One-step estimator paths for concave regularization

Abstract

The statistics literature of the past 15 years has established many favorable properties for sparse diminishing-bias regularization: techniques which can roughly be understood as providing estimation under penalty functions spanning the range of concavity between L_0 and L_1 norms. However, lasso L_1-regularized estimation remains the standard tool for industrial 'Big Data' applications because of its minimal computational cost and the presence of easy-to-apply rules for penalty selection. In response, this article proposes a simple new algorithm framework that requires no more computation than a lasso path: the path of one-step estimators (POSE) does L_1 penalized regression estimation on a grid of decreasing penalties, but adapts coefficient-specific weights to decrease as a function of the coefficient estimated in the previous path step. This provides sparse diminishing-bias regularization at no extra cost over the fastest lasso algorithms. Moreover, our 'gamma lasso' implementation of POSE is accompanied by a reliable heuristic for the fit degrees of freedom, so that standard information criteria can be applied in penalty selection. The methods are illustrated in extensive simulations and in application of logistic regression to evaluating the performance of hockey players.

1 Introduction

For regression in high-dimensions, it is useful to regularize estimation through a penalty on coefficient size. L_1 regularization (i.e., the lasso of Tibshirani, 1996) is especially popular, with costs that are non-differentiable at their minima and can lead to coefficient solutions of exactly zero. A related approach is concave penalized regularization (e.g. SCAD from Fan and Li 2001 or MCP from Zhang 2010a) with cost functions that are also spiked at zero but flatten for large values (as opposed to the constant increase of an L_1 norm). This yields sparse solutions where large non-zero values are estimated with little bias.

The combination of *sparsity* and *diminishing-bias* is appealing in many settings, and a large literature on concave penalized estimation has developed over the past 15 years. For example, many authors (e.g., from Fan and Li 2001 and Fan and Peng 2004) have contributed work on their *oracle properties*, a class of results showing conditions under which coefficient estimates through concave penalization, or in related schemes, will be the same as if you knew the sparse 'truth' (either asymptotically or with high probability). From an information compression perspective, the increased sparsity encouraged by diminishing-bias penalties (since single large coefficients are allowed to account for the signals of other correlated covariates) leads to lower memory, storage, and communication requirements. Such savings are very important in distributed computing schemes (e.g., Taddy, 2015).

Unfortunately, exact solvers for concave penalized estimation all require significantly more compute time than a standard lasso. This has precluded their use in settings – e.g., text or web-data analysis – where both n (the number of observations) and p (covariate dimension) are very large. As we review in Section 3, recent literature recommends the use of approximate solvers. These approximations take the form of iteratively-weighted-L_1 regularization, where the coefficient-specific weights are based upon results from previous iterations of the approximate solver. Work on one-step estimation (OSE), e.g. by Zou and Li (2008), shows that even a single step of such weighted-L_1 regularization is enough to get solutions that are close to optimal, so long as the pre-estimates are *good enough* starting points. The crux of success is finding starts that are, indeed, good enough.

This article provides a complete framework for sparse diminishing-bias regularization that combines ideas from OSE with the concept of a *regularization path* – a general technique, most famously associated with the LARS algorithm (Efron et al., 2004), that estimates a sequence of models under decreasing amounts of regularization. So long as the estimates do not change too quickly along the path, such algorithms can be very fast to run and are an efficient way to obtain a high-quality *set* of models to choose amongst.

A path of one-step estimators (POSE; Algorithm 1) provides L_1 penalized regression on a grid of decreas-

Preliminary work. Under review by AISTATS 2016.
Do not distribute.

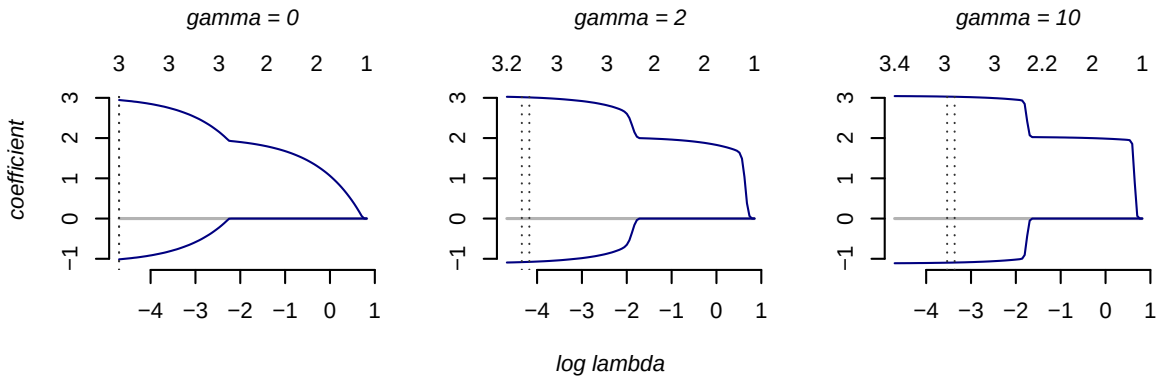

Figure 1: Gamma lasso estimation on $n = 10^3$ observations of $y_i = 4 + 3x_{1i} - x_{2i} + \varepsilon_i$, where $\varepsilon_i \overset{ind}{\sim} \mathrm{N}(0, 1)$ and $\{x_{1i}, x_{2i}, x_{3i}\}$ are marginally standard normal with correlation of 0.9 between covariates (x_{3i} is spurious). The penalty path has $T = 100$ segments, $\lambda^1 = n^{-1} |\sum_i x_{1i} y_i|$, and $\lambda^{100} = 0.01\lambda^1$. Degrees of freedom are on top and vertical lines mark AICc and BIC selected models (see Section 4).

ing penalties, but adapts coefficient-specific weights to decrease as a function of the coefficient estimated in the previous path step. POSE takes advantage of the natural match between path algorithms and one-step estimation: OSE relies upon inputs being close to the optimal solution, which is precisely the setting where path algorithms are most efficient. We formalize 'close' with a novel result in Theorem 3.1 that relates weighted-L_1 to L_0 regularization.

This framework allows us to provide

- a *path* of coefficient fits, each element of which corresponds to sparse diminishing-bias regularization estimation under a different level of penalization; where
- obtaining the path of coefficient fits requires no more computation than a state-of-the-art L_1 regularization path algorithm; and
- there are reliable closed-form rules for selection of the optimal penalty level along this path.

The last capability here is derived from a Bayesian interpretation for our *gamma lasso* implementation of POSE from which we are able to construct heuristic information criteria for penalty selection. We view such tools as an essential ingredient for practical applicability in large-scale industrial machine learning where, e.g., cross-validation is not always viable or advisable.

The remainder of this paper is outlined as follows. Section 2 presents the general regularized regression problem and introduces POSE, our path of one-step estimators algorithm, and the gamma lasso (GL), our implemented version of POSE. Section 3 gives an overview on the relationship between concave and weighted-L_1 regularization. Section 4 provides a Bayesian model interpretation for the gamma lasso, and derives from this model a set of information criteria that can be applied in penalty selection along the regularization path. Finally, we present two empirical studies: an extensive simulation experiment in Section 5, and in Section 6 we investigate the data analysis question: given all goals in the past decade of NHL hockey, what can we say about individual player contributions?

