One-step estimator paths for concave regularization

Matt Taddy
The University of Chicago Booth School of Business

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, 2013a).

Unfortunately, exact solvers for concave penalized estimation all require significantly more compute time than a standard lasso. In our experience, this has precluded their use in settings – e.g., in text or web-data analysis – where both n (the number of observations) and p (the number of covariates) are very large. As we review below, recent literature recommends the use of approximate solvers. Certainly, this is necessary for data of the size we encounter in analysis of, say, internet commerce. These approximations take the form of iteratively-weighted-L_1 regularization, where the coefficient-specific weights are based upon estimates of the coefficients taken from previous iterations of the approximate solver. This literature (e.g., Zou and Li, 2008; Fan et al., 2014) holds that even a single step of 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 with such one-step estimation (OSE) is finding starts that are, indeed, good enough.

This article provides a complete framework for sparse

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

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 model estimates do not change too much from one level of regularization to the next, such path 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; detailed in Section 3) algorithm 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. 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. 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 pat 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 (see Section 5) from which we are able to construct heuristic information criteria for penalty selection. We view such penalty-selection 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 4 is a survey of sparse diminishing-bias regularization: we review the connection between concave and weighted-L_1 penalties in 4.1, and present novel results on the distance between weighted-L_1 and L_0 minimization in 4.2. The POSE algorithm is introduced in Section 3, with our gamma lasso specification in 3.1 and its motivation from Bayesian foundations in 3.2. Section **??** covers implementation issues key to any path-based strategy: speed and stability in **??**.1, and model selection in **??**.2. Finally, Section **??** presents two empirical studies: a simulation experiment in **??**.1, and in **??**.2 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 Sparse regularization paths and diminishing bias

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 $l(\alpha, \boldsymbol{\beta}) = l(\boldsymbol{\eta})$ an unregularized objective proportional to 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 then the solution to

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

where $\lambda > 0$ controls overall penalty magnitude and are $c_j()$ are coefficient cost functions.

A few common cost functions are shown in Figure 1. 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 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.

The 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 obtained while moving from high to low penalization along $\lambda^1 > \lambda^2 \ldots > \lambda^T$ (e.g., LARS in Efron et al., 2004, is a well known example). These paths begin at λ^1 set to infimum λ such that (1) is minimized at $\hat{\boldsymbol{\beta}} = \mathbf{0}$ (see Appendix **??**), and proceed down to some pre-specified λ^T (e.g., $\lambda^T = 0.01\lambda^1$).

2.1 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 in

[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.

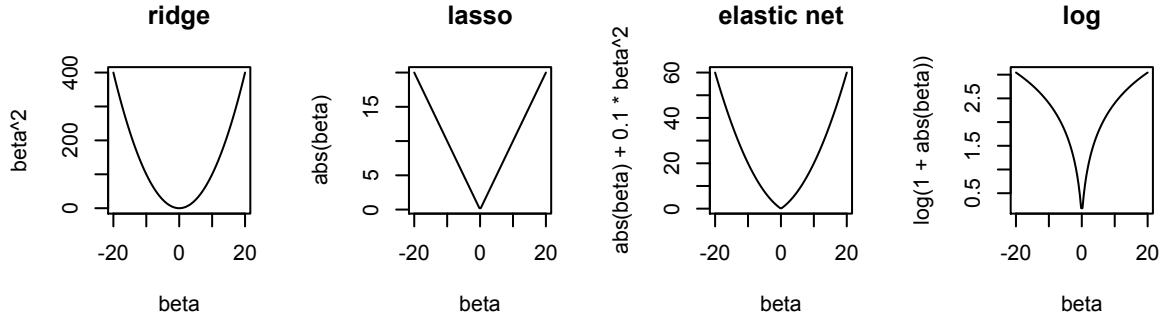

Figure 1: From left to right, L_2 costs (ridge, Hoerl and Kennard, 1970), L_1 (lasso, Tibshirani, 1996), the 'elastic net' mixture of L_1 and L_2 (Zou and Hastie, 2005), and the log penalty (Candes et al., 2008).

