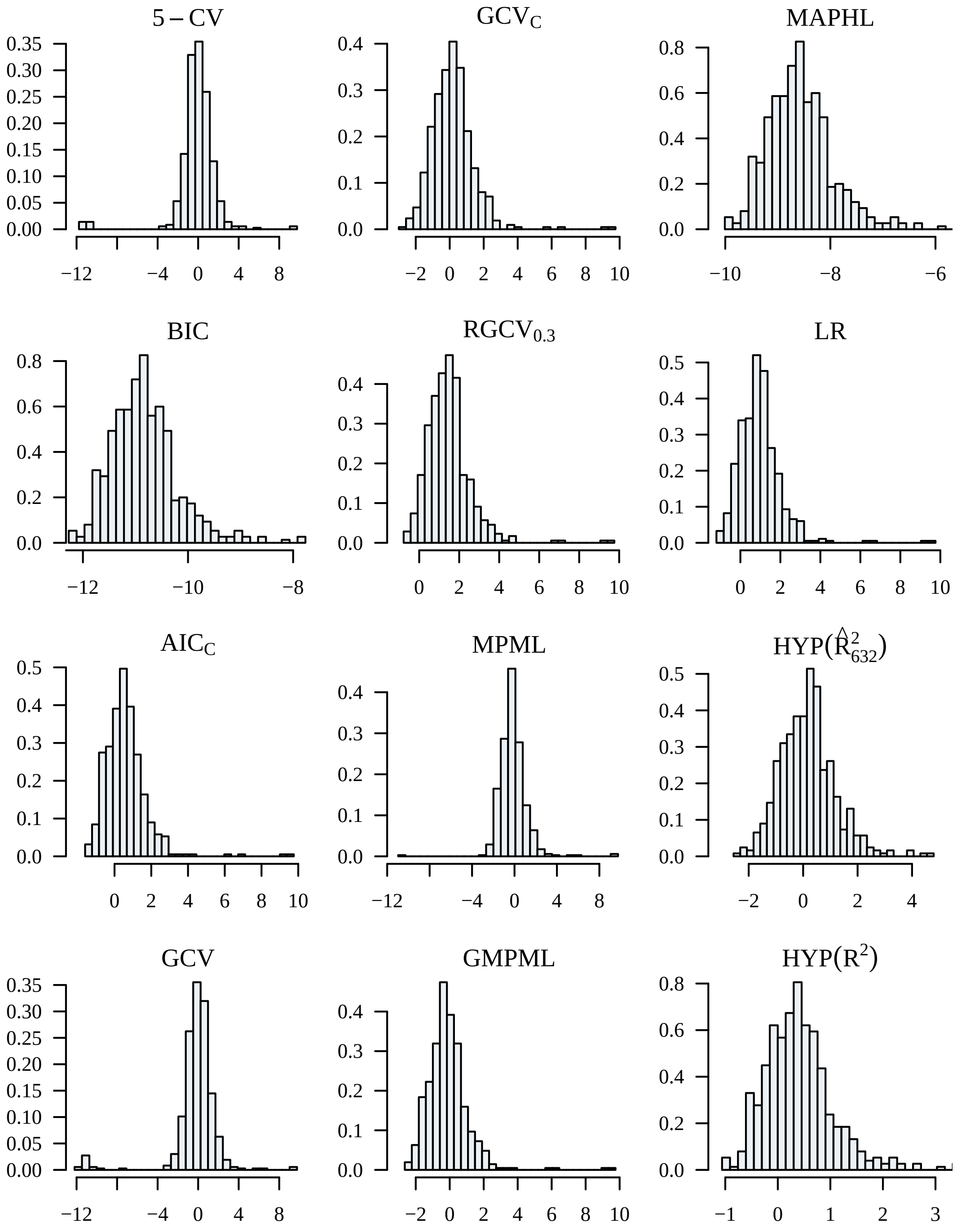

Figure 3: Histograms of $\ln(\lambda/\lambda_{\text{opt}})$ for $p = 4000$, $n = 100$, $\pi = 0.3$, $R^2 = 0.2$, and correlated covariates. $\ln(\lambda/\lambda_{\text{opt}}) = 0$ means that λ was chosen to yield optimal shrinkage. All methods are described in Table 1.