2 Paths of one-step estimators

Denote n response observations as $\mathbf{y} = [y_1, \ldots, y_n]'$ and the associated matrix of p covariates as $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]'$, with rows $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]'$ and columns $\boldsymbol{x}_j = [x_{1j}, \ldots, x_{nj}]'$.[1] Write $\eta_i = \alpha + \mathbf{x}_i'\boldsymbol{\beta}$ as the linear equation for observation i, and denote with $l(\alpha, \boldsymbol{\beta}) = l(\boldsymbol{\eta})$ the negative log likelihood. For example, in Gaussian (linear) regression, $l(\boldsymbol{\eta})$ is the sum-of-squares $0.5 \sum_i (y_i - \eta_i)^2$ and in binomial (logistic) regression, $l(\boldsymbol{\eta}) = -\sum_i [\eta_i y_i - \log(1 + e^{\eta_i})]$ for $y_i \in [0, 1]$. A penalized estimator is the solution to

$$\underset{\alpha, \beta_j \in \mathbb{R}}{\operatorname{argmin}} \left\{ l(\alpha, \boldsymbol{\beta}) + n\lambda \sum_{j=1}^{p} c(|\beta_j|) \right\}, \qquad (1)$$

where $\lambda > 0$ controls overall penalty magnitude and $c()$ is the coefficient cost function.

A few common cost functions are: L_2 $c(|\beta|) \propto \beta^2$ (ridge, Hoerl and Kennard, 1970), L_1 $c(|\beta|) \propto |\beta|$ (lasso, Tibshirani, 1996), the 'elastic net' mixture of L_1 and L_2 (Zou and Hastie, 2005), and the log penalty $c(|\beta|) \propto \log(1 + \gamma|\beta|)$ (Candes et al., 2008). Those that have a non-differentiable spike at zero (all but ridge) lead to sparse estimators, with some coefficients set to exactly zero. The curvature of the penalty away from zero dictates then the weight of shrinkage imposed on the nonzero coefficients: L_2 costs increase

[1] Since the size of penalized β_j depends upon the units of x_{ij}, it is common to scale the coefficient by $\mathrm{sd}(\boldsymbol{x}_j)$, the standard deviation of the j^{th} column of $\mathbf{X}$; this is achieved if x_{ij} is replaced by $x_{ij}/\mathrm{sd}(\boldsymbol{x}_j)$ throughout.

with coefficient size, lasso's L_1 penalty has zero curvature and imposes constant shrinkage, and as curvature goes towards $-\infty$ one approaches the L_0 penalty of subset selection. In this article we are primarily interested in *concave cost functions*, like the log penalty, which span the range between L_1 and L_0 penalties.

Penalty size, λ, acts as a *squelch*: it suppresses noise to focus on the true input signal. Large λ lead to very simple model estimates, while as $\lambda \to 0$ we approach maximum likelihood estimation (MLE). Since you don't know optimal λ, practical application of penalized estimation requires a *regularization path*: a $p \times T$ field of $\hat{\boldsymbol{\beta}}$ estimates, say $\hat{\boldsymbol{\beta}}|_\lambda$, obtained while moving from high to low penalization along $\lambda^1 > \lambda^2 \ldots > \lambda^T$. These paths begin at λ^1 set to infimum λ such that (1) is minimized at $\hat{\boldsymbol{\beta}}|_\lambda = \mathbf{0}$, and proceed down to some pre-specified λ^T (e.g., $\lambda^T = 0.01\lambda^1$).

Our path of one-step estimators (POSE) framework is in Algorithm 1. In this, we are assuming a penalty specification such that $\lim_{b\to 0} c'(|b|) = 1$ and that the cost function is differentiable for $b \neq 0$.

Algorithm 1 POSE

Initialize $\lambda^1 = \inf\left\{\lambda : \hat{\boldsymbol{\beta}}|_\lambda = \mathbf{0}\right\}$, so that $\hat{\boldsymbol{\beta}}_1 = \mathbf{0}$.

Set step size $0 < \delta < 1$.
for $t = 2 \ldots T$:

$$\lambda^t = \delta\lambda^{t-1}$$

$$\omega_j^t = \begin{cases} c'(|\hat{\beta}_j^{t-1}|) \text{ for } j \in \hat{S}_t \\ 1 \text{ for } j \in \hat{S}_t^c \end{cases} \quad (2)$$

$$\left[\hat{\alpha}, \hat{\boldsymbol{\beta}}\right]^t = \underset{\alpha,\beta_j\in\mathbb{R}}{\operatorname{argmin}} \; l(\alpha, \boldsymbol{\beta}) + n\sum_j \lambda^t \omega_j^t |\beta_j| \quad (3)$$

Section 3 will detail how POSE relates to concave regularization. However, for some quick intuition, consider POSE with a concave cost function (such as the log penalty in Figure 2). The derivative $c'(|\hat{\beta}|)$ will be positive but decreasing with larger values of $\hat{\beta}$, such that the *weight* on the L_1 penalty for $\hat{\beta}_j^t$ will *diminish* with the size of $|\hat{\beta}_j^t|$. This implies that coefficient estimates later in the path will be less biased towards zero if that coefficient has a large value earlier in the path.

2.1 The gamma lasso

The gamma lasso (GL) specification for POSE is based upon the log penalty,

$$c(\beta_j) = \log(1 + \gamma|\beta_j|), \quad (4)$$

where $\gamma > 0$. This penalty is concave with curvature $-1/(\gamma^{-1} + |\beta_j|)^2$ and it spans the range from L_0 ($\gamma \to \infty$) to L_1 ($\gamma \to 0$) costs. It appears under a variety of parameterizations and names in the literature; see Mazumder et al. (2011) and applications in Friedman (2008), Candes et al. (2008), Cevher (2009), Taddy (2013) and Armagan et al. (2013).

GL simply replaces line (2) in Algorithm 1 with

$$\omega_j^t = \left(1 + \gamma|\hat{\beta}_j^{t-1}|\right)^{-1} \quad j = 1 \ldots p \quad (5)$$

Behavior of the resulting paths is governed by γ, which we refer to as the penalty *scale*. Under $\gamma = 0$, GL is just the usual lasso. Bias diminishes faster for larger γ and, at the extreme, $\gamma = \infty$ yields a subset selection routine where a coefficient is unpenalized in all segments after it first becomes nonzero. Figure 1 shows solutions in a simple problem.

Each gamma lasso path segment is solved through coordinate descent (see supplement). The algorithm is implemented in `c` as part of the `gamlr` package for R. Usage of `gamlr` mirrors that of its convex penalty analogue `glmnet` (Friedman et al., 2010), the fantastic and widely used package for costs between L_1 and L_2 norms. In the lasso case ($\gamma = 0$), the two algorithms are essentially equivalent.