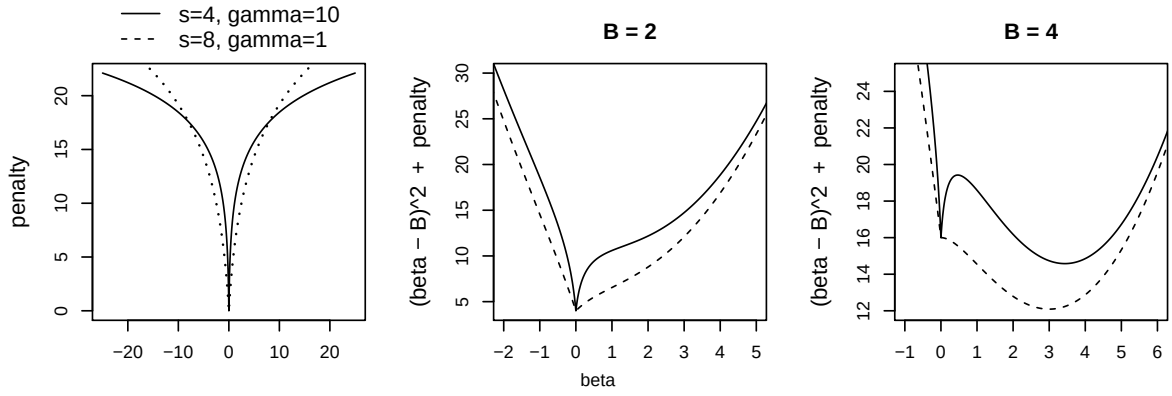

Figure 2: Log penalties $c(\beta) = s\log(1+\gamma|\beta|)$ and penalized objectives $(\beta - B)^2 + c(\beta)$.

the introduction. 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 for illustration.

This concavity makes computation difficult. For example, one run of SCAD via the `ncvreg` R package (Breheny and Huang, 2011) for the simulation in Section **??**.1 requires around 10 minutes, compared to 1-2 seconds for the lasso. The slowness is due to the difficulty of concave minimization: finding global, or even good local, minima is an extremely tough task. 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 under decreasing L_1 penalty magnitude and, for each segment on this path, adapts coefficient estimates along a second path of increasing penalty concavity. While far more efficient than previous alternatives, `sparsenet` relies upon solution over a large set of penalty specifications[2] and its compute cost remains much higher than for lasso (e.g., 15-20 seconds in our simulation).

A relatively fast class of solvers for concave penalized estimators uses local linear approximation (LLA; e.g., Candes et al., 2008). LLA replaces the concave cost function c_j with its tangent at the current estimate, $c_j'(\hat{\beta}_j)\beta_j$. The objective is then just a weighted L_1 penalized loss (solvable, say, as in Appendix **??**), and one 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 does well in practice. Importantly, they show that LLA does well even (or especially) 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. Such early-stopping can be as good as the full-step 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. He shows that just one or two steps here is sufficient for obtaining oracle support recovery

[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.

properties under much weaker conditions than required by a standard lasso. 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 $\hat{\beta}_j^0 = \text{cor}(\mathbf{x}_j, \mathbf{y})$.[3]

OSE LLA, or a two-step estimator starting from $\hat{\boldsymbol{\beta}} = 0$ as suggested in Fan et al. (2014) and Wang et al. (2013), or any version of the adaptive lasso, are *all just weighted-L_1 minimization*. These methods differ only in how the weights are constructed. Regardless of which you prefer, weighted-L_1 penalties are likely to play a central role in diminishing-bias penalization whenever the dataset is too large for use of exact concave penalty solvers.

3 Path of one-step estimators

Our path of one-step estimators (POSE), in Algorithm 1, uses solutions along the sequence of decreasing penalty sizes, λ^t, as the basis for LLA weights at the next path step. 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 $b \neq 0$. This yields a path of one-step LLA penalized coefficient estimates.

