Statistical Science
2020, Vol. 35, No. 4, 614–622
https://doi.org/10.1214/20-STS809
Main article: https://doi.org/10.1214/19-STS701, https://doi.org/10.1214/19-STS733

A Look at Robustness and Stability of ℓ_1- versus ℓ_0-Regularization: Discussion of Papers by Bertsimas et al. and Hastie et al.

Yuansi Chen, Armeen Taeb and Peter Bühlmann

Abstract. We congratulate the authors Bertsimas, Pauphilet and van Parys (hereafter BPvP) and Hastie, Tibshirani and Tibshirani (hereafter HTT) for providing fresh and insightful views on the problem of variable selection and prediction in linear models. Their contributions at the fundamental level provide guidance for more complex models and procedures.

Key words and phrases: Distributional robustness, high-dimensional estimation, latent variables, low-rank estimation, variable selection.

1. A BIT OF HISTORY ON SUBSET SELECTION

ℓ_0 regularization in linear and other models has taken a prominent role, perhaps because of Occam's razor principle of simplicity. Information criteria like AIC (Akaike, 1973) and BIC (Schwarz, 1978) have been a breakthrough for complexity regularization with ℓ_0 regularization, see also Kotz and Johnson (1992). Miller's book on subset selection (Miller, 1990) provided a comprehensive view of the state-of-the-art 30 years ago. But the landscape has changed since then.

Breiman introduced the nonnegative garrote for better subset selection (Breiman, 1995), mentioned instability of forward selection (Breiman, 1996b) and promoted bagging (Breiman, 1996a) as a way to address it. A bit earlier, Basis pursuit has been invented by Chen and Donoho (1994), just before Tibshirani (1996) proposed the famous Lasso with ℓ_1 norm regularization that has seen massive use in statistics and machine learning; see also Chen, Donoho and Saunders (2001) and Fuchs (2004). The view and fact that ℓ_1 norm regularization can be seen as a powerful convex relaxation (Donoho, 2006) for the ℓ_0 problem has perhaps overshadowed that there are other statistical aspects in favor of ℓ_1 norm regularization, as pointed out now again by HTT in their current paper.

Yuansi Chen is Postdoctoral Fellow, Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland (e-mail: chen@stat.math.ethz.ch). Armeen Taeb is Postdoctoral Fellow, Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland (e-mail: armeen.taeb@stat.math.ethz.ch). Peter Bühlmann is Professor, Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

The scientific debate on whether to use ℓ_0 or ℓ_1 regularization to induce sparsity has always been active. One way to tackle the computational shortcoming of the exact ℓ_0 approach is through greedy algorithms. Tropp (2004) justified the good old greedy forward selection from a theoretical view point with no noise and Cai and Wang (2011) generalize this result to noisy situations, both assuming fairly restrictive coherence conditions on the design; the conditions in the latter work for noisy problems are weakened in Zhang (2011). Also matching pursuit (Mallat and Zhang, 1993) has been a contribution in favor of an algorithmic forward approach, perhaps mostly for computational reasons at that time. The similarity of matching pursuit to ℓ_1 norm penalization was made more clear by Least Angle regression (Efron et al., 2004) and L_2 boosting (Bühlmann, 2006). It is worth mentioning the recent work by Bertsimas, King and Mazumder (2016) illustrates that solving moderate-size exact ℓ_0 problem is no longer as slow as people used to think. Apart from their smart implementations, the speed-up they are able to attain benefits from tremendous progress in modern optimization theory and in physical computing powers.

On choosing between the ℓ_0 and the ℓ_1 approach, both BPvP and HTT have provided valuable statistical and optimization comparisons and discussions. To consolidate the contributions in both papers in a principled manner, here we outline what we believe are the three aspects that a data analyst must consider when deciding between ℓ_0 and ℓ_1 regularization: namely the application for which they will be employed, the optimization guarantees (computation speed and certificates of optimality) that are desired, and the statistical properties that the model might possess.