3 Weighted-L_1 approximations to concave penalization

Concave penalties such as the log penalty, which have a gradient that is decreasing with absolute coefficient size, yield the 'diminishing-bias' property discussed above. It is *the* reason why one would use concave penalization instead of L_1 or convex alternatives.

Unfortunately, such penalties can overwhelm the convex likelihood and produce a concave minimization objective; see Figure 2. This makes computation difficult. For example, one run of SCAD via the `ncvreg` R package (Breheny and Huang, 2011) for the simulation in Section 5 requires around 10 minutes, compared to 1-2 seconds for lasso (or gamma lasso). The most efficient exact solver that we've found is the `sparsenet` of Mazumder et al. (2011), also implemented in R, which first fits a lasso path and, for each segment on this path, adapts coefficient estimates along a second path of increasing penalty concavity. However, `sparsenet` relies upon solution over a large set of specifications[2] and its compute cost remains much higher than for the [gamma] lasso.

Local linear approximation (LLA; e.g., Candes et al., 2008) algorithms replace the concave cost function c

[2]POSE shares with `sparsenet` the idea of moving along a path of closely related specifications, but does not require a grid in both cost size and concavity. Intuitively, POSE runs a path diagonally through this grid.

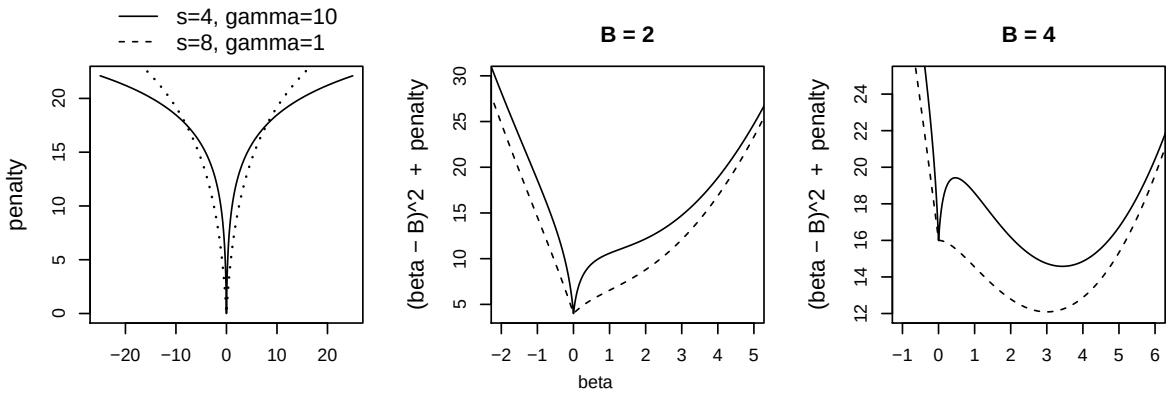

Figure 2: Log penalties $c(\beta) = s\log(1+\gamma|\beta|)$ and penalized objectives $(\beta - B)^2 + c(\beta)$.

with its tangent at the current estimate, $c'(\hat{\beta}_j)\beta_j$. The objective is then just a weighted L_1 penalized loss. An exact LLA solver iterates between updating $c'(\hat{\beta})$ and solving the implied L_1 penalized minimization problem. Zou and Li (2008) present numerical and theoretical evidence that LLA can provide near-optimal solutions even if you *stop it after one iteration*. This is an example of one-step estimation (OSE), a technique inspired by Bickel (1975) that amounts to taking as your estimator the first step of an iterative approximation to some objective. Early-stopping can be as good as the full solution *if* the initial estimates are good enough.

OSE and similar ideas have had a resurgence in the concave penalization literature recently, motivated by the need for faster estimation algorithms. Fan et al. (2014) consider early-stopping of LLA for folded concave penalization and show that, under strong sparsity assumptions about true $\boldsymbol{\beta}$ and given appropriate initial values, OSE LLA is with high probability an oracle estimator. Zhang (2010,2013) investigates 'convex relaxation' iterations, where estimates under convex regularization are the basis for weights in a subsequent penalized objective. Wang et al. (2013) propose a two step algorithm that feeds lasso coefficients into a linear approximation to folded concave penalization. These OSE methods are all closely related to the adaptive lasso (AL; Zou, 2006), which does weighted-L_1 minimization under weights $\omega_j = 1/|\hat{\beta}_j^0|$, where $\hat{\beta}_j^0$ is an initial guess at the coefficient value. The original AL paper advocates using MLE estimates for initial values, while Huang et al. (2008) suggest using marginal regression coefficients $\hat{\beta}_j^0 = \mathrm{cor}(\mathbf{x}_j, \mathbf{y})$.

The main point here is that OSE LLA, or a two-step estimator starting from $\hat{\boldsymbol{\beta}} = 0$, or any version of the adaptive lasso, are all interpretable as weighted-L_1 penalization with weights equal to something like $c'(\beta^0)$ for initial coefficient guess β^0. POSE and GL take advantage of an available source of initial values in any path estimation algorithm – the solved values from the previous path iteration. Our simulations in Section 5 show that this efficient strategy works as well or better than expensive exact solvers. In the next section, we provide some theoretical intuition on why it works.

3.1 Comparison between weighted-L_1 and L_0 penalized estimation

Our oracle benchmark is estimation under L_0 costs, $c(\beta_j) = \mathbb{1}_{\{\beta_j \neq 0\}}$, for which global solution is impractical. We are interested in weighted-L_1 penalization as a way to obtain fits that are as sparse as possible without compromising predictive ability, regardless of the underlying data generating process (or 'true' sparsity).

For $S \subset \{1 \dots p\}$ with cardinality $|S| = s$ and complement $S^c = \{1 \dots p\} \setminus S$, denote vectors restricted to covariates in S as $\beta_S = [\beta_j : j \in S]'$, matrices as $\mathbf{X}_S$, etc. Use $\boldsymbol{\beta}^S$ to denote the coefficients for ordinary least-squares (OLS) restricted to S: that is, $\boldsymbol{\beta}_S^S = (\mathbf{X}_S'\mathbf{X}_S)^{-1}\mathbf{X}_S'\mathbf{y}$ and $\beta_j^S = 0 \; \forall \; j \notin S$. Moreover, $\mathbf{e}^S = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^S = (\mathbf{I} - \mathbf{H}^S)\mathbf{y}$ are residuals and $\mathbf{H}^S = \mathbf{X}_S(\mathbf{X}_S'\mathbf{X}_S)^{-1}\mathbf{X}_S'$ the projection matrix from OLS on S. Use $|\cdot|$ and $\|\cdot\|$ for L_1 and L_2 norms.

We use the following result for iterative *stagewise* regression, with proof in the supplemental appendix.