Algorithm 1 POSE

Initialize $\hat{\boldsymbol{\beta}}^0 = \mathbf{0}$, so that $\hat{S}_0 = \varnothing$.

Set $\lambda^1 > 0$ with step size $0 < \delta < 1$.
for $t = 1 \dots T$:

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

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

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

From an engineering standpoint, POSE has the same appeal as any successful path algorithm: if the estimates change little from iteration t to $t+1$, then you will be able to quickly solve for a large set of candidate specifications. Following the discussion of Section 4.1, such algorithms are a natural match with one-step estimation: OSE relies upon inputs being close to the optimal solution, which is precisely the setting where path algorithms are most efficient. More rigorously, Theorem **??** applied to POSE yields $\hat{S}_{t-1} \cap S^c = \varnothing \Rightarrow \omega_{S^c}^{t\,\min} = 1$. Thus so long as λ is large enough, Section 4.2 demonstrates that fast diminishing ω_j will help control false discovery and improve prediction. Of course, the moment $\hat{S}_t \cap S^c \neq \varnothing$, diminishing-bias allows spurious covariates to enter with little shrinkage and can move the fit arbitrarily far away from L_0-optimality – that is, with λ too small the diminishing bias hurts your ability to estimate and predict. This is why it is essential to have a path of candidate λ^t to choose amongst.

[3] We include marginal AL in our study of Section **??**.1.

4 Sparse regularization paths and diminishing bias

5 The gamma lasso

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

$$c(\beta_j) = \log(1 + \gamma|\beta_j|), \tag{4}$$

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 (see Figure 2). 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 (2013b) and Armagan et al. (2013).

GL – POSE under the log penalty – leads to line (2) being replaced by

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

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 3 shows solutions in a simple problem.

Each gamma lasso path segment is solved through coordinate descent, as detailed in Appendix **??**. The algorithm is implemented in `c` as part of the `gamlr` package for R. The software has detailed documentation and versioned source code is at `github.com/mataddy/gamlr`. 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.

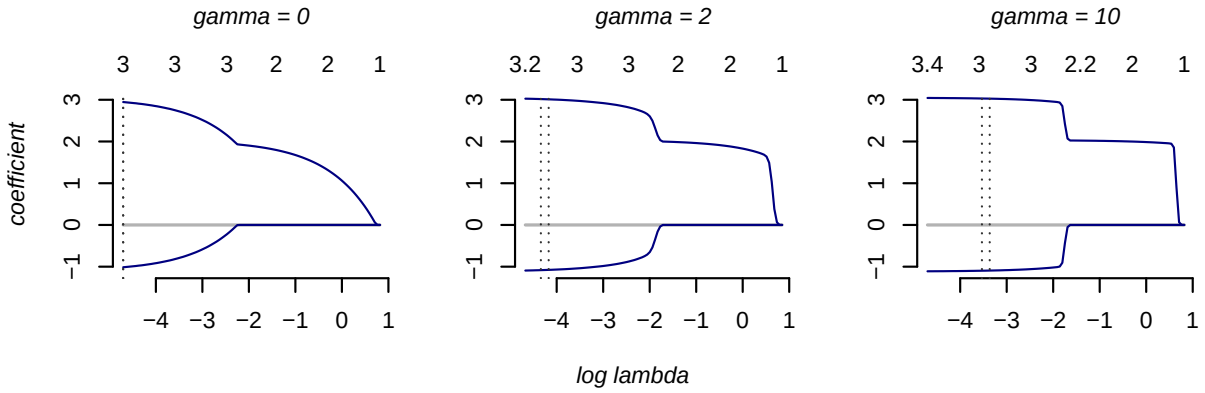

Figure 3: Gamma lasso estimation on $n = 10^3$ observations of $y_i = 4 + 3x_{1i} - x_{2i} + \varepsilon_i$, where $\varepsilon_i \overset{ind}{\sim} \mathrm{N}(0,1)$ and $\{x_{1i}, x_{2i}, x_{3i}\}$ are marginally standard normal with correlation of 0.9 between covariates (x_{3i} is spurious). The penalty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