The goal is to find genes associated with expression of the *BBS11/TRIM32* gene, which is causative for Bardet-Biedl syndrome (Chiang et al., 2006). Following the strategy of Huang

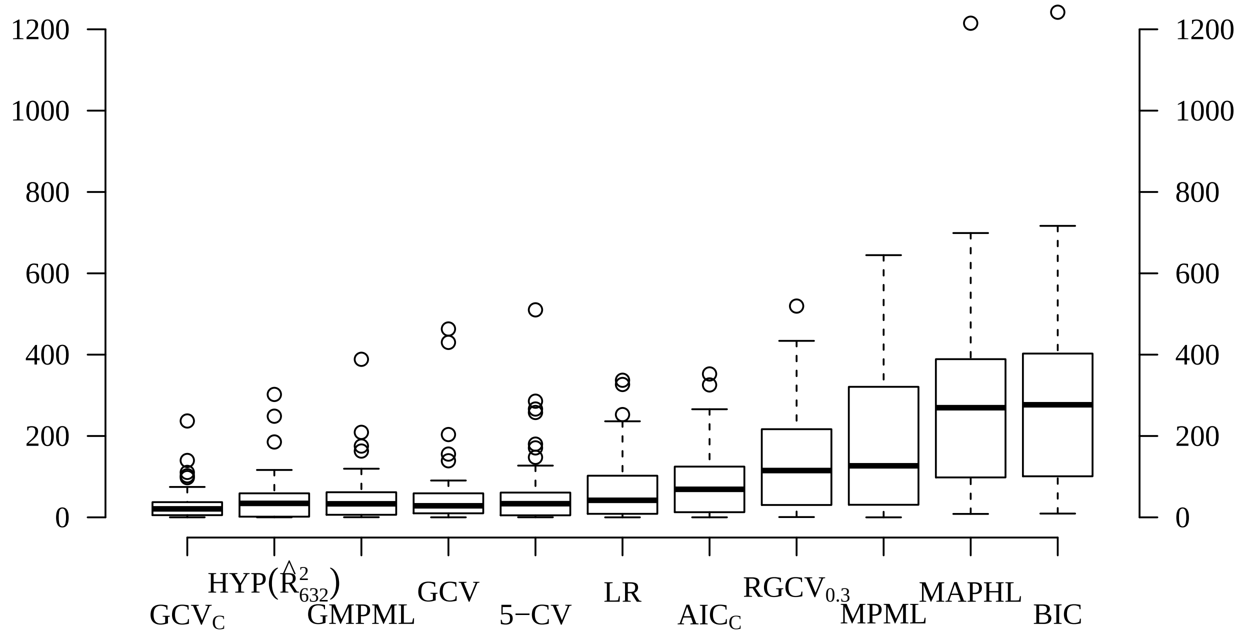

Figure 4: Boxplots of rMSPEs from 1,000 training/testing partitions of the Bardet-Biedl data, ordered from left to right in terms of average rMSPE.

et al. (2008), we considered 18,975 probesets that were sufficiently expressed, and, of these, reduced the number further to the $p = 3,000$ probes displaying the largest variation. We randomly selected $n = 80$ arrays as our training data, fit all methods to these data, and measured rMSPE based on the remaining 40 arrays, repeating this 1,000 times. Figure 4 gives the boxplots of these 1,000 rMSPEs, ordered from left to right in terms of the average rMSPE. The best-performing method is GCV_C with average rMSPE of 32.6, followed by $\text{HYP}(\hat{R}^2_{632})$ (47.0), GMPML (52.7), GCV (57.4), and 5-CV (64.5). These rankings correspond closely to those from the simulation study. The average empirical R^2 from the 40 left-out arrays for these top five methods ranged from 0.401–0.406, whereas the average value of $\hat{R}^2_{632}$ was 0.611, which suggests that $\hat{R}^2_{632}$ is upwardly biased of the "true" R^2. Despite this, the $\text{HYP}(\hat{R}^2_{632})$ method has small prediction error.