Lemma 3.1. *Say* $\mathrm{MSE}_S = \|\mathbf{X}\boldsymbol{\beta}^S - \mathbf{y}\|^2/n$ *and* $\mathrm{cov}(\boldsymbol{x}_j, \mathbf{e}^S) = \boldsymbol{x}_j'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^S)/n$ *are sample variance and covariances. Then for any* $j \in 1 \dots p$,

$$\mathrm{cov}^2(\boldsymbol{x}_j, \mathbf{e}^S) \leq \mathrm{MSE}_S - \mathrm{MSE}_{S \cup j}$$

In addition, we need to define *restricted eigenvalues* (RE) on the gram matrix $\mathbf{X}'\mathbf{X}/n$.[3]

Definition 3.1. *The restricted eigenvalue is* $\phi^2(L, S) = \min_{\{\mathbf{v}:\mathbf{v}\neq\mathbf{0},\ |\mathbf{v}_{S^c}| \leq L\sqrt{s}\|\mathbf{v}_S\|\}} \frac{\|\mathbf{X}\mathbf{v}\|^2}{n\|\mathbf{v}\|^2}$.

Finally, we bound the distance between prediction rules from L_0 and weighted-L_1 penalized estimation.

[3] This RE matches the 'adaptive restricted eigenvalues' of Buhlmann and van de Geer (2011). Similar quantities are common in the theory of regularized estimators; see also Raskutti et al. (2010) and Bickel et al. (2009).

Theorem 3.1. *Consider squared-error loss* $l(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\|^2$*, and suppose* $\boldsymbol{\beta}^\nu$ *minimizes the* L_0 *penalized objective* $l(\boldsymbol{\beta})+n\nu\sum_{j=1}^p \mathbb{1}_{\{\beta_j\neq 0\}}$ *with* $\boldsymbol{\beta}^\nu = \boldsymbol{\beta}^S$ *and* $|S| = s < n$*. Write* $\hat{\boldsymbol{\beta}}$ *as solution to the weighted-*L_1 *minimization* $l(\boldsymbol{\beta}) + n\lambda\sum_j \omega_j|\beta_j|$*.*

Then $\omega_{S^c}^{\min}\lambda > \sqrt{2\nu}$ *while* $\phi^2(L,S) > 0$ *implies*

$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu)\|^2}{n} \leq \frac{4\lambda^2\|\boldsymbol{\omega}_S\|^2}{\phi^2(L,S)} \tag{6}$$

with $L = \frac{\|\boldsymbol{\omega}_S\|}{\sqrt{s}}\left(\omega_{S^c}^{\min} - \sqrt{2\nu}/\lambda\right)^{-1}$ *for the RE.*

Proof. From the definitions of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^\nu = \boldsymbol{\beta}^S$,

$$\begin{aligned}\frac{1}{2}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}\|^2 + n\lambda\sum_j \omega_j|\hat{\beta}_j| \qquad (7)\\ = \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu)\|^2}{2} + \frac{\|\mathbf{e}^S\|^2}{2} - \hat{\mathbf{y}}'\mathbf{e}^S + n\lambda\sum_j \omega_j|\hat{\beta}_j| \\ \leq \frac{1}{2}\|\mathbf{e}^S\|^2 + n\lambda\sum_{j\in S}\omega_j|\beta_j^S|\end{aligned}$$

Since $\hat{\mathbf{y}}'\mathbf{e}^S = \hat{\mathbf{y}}'(\mathbf{I}-\mathbf{H}^S)\mathbf{y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\nu) = \sum_{j\in S^c}\hat{\beta}_j\boldsymbol{x}_j'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\nu)$, we can apply Lemma 3.1 followed by $\boldsymbol{\beta}^\nu$ being optimal under L_0 penalty ν to get

$$\left(\frac{\boldsymbol{x}_j'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\nu)}{n}\right)^2 \leq \mathrm{MSE}_S - \mathrm{MSE}_{S\cup j} < 2\nu \quad \forall\, j \tag{8}$$

so that $|\hat{\mathbf{y}}'\mathbf{e}^S| = |\hat{\boldsymbol{\beta}}_{S^c}\mathbf{X}_{S^c}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\nu)| < n\sqrt{2\nu}|\hat{\boldsymbol{\beta}}_{S^c}|$. Applying this inside (7),

$$\begin{aligned}\frac{1}{2}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu)\|^2 + n\left(\omega_{S^c}^{\min}\lambda - \sqrt{2\nu}\right)|\hat{\boldsymbol{\beta}}_{S^c}| \qquad (9)\\ \leq\ n\lambda\sum_{j\in S}\omega_j|\hat{\beta}_j - \beta_j^\nu| \ \leq\ n\lambda\|\boldsymbol{\omega}_S\|\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^\nu\|.\end{aligned}$$

Given $\omega_{S^c}^{\min}\lambda > \sqrt{2\nu}$, difference $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu$ is in the RE support for $L = \frac{\|\boldsymbol{\omega}_S\|}{\sqrt{s}}(\omega_{S^c}^{\min} - \sqrt{2\nu}/\lambda)^{-1}$ and thus $\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^\nu\| \leq \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu)/\sqrt{n}\|/\phi(L,S)$. Finally, applying this inside (9) yields

$$\frac{1}{2}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu)\|^2 \ \leq\ \frac{\sqrt{n}\lambda\|\boldsymbol{\omega}_S\|\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu)\|}{\phi(L,S)}. \tag{10}$$

Dividing each side by $\sqrt{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\nu)\|/2$ and squaring gives the result. □

Remarks

- Theorem 3.1 is finite sample exact. Distinguishing it from related results in the literature, it is also completely *non-parametric* – it makes no reference to the true distribution of $\mathbf{y}|\mathbf{X}$. Indeed, if we make such assumptions, Theorem 3.1 provides bounds on the distance between a weighted-lasso and optimal prediction. The next remark is an example.

- Assume that $\mathbf{y} \sim (\boldsymbol{\eta}, \sigma^2\mathbf{I})$ – y_i independent with mean μ_i and shared variance σ^2. The C_p formula of Mallows (1973) gives $\mathrm{MSE}_S + 2s\sigma^2/n$ as an unbiased estimate of residual variance. Following Efron (2004), this implies $\nu = \sigma^2/n$ is the optimal L_0 penalty for minimizing prediction error. Theorem (3.1) applies directly, with $L = \frac{\|\boldsymbol{\omega}_S\|}{\sqrt{s}}\left(\omega_{S^c}^{\min} - \sqrt{2}\sigma/(\lambda\sqrt{n})\right)^{-1}$, to give a bound on the distance between weighted-L_1 estimation and C_p-optimal prediction. Note that, since the condition on minimum S^c weights has become $\omega_{S^c}^{\min} > (\sigma/\lambda)\sqrt{2/n}$, comparison to C_p suggests we can use larger γ with large n or small σ.

- Plausibility of the restricted eigenvalue assumption $\phi(L,S) > 0$ depends upon L. It is less restrictive if we can reduce $\|\boldsymbol{\omega}_S\|$ without making $\omega_{S_c}^{\min}$ small. *This is a key motivation for the POSE algorithm:* if covariates with nonzero $\hat{\beta}_j$ for large λ (i.e., early in the path) can be assumed to live in S, then increasing $\gamma > 0$ will improve prediction. Of course, the moment that λ becomes small enough that elements of S_c get nonzero $\hat{\beta}$, then larger γ can lead to over-fit. For this reason it is essential that we have tools for choosing optimal λ.