5.1 Bayesian motivation

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

$$\beta_j \sim \mathrm{La}\left(\tau_j\right) = \frac{\tau_j}{2}\exp\left[-\tau_j|\beta_j|\,\right]. \tag{6}$$

Typically, scale parameters $\tau_1 = \ldots = \tau_p$ are set as a single shared value, say $n\lambda/\phi$ where ϕ is the exponential family dispersion (e.g. Gaussian variance σ^2 or 1 for the binomial). Posterior maximization under the prior in (6) is then lasso estimation (e.g., Park and Casella, 2008).

Instead of working from 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 $\mathrm{var}(\tau_j) = s\gamma^2$. Then the *joint* prior for both coefficient and scale is

$$\pi(\beta_j, \tau_j) = \mathrm{La}\left(\beta_j;\ \tau_j\right)\mathrm{Ga}\left(\tau_j;\ s, \gamma^{-1}\right) = \frac{1}{2\Gamma(s)}\left(\frac{\tau_j}{\gamma}\right)^s \exp\left[-\tau_j\left(1/\gamma + |\beta_j|\right)\right]. \tag{7}$$

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 (7), where we've suppressed α for simplicity. By taking negative logs and removing constants, this is equivalent to solving

$$\min_{\beta_j\in\mathbb{R},\ \tau_j\in\mathbb{R}^+} \ \phi^{-1}l(\boldsymbol{\beta}) + \sum_j \left[\tau_j(\gamma^{-1} + |\beta_j|) - s\log(\tau_j)\right]. \tag{8}$$

It is straightforward to show that (8) is equivalent to the log-penalized objective

$$\min_{\beta_j\in\mathbb{R}} \ \phi^{-1}l(\boldsymbol{\beta}) + \sum_j s\log(1+\gamma|\beta_j|) \tag{9}$$

PROPOSITION 5.1. *$\hat{\boldsymbol{\beta}}$ solves (9) if and only if it is also in the solution to (8).*

Proof. The conditional posterior mode for each τ_j given β_j is $\tau(\beta_j) = \gamma s/(1+\gamma|\beta_j|)$. Any joint solution $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}]$ for (8) thus consists of $\hat{\tau}_j = \tau(\hat{\beta}_j)$; otherwise, it is always possible to decrease the objective by replacing $\hat{\tau}_j$. Setting each $\tau_j = \tau(\beta_j)$ in (8) and removing constant terms yields (9). Moreover, the solution to (8) solves (9): otherwise, there would need to be a point on the profile slice of (8) defined by $\tau_j = \tau(\hat{\beta}_j)$ that is lower than its minimum. □

For a Bayesian it is odd to be solving for $\boldsymbol{\tau}$ rather than marginalizing over its uncertainty. However, recognizing the form of a gamma density in (7), $\pi(\beta_j, \tau_j)$ integrates over τ_j to yield the marginal prior $\pi(\beta_j) = 0.5s\left(1+\gamma|\beta_j|\right)^{-(s+1)}$. This is the generalized double Pareto density, as in Armagan et al. (2013). Since $-\log\pi(\beta_j) \propto (s+1)\log(1+\gamma|\beta_j|)$, the *profile* MAP solution to (8) is also the *marginal* MAP for $\boldsymbol{\beta}$ under $\mathrm{Ga}(s, 1/\gamma)$ priors on each τ_j.

References

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

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

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

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., 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, June). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics 42*(3), 819–849.

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.

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.

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.

Taddy, M. (2013a). Distributed multinomial regression. arXiv:1311.6139.

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

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

Wang, L., Y. Kim, and R. Li (2013, October). Calibrating nonconvex 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, November). 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. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1509–1533.