7 Discussion

We have examined strategies for choosing the ridge parameter λ when the sample size n is small relative to p. Our small-sample modification to GCV, called GCV_C, is conceptually trivial but uniformly dominates GCV in our simulation study. This corrected GCV may be applied in other shrinkage or smoothing situations that would otherwise use the standard GCV, such as smoothing splines (Wahba, 1985) or adaptively-weighted linear combinations of linear regression estimates (Boonstra et al., 2013).

We also proposed a novel approach using what we call hyperpenalties, which add another level of shrinkage, that of λ itself, by extending the hierarchical model. A hyperpenalty based on the Gamma density with mean $p(1/\hat{R}^2 - 1)$ was shown to work well in the context of ridge regression. The approach is based on the observation that the optimal tuning parameter λ is approximated by the expression $p(1/R^2 - 1)$, and we proposed a simple strategy for estimating the unknown R^2. Relative to existing methods, our implementation can offer superior prediction and protection against choosing extreme values of λ. One area for improvement of this approach lies in the high-R^2 scenarios, for which it is clear that $p(1/R^2 - 1)$ does not approximate the optimal tuning parameter. However, it is unusual in a high-dimensional regression to expect R^2 larger than 0.6 or 0.7.

Another advantage of the hyperpenalty approach is its applicability in missing data problems: when implementing the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), it is not clear how one might do ridge regression concurrently using goodness-of-fit or marginal likelihood approaches to select λ. On the other hand, by taking advantage of the conditional independence, specified by the hierarchical framework, between λ and any missing data given the remaining parameters, it is conceptually straightforward to embed a maximization step for λ, like expression (23), within a larger EM algorithm. This remains the focus of our current research.

Acknowledgments

This work was supported by the National Science Foundation [DMS1007494] and the National Institutes of Health [CA129102]. The authors thank an associate editor and two reviewers for their constructive and helpful comments and Jian Huang for generously sharing the Bardet-Biedl data. Code for the simulation study was written in `R` (R Core Team, 2012), using the `Matrix` package (Bates and Maechler, 2013) to construct block-diagonal matrices, and is available at `http://www-personal.umich.edu/~philb`.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pages 267–281.

Armagan, A. and Zaretzki, R. L. (2010). Model selection via adaptive shrinkage with t priors. *Computational Statistics*, 25:441–461.

Bates, D. and Maechler, M. (2013). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.0-12.

Boonstra, P. S., Taylor, J. M. G., and Mukherjee, B. (2013). Incorporating auxiliary information for improved prediction in high-dimensional datasets: an ensemble of shrinkage approaches. *Biostatistics*, 14:259–272.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer, New York, 2nd edition.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. (2006). Homozygosity mapping with snp arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences*, 103:6287–6292.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

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. (2001). Selection criteria for scatterplot smoothers. *Annals of Statistics*, 29:470–504.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

Fu, W. J. (1998). Penalized regressions: The bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7:397–416.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223.

Hardin, J., Garcia, S. R., Golan, D., et al. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, 7:1733–1762.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer, New York, 2nd edition.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the*

Royal Statistical Society: Series B, 60:271–293.

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

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B*, 58:619–678.

Li, K.-C. (1986). Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14:1101–1112.

Lukas, M. A. (2006). Robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems*, 22:1883–1902.

R Core Team (2012). *R: A Language and Environment for Statistical Computing.* Vienna, Austria.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103:14429–14434.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sin, C.-Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71:207–225.

Strawderman, R. L. and Wells, M. T. (2012). On hierarchical prior specifications and penalized likelihood. In Fourdrinier, D., Éric Marchand, and Rukhin, A. L., editors, *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, volume 8, pages 154–180. Institute of Mathematical Statistics.

Takada, Y. (1979). Stein's positive part estimator and Bayes estimator. *Annals of the Institute of Statistical Mathematics*, 31:177–183.

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

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244.

Tran, M. N. (2009). Penalized maximum likelihood for choosing ridge parameter. *Communications in Statistics*, 38:1610–1624.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, 13:1378–1402.

Wahba, G. and Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. *Statistics & Probability Letters*, 25:105–111.

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78:81–89.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301–320.