- In the supplemental material, we also adapt standard results from Wainwright (2006, 2009) to show how reducing $\|\boldsymbol{\omega}_S\|$ without making $\omega_{S_c}^{\min}$ small can lead to lower false discovery rates with respect to the L_0 oracle. However, such results depend upon design restrictions that are less realistic in practice.

4 Penalty selection

Lasso, the gamma lasso, and related sparse regularization estimators do not actually *do* model selection; rather, they can be used to obtain paths of estimates corresponding to different levels of penalization. Each penalty level corresponds to a different 'model' and we must select the optimal choice from these candidates.

K-fold cross-validation (CV; e.g., Efron, 2004) is the most common technique for penalty selection, and it does a good job. However, there are many scenarios where we might want an analytic alternative to CV. For example, if a single fit is expensive then doing it K times will be impractical. More subtly, truly Big Data are distributed: they are too large to store on a single machine. Algorithms can be designed to work in parallel on subsets (e.g., Taddy, 2015) but a bottleneck results if you need to communicate across machines for CV experimentation. Finally, CV can lead to over-fit for unstable algorithms whose results change dramatically in response to data jitter; see Breiman (1996) for a classic discussion. Such instability arises for large γ combined with small λ in our GL algorithm and, more generally, in many concave regularized estimators; see the supplemental material for a detailed overview.

An important feature of the standard L_1 lasso is that it comes with a simple approximation for the estimation degrees-of-freedom (*df*) at any λ: the number of nonzero estimated coefficients (see Zou et al., 2007). These *df* can then combined with the fitted deviance for penalty selection via common information criteria.

This section derives the gamma lasso as approximately maximizing the posterior for a hierarchical Bayesian model, and uses this interpretation to obtain a heuristic degrees-of-freedom for each estimate along the GL path. These GL *df* can be input to information criteria for model selection. In particular, our extensive simulations demonstrate that the GL *df* can be used with the corrected AICc of Hurvich and Tsai (1989) to obtain out-of-sample predictive performance that is as good or better than that from cross-validated predictors.

4.1 Bayesian model interpretation

Consider a model where each β_j is assigned a Laplace distribution prior with scale $\tau_j > 0$,

$$\beta_j \sim \mathrm{La}\left(\tau_j\right) = \frac{\tau_j}{2}\exp\left[-\tau_j|\beta_j|\,\right]. \tag{11}$$

Typically, scale parameters $\tau_1 = \ldots = \tau_p$ are set as a single value, say $n\lambda/\phi$ where ϕ is the dispersion (e.g. Gaussian variance σ^2 or 1 for the binomial). Posterior maximization under the prior in (11) is L_1 regularized estimation (e.g., Park and Casella, 2008).

Instead of a single shared scale, assume an independent gamma $\mathrm{Ga}(s, 1/\gamma)$ hyperprior with 'shape' s and 'scale' γ for each τ_j, such that $\mathbb{E}[\tau_j] = s\gamma$ and $\mathrm{var}(\tau_j) = s\gamma^2$. The *joint* coefficient-scale prior is

$$\begin{aligned}\pi(\beta_j, \tau_j) &= \mathrm{La}\left(\beta_j;\ \tau_j\right)\mathrm{Ga}\left(\tau_j;\ s, \gamma^{-1}\right) \\ &= \frac{1}{2\Gamma(s)}\left(\frac{\tau_j}{\gamma}\right)^s \exp\left[-\tau_j(\gamma^{-1}+|\beta_j|)\right].\end{aligned} \tag{12}$$

The gamma hyperprior is conjugate here, implying a $\mathrm{Ga}\left(s+1,\ 1/\gamma+|\beta_j|\right)$ posterior for $\tau_j \mid \beta_j$ with conditional posterior mode (MAP) at $\hat{\tau}_j = \gamma s/(1+\gamma|\beta_j|)$.

Consider joint MAP estimation of $[\boldsymbol{\tau}, \boldsymbol{\beta}]$ under the prior in (12), where we've suppressed α for simplicity. By taking negative logs and removing constants, this is equivalent to solving

$$\underset{\beta_j\in\mathbb{R},\ \tau_j\in\mathbb{R}^+}{\operatorname{argmin}} \frac{l(\boldsymbol{\beta})}{\phi} + \sum_j \left[\tau_j(\gamma^{-1}+|\beta_j|) - s\log(\tau_j)\right], \tag{13}$$

and is straightforward to show (supplemental) that this is equivalent to the log-penalized objective

$$\min_{\beta_j\in\mathbb{R}}\ \phi^{-1}l(\boldsymbol{\beta}) + \sum_j s\log(1+\gamma|\beta_j|). \tag{14}$$

4.2 Degrees of freedom

For prediction rules, say $\hat{y}_i$, that are suitably stable (i.e., Lipschitz; see Zou et al., 2007), the SURE framework of Stein (1981) applies and $df = \mathbb{E}\left[\sum_i \partial\hat{y}_i/\partial y_i\right]$. Consider a single coefficient β estimated via least-squares under L_1 penalty τ. Write gradient at zero $g = -\sum_i x_i y_i$ and curvature $h = \sum_i x_i^2$ and set $\varsigma = -\mathtt{sign}(g)$. The prediction rule is $\hat{y} = x(\varsigma/h)(|g|-\tau)_+$ with derivative $\partial\hat{y}_i/\partial y = x_i^2/h\mathbb{1}_{[|g|<\tau]}$, so that the SURE expression yields $df = \mathbb{E}\left[\mathbb{1}_{[|g|<\tau]}\right]$. This expectation is taken with respect to the *unknown true* distribution over $\mathbf{y}|\mathbf{X}$, not that estimated from the observed sample. However, one can evaluate this expression at observed gradients as an unbiased estimator for the true *df* (e.g., Zou et al., 2007).

This motivates our heuristic *df* in weighted-L_1 regularization: the *prior* expectation for the number L_1 penalty dimensions, $\tau_j = \lambda\omega_j$, that are less than their corresponding absolute gradient dimension. Referring to our Bayesian model above, each τ_j is $iid\ \mathrm{Ga}(s, 1/\gamma)$ in the prior, leading to the GL degrees of freedom

$$df^t = \sum_j \mathrm{Ga}(|g_j|;\ n\lambda^t/(\gamma\phi), 1/\gamma), \tag{15}$$

where $\mathrm{Ga}(\ \cdot\ ;\ \mathrm{shape}, 1/\mathrm{scale})$ is the Gamma cumulative distribution function and g_j is an estimate of the j^{th} coefficient gradient evaluated at $\hat{\beta}_j = 0$.[4]

For orthogonal covariates, g_j is just the marginal gradient at zero. In the non-orthogonal case, where $g_j = g_j(0)$ becomes a function of all of the elements of $\hat{\boldsymbol{\beta}}$, we plug in the most recent g_j at which $\hat{\beta}_j^t = 0$: this requires no extra computation and has the advantage of maintaining $df = \hat{p}^t$ for $\gamma = 0$.

4.3 Selection via information criteria

An information criterion is an attempt to approximate divergence between the unknown true data generating process and our parametric approximation; see the supplement for an overview. These take the form

$$l(\hat{\boldsymbol{\beta}}) + k(df) \tag{16}$$

where k is the cost on the degrees-of-freedom associated with $\hat{\boldsymbol{\beta}}$ and l is the negative log likelihood. The AIC of Akaike (1973) uses $k(df) = df$ while the BIC of Schwarz (1978) uses $k(df) = \log(n)df/2$.

As detailed in Flynn et al. (2013), the corrected AICc with $k(df) = df \times n/(n-df-1)$ does a better job than the AIC or BIC in choosing the optimal model for prediction when *df* is large. Alternatively, the BIC is often preferred for accurate support recovery or avoiding false discovery; see, e.g., Zou et al. (2007).

[4] Note that the number of unpenalized variables (e.g., 1 for α) should also be added to the total estimation *df*.

		% Worse than oracle C_p				
		AICc			CV	
$\frac{\mathrm{sd}(\eta)}{\sigma}$	d	GL0	GL2	GL10	MCP	C_p R^2
2	10	2.5	**1.3**	3.8	**1.3**	.79
2	50	7.8	**3.9**	6.5	**3.9**	.77
2	100	12.0	8.0	8.0	**6.7**	.75
1	10	8.3	6.2	14.6	**4.2**	.48
1	50	27.3	**16.0**	18.2	18.2	.44
1	100	35.0	**26.8**	**26.8**	**26.8**	.41
1/2	10	33.3	**27.8**	61.1	**27.8**	.18
1/2	50	**71.4**	92.9	100.0	**71.4**	.14
1/2	100	**80.0**	110.0	110.0	90.0	.10

5 Simulation experiment

We consider samples of size $n = 1000$ from

$$y \sim \mathrm{N}\left(\mathbf{x}'\boldsymbol{\beta}, \sigma^2\right) \text{ where } \beta_j = \frac{\exp\left(-\frac{j}{\mathsf{d}}\right)}{j},\ j = 1 \dots p,$$
$$\text{and } \mathbf{x} = \mathbf{u} * \mathbf{z},\ \ \mathbf{u} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right),\ \ z_j \overset{ind}{\sim} \mathrm{Bin}(0.5). \tag{17}$$

Each simulation draws means $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, and two independent response samples $\mathbf{y}, \tilde{\mathbf{y}} \sim \mathrm{N}(\boldsymbol{\eta}, \sigma^2\mathbf{I})$. Multicollinearity is parametrized via $\Sigma_{jk} = \rho^{|j-k|}$ with $\rho = 0.5$. We define σ^2 through *signal-to-noise* ratios $\mathrm{sd}(\boldsymbol{\eta})/\sigma$ of $1/2$, 1, and 2; for coefficient decay rates we consider d of 10, 50, and 100.[5]

The regression in (17) is obviously *dense*: true coefficients are all nonzero. However, they decay in magnitude along the index j and it will be useless to estimate many of them in a $p = n$ regression. Our sparse oracle comparator is the C_p optimal L_0 penalized solution

$$\boldsymbol{\beta}^\star = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\sigma^2 \sum_j \mathbb{1}_{\{\beta_j \neq 0\}} \right\}, \tag{18}$$

which is solvable here by searching through OLS regression on $\mathbf{X}_{\{1\dots j\}}$ for $j = 1 \dots p$.

We consider `gamlr` runs of GL with γ of 0 (lasso), 2, and 10 and marginal AL, as well as `sparsenet`'s MCP penalized regression. Covariate penalties are standardized by $\mathrm{sd}(\boldsymbol{x}_j)$, and paths run through 100 λ values down to $\lambda^{100} = 0.01\lambda^1$. On an Intel Xeon E5-2670 core, 5-fold CV with $\mathrm{sd}(\eta)/\sigma = 1$ and $\rho = 1/2$ requires 1-2 seconds for lasso and marginal AL, 2 and 3 seconds for GL with $\gamma = 2$ and $\gamma = 10$, and 15-20 seconds for Sparsenet. [6]

[5] Results for many additional configurations, including different covariate sparsity and multicollinearity, are available in the supplement.

[6] `ncvreg` SCAD required ten minutes for a single run and is thus impractical for the applications under consideration. But, in a small study, CV.min selected SCAD performs quite well in prediction – similarly to the best CV.min methods for each data configuration – with relatively high values for both false discovery and sensitivity.

Figures 3 and 4 illustrate GL paths for a single dataset, with $\mathrm{sd}(\eta)/\sigma = 1$ and $\rho = 1/2$. In Figure 3, increasing γ leads to 'larger shouldered' paths where estimates move quickly to MLE for the nonzero-coefficients. Degrees of freedom, calculated as in (15), are along the top of each plot; equal λ have higher df^t for higher γ since there is less shrinkage of $\hat{\beta}_j \neq 0$. Figure 4 shows CV and AICc error estimates. The two criteria roughly track each other, although AICc more heavily penalizes over-fit and at $\gamma = 10$ their minima do not match. Notice that as γ increases, the CV error increases more quickly after it's minimum; this indicates that the consequences of over-fit are worse under faster diminishing-bias.

6 Hockey example

This section attempts to quantify the performance of hockey players. It extends analysis in Gramacy et al. (2013). The current version includes data about who was on the ice for every goal in the National Hockey League (NHL) back to the 2002-2003 season, including playoffs. The data are in the `gamlr` package; there are 69449 goals and 2439 players.

The logistic regression model of player contribution is, for goal i in season s with away team a and home team h,

$$\text{logit}\left[\text{p(home team scored goal } i)\right] \tag{19}$$
$$= \alpha_0 + \alpha_{sh} - \alpha_{sa} + \mathbf{u}_i'\boldsymbol{\phi} + \mathbf{x}_i'\boldsymbol{\beta}_0 + \mathbf{x}_i'\boldsymbol{\beta}_s, \tag{20}$$

Vector $\mathbf{u}_i$ holds indicators for various special-teams scenarios (e.g., a home team power play), and $\boldsymbol{\alpha}$ provides matchup/season specific intercepts. Vector $\mathbf{x}_i$ contains player effects: $x_{ij} = 1$ if player j was on the home team and on ice for goal i, $x_{ij} = -1$ for away player j on ice for goal i, and $x_{ij} = 0$ for everyone not on the ice. Coefficient $\beta_{0j} + \beta_{sj}$ is the season-s effect of player j on the log odds that, given a goal has been scored, the goal was scored by their team. These effects are 'partial' in that they control for who else was on the ice, special teams scenarios, and team-season fixed effects – a player's β_{0j} or β_{sj} only need be nonzero if that player effects play above or below the team average for a given season.

We estimate gamma lasso paths of $\boldsymbol{\beta}$ for the model in (19) *with $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ left unpenalized.* In contrast to the default for most analyses, our coefficient costs are *not* scaled by covariate standard deviation. Doing the usual standardization would have favored players with little ice time. The algorithm is run for $\log_{10}\gamma = -5 \dots 2$, plus the $\gamma = 0$ lasso.[7]

Joint $[\gamma, \lambda]$ surfaces for AICc and BIC are in Figure 5.

[7] On this data, $\gamma = \infty$ subset selection yields perfect separation and infinite likelihood.

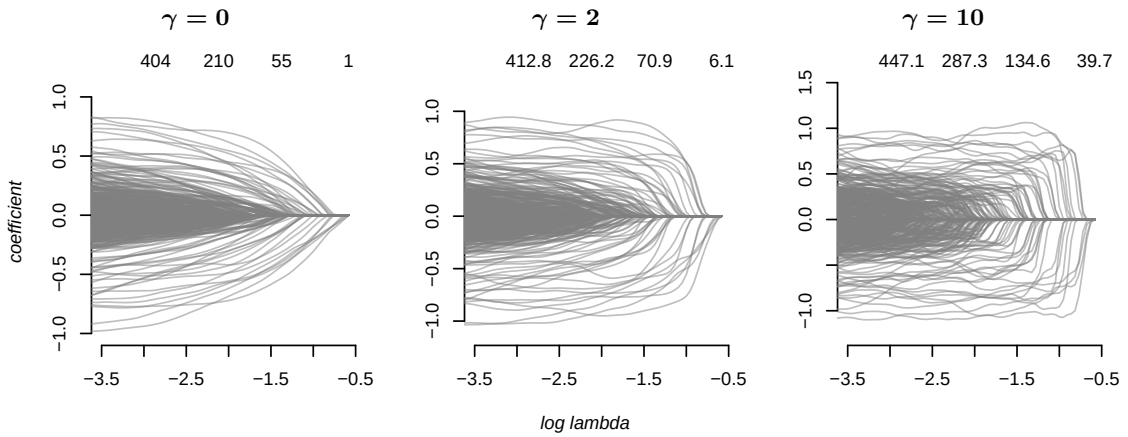

Figure 3: Regularization paths for simulation example. Degrees of freedom df^t are along the top.

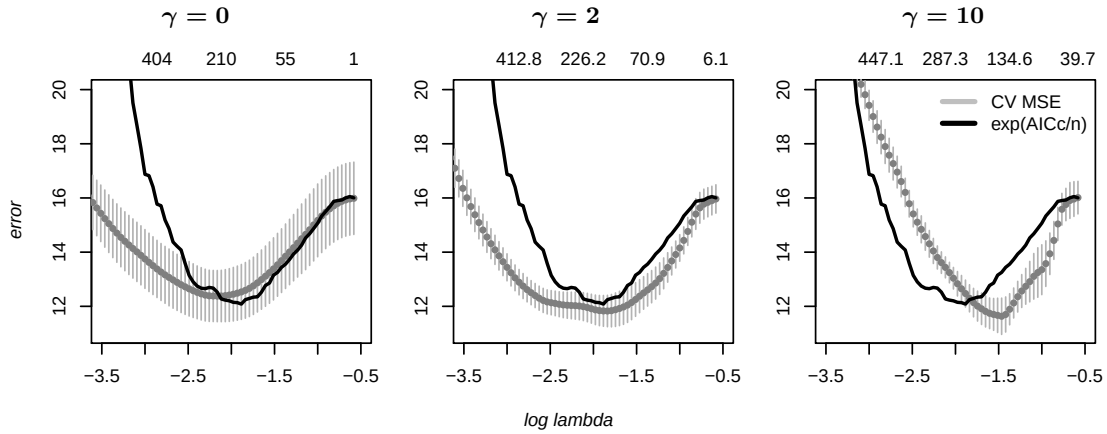

Figure 4: 5-fold CV and AICc for a simulation example. Points-and-bars show mean OOS MSE $\pm$1se.

AICc favors denser models with low λ but not-to-big γ, while the BIC prefers very sparse but relatively unbiased models with large λ and small γ. Both criteria are strongly adverse to any model at $\gamma = 100$, which is where timings explode in Figure **??**. Ten-fold CV results are shown in Figure 6 for γ of 0, 1, and 10. The OOS error minima are around the same in each case – average deviance slightly above 1.16 – but errors increase much faster away from optimality with larger γ. We also see that AICc selection is always between the CV.min and CV.1se selctions: at $\gamma = 0$ AICc matches the CV.1se choice, while at $\gamma = 10$ it has moved right to the CV.1se selection. Our heuristic from Section **??**.2 might be over-estimating *df* for large-γ models (especially under this very collinear design), but one would also suspect that CV estimates of minimum deviance are biased downward more dramatically for larger γ than for low-variance small-γ estimators.

The motivating application for this example was to devise a better versions of hockey's 'plus-minus' (PM) statistic: number of goals *for* minus *against* each player's team while he is on the ice. To convert from player effects $\beta_{0j} + \beta_{sj}$ to the scale of 'plus/minus', we first state that in absence of any other relevant information about each goal (not team, home/away, etc), the probability a goal was scored by his team given player j is on ice becomes $p_j = e^{\beta_j}/(1 + e^{\beta_j})$ and our 'partial plus/minus' (PPM) is

$$\text{ppm}_j = N_j(p_j - (1 - p_j)) = N_j(2p_j - 1)$$

where N_j is the number of goals for which he was on-ice. This measures quality and quantity of contribution, controlling for confounding information, and lives on the same scale as PM.

Table 1 contains the estimated PPM values for the 2013-2014 season under various γ levels, using AICc selection. We see that, even if changing concavity (γ) has little effect on minimum CV errors, larger γ yield more sparse models and different conclusions about player contribution. At the $\gamma = 0$ lasso, there are 305 nonzero player effects (individuals measurably different from their team's average ability) and the list includes young players who have had very strong starts to their careers. For example, Ondrej Palat and Tyler Toffoli both played their first full seasons in the NHL in 2013-2014. As γ increases to 1, these young guys drop in rank while more proven stars (e.g., Sidney

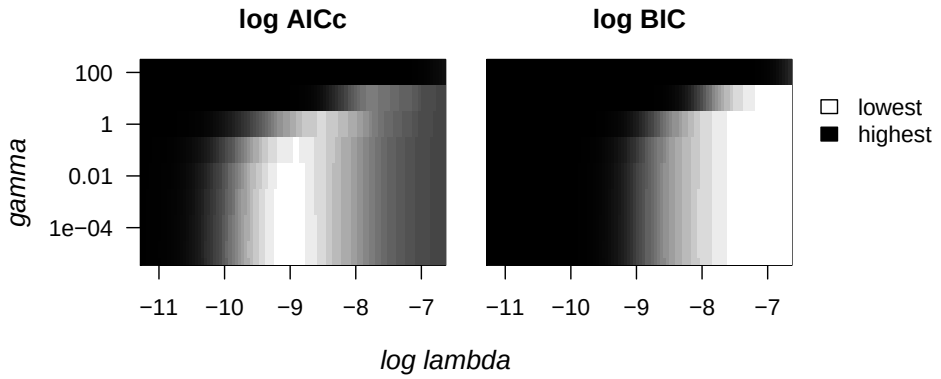

Figure 5: Hockey example AICc and BIC surfaces, rising from white to black on log scale.

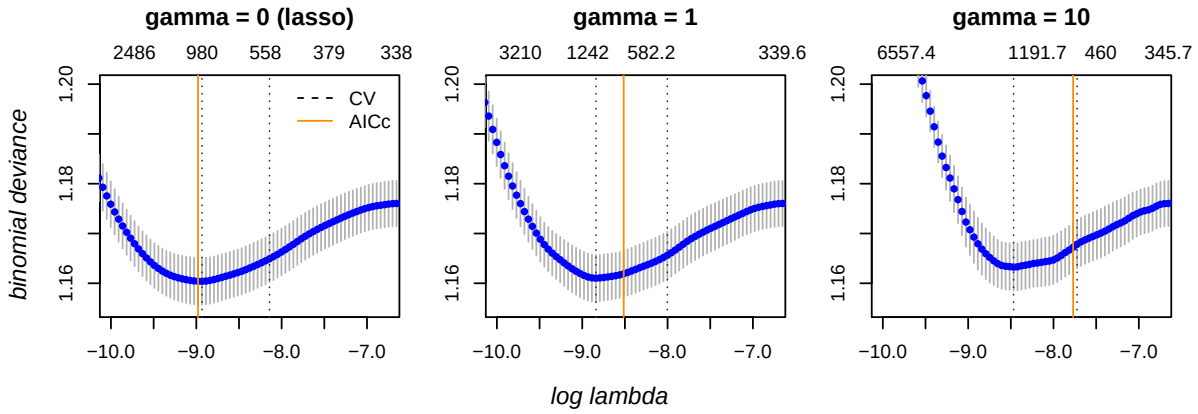

Figure 6: Hockey example 10-fold CV: mean OOS deviance $\pm$1se, with minimum-error and 1SE selection rules marked with black dotted lines, and solid orange line showing AICc selection.

Crosby and Jonathan Toews) move up the list. Finally, at $\gamma = 10$ only big-name stars remain amongst the 64 nonzero player effects.

7 Discussion

Concave penalized estimation in Big data, where exact solvers are too computationally expensive, reduces largely to weighted-L_1 penalization. This review has covered a number of topics that we think relevant for such schemes. Apart from the simulation study, we have not provided extensive comparison of the many available weighting mechanisms. However, we feel that path adaptation is an intuitively reasonable source of weights. In any case, POSE has the advantage that using a regularization path to supply penalty weights is computationally efficient. To scale for truly Big data, L_1 weights need be constructed at practically no cost on top of a standard lasso run. Beyond `gamlr` and marginal regression adaptive lasso, we have found no other software for sparse diminishing bias estimation where this standard is met.

References

Akaike, H. (1973). Information theory and the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest.

Armagan, A., D. B. Dunson, and J. Lee (2013). Generalized double pareto shrinkage. *Statistica Sinica 23*, 119.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association 70*, 428–434.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Breheny, P. and J. Huang (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics 5*(1), 232–253.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics 24*(6), 2350–2383.

Buhlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data*. Springer.

	lasso			$\gamma = 1$			$\gamma = 10$		
		PPM	PM		PPM	PM		PPM	PM
1	Ondrej Palat	33.8	38	Sidney Crosby	29.2	52	Sidney Crosby	32.6	52
2	Sidney Crosby	31.2	52	Ondrej Palat	29	38	Jonathan Toews	22.8	35
3	Henrik Lundqvist	25.8	9	Jonathan Toews	21.4	35	Joe Thornton	22	34
4	Jonathan Toews	24	35	Joe Thornton	21	34	Anze Kopitar	22	39
5	Andrei Markov	23.1	34	Andrei Markov	20.9	34	Andrei Markov	20.7	34
6	Joe Thornton	21.4	34	Henrik Lundqvist	19.8	9	Alex Ovechkin	18.1	16
7	Anze Kopitar	20.6	39	Anze Kopitar	19.5	39	Pavel Datsyuk	16.6	13
8	Tyler Toffoli	18.9	31	Pavel Datsyuk	16.1	13	Ryan Getzlaf	15.8	16
9	Pavel Datsyuk	17.7	13	Logan Couture	15.9	29	Henrik Sedin	15.2	7
10	Ryan Nugent-hopkins	17.4	18	Alex Ovechkin	15.8	16	Marian Hossa	14.9	21
11	Gabriel Landeskog	16.6	36	Marian Hossa	14.4	21	Alexander Semin	14.7	-1
12	Logan Couture	16.5	29	Alexander Semin	14.2	-1	Jaromir Jagr	14.5	28
13	Alex Ovechkin	15.8	16	Matt Moulson	13.9	22	Logan Couture	14.2	29
14	Marian Hossa	15.4	21	Tyler Toffoli	13.3	31	Matt Moulson	13.7	22
15	Alexander Semin	14.8	-1	David Perron	12.7	2	Mikko Koivu	13	12
16	Zach Parise	14.7	21	Mikko Koivu	12.5	12	Joe Pavelski	12.6	33
17	Frans Nielsen	13.5	8	Frans Nielsen	12.3	8	Steven Stamkos	12.6	24
18	Mikko Koivu	13.4	12	Ryan Getzlaf	12.1	16	Frans Nielsen	12.5	8
19	Matt Moulson	13.4	22	Ryan Nugent-hopkins	11.9	18	Marian Gaborik	12.3	29
20	David Perron	13.1	2	Jaromir Jagr	11.8	28	Zach Parise	12.2	21
		305 nonzero effects			*204 nonzero effects*			*64 nonzero effects*	

Table 1: Top 20 AICc selected player 'partial plus-minus' (PPM) values for the 2013-2014 season, under $\gamma = 0, 1, 10$. The number of nonzero player effects for each γ are noted along the bottom.

Candes, E. J., M. B. Wakin, and S. P. Boyd (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications 14*, 877–905.

Cevher, V. (2009). Learning with compressible priors. In *Neural Information Processing Systems (NIPS)*.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association 99*, 619–632.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*, 928–961.

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics 42*(3), 819–849.

Flynn, C., C. Hurvich, and J. Simonoff (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association 108*, 1031–1043.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Friedman, J. H. (2008). Fast sparse regression and classification. Technical Report, Dept. of Statistics, Stanford University.

Gramacy, R. B., S. T. Jensen, and M. Taddy (2013). Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports 9*.

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

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica 18*(4), 1603.

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

Mallows, C. L. (1973). Some comments on CP. *Technometrics 15*, 661–675.

Mazumder, R., J. H. Friedman, and T. Hastie (2011). SparseNet : Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association 106*, 1125–1138.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*, 681–686.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research 11*, 2241–2259.

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

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*, 1135–1151.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association 108*, 755–770.

Taddy, M. (2015). Distributed multinomial regression. *The Annals of Applied Statistics*. To appear.

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

Wainwright, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. *UC Berkeley Technical Report*.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using L1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Wang, L., Y. Kim, and R. Li (2013). Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics 41*(5), 2505–2536.

Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*, 894–942.

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research 11*, 1081–1107.

Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli 19*(5B), 2277–2293.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2007). On the degrees of freedom of the lasso. *The Annals of Statistics 35*, 2173–2192.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1509–1533.