2. THREE ASPECTS OF THE PROBLEM

2.1 Application Aspect

Without a concrete data problem, the application aspect is perhaps not clearly distinguishable from a statistical data analysis viewpoint. Whether one would prefer ℓ_0 or ℓ_1 norm regularization is problem dependent. For example, for problems where the cardinality of model parameters is naturally constrained, a truly sparse approximation is reasonable and one may prefer ℓ_0 regularization. While ℓ_0 regularization is more natural from a modeling perspective, data analysts have often resorted to ℓ_1 regularization as a convex surrogate to speed up computations. However, the ℓ_1 surrogate can become problematic if the solution of the ℓ_1 problem were interpreted as that of the exact ℓ_0 problem. In particular, it is well known that ℓ_1 tends to overshrink the fitted estimates (Zou, 2006). That is, the ℓ_1 estimates are overly biased toward zero. This phenomenon can lead to poorly fitted models in practice. For example, ℓ_1 procedures for deconvolving calcium imaging data to determine the location of neuron activity lead to incorrect estimates (Friedrich, Zhou and Paninski, 2017), which has motivated exact solvers for the ℓ_0-penalized problem in this context (Jewell and Witten, 2018). In directed acyclic graph (DAG) estimation, ℓ_0 regularization is clearly preferred over ℓ_1 regularization. In particular, ℓ_0 regularization preserves the important property that Markov-equivalent DAGs have the same penalized likelihood score, while this is not the case for ℓ_1 regularization (van de Geer and Bühlmann, 2013).

The two examples above illustrate that many additional desiderata—other than computational complexity—may arise when deciding between ℓ_0 and ℓ_1 regularization. Depending on the application, a practitioner may consider: certificates of optimality (e.g., in mission critical applications), statistical prediction guarantees, feature selection guarantees, stability of the estimated model, distributional robustness and generalization of the theoretical understandings in linear models to more complex models. Our main objective in this discussion is to raise awareness of the above considerations when choosing between ℓ_0 and ℓ_1 regularization.

2.2 Optimization Aspect

Exact ℓ_0 regularization has long been abandoned because of the computational hardness. One could optimize using branch-and-bound techniques for moderate dimensions (Gatu and Kontoghiorghes, 2006, Hofmann, Gatu and Kontoghiorghes, 2007, Bertsimas and Shioda, 2009) or, as most statisticians did, resort to some forward–backward heuristics. These approaches either provide no guarantee for finding the optimal solution or are not computationally tractable for large-scale problems. On the other hand, ℓ_1 regularization is convex. When ℓ_1 regularization is combined with a convex loss, the optimization program can be solved efficiently using standard subgradient-based methods. Gradient methods are ubiquitous and ready-to-use in today's deep learning research and popular computing packages such as Tensorflow (Abadi et al., 2015) or PyTorch (Paszke et al., 2019). These publicly available computing packages make ℓ_1 regularization easily deployable in applications beyond linear models.

Of course, finding the global optimum may not be necessary for good statistical performance. The search for greedy algorithms for ℓ_0 regularization and the study of the corresponding statistical performance has been and still is an active research field (see Tropp, 2004, Cai and Wang, 2011 and Zhang, 2011 mentioned in Section 1). However, in order to provide both computational and statistical guarantees for these greedy algorithms, often fairly restrictive coherence conditions on the design matrix have to be assumed.

The computational difficulty of exact ℓ_0 regularization is not completely insurmountable. As pointed by the recent work of Bertsimas, King and Mazumder (2016), with the advance in physical computing powers and in the modern optimization theory on mixed-integer programming (MIO), data analysts now have a rigorous approach to employ ℓ_0 regularization efficiently in regression problems that would have taken years to compute decades ago. However, as pointed out by BPvP and HTT, the mixed-integer programming formulation of the ℓ_0 problem is the not yet as fast as Lasso, especially for low-SNR regimes and large problem sizes. In such scenarios, one often has to trade off between getting closer to the global minimum and terminating the program early under a time limit. As an example, HTT points out for problem size $n = 500$, $p = 100$, the method originally introduced in Bertsimas, King and Mazumder (2016) still requires an hour to certify optimality. The addition of ℓ_2 regularization by BPvP does improve computations in the low-SNR regime, in particular with their new convex relaxation SS for the $\ell_0 + \ell_2$ framework. As a convex relaxation, SS is a heuristic solution and is able to solve sparse linear regression of problem size $n = 10{,}000$, $p = 100{,}000$ in less than 20 seconds, putting it on par with Lasso. Overall, the contribution of Berstimas and coauthors in the series of related papers Bertsimas, King and Mazumder (2016), Bertsimas and King (2017), Bertsimas and Van Parys (2020) enables the possibility of ℓ_0 regularization in high-dimensional data analysis, and inspires future research to further integrate ℓ_0-based procedures.

2.3 Statistical Aspect

The statistical aspect is rooted in the goal to do accurate or "optimal" information extraction in the context of

uncertainty and aiming for stability and replicability. This inferential aspect may have very different aims, and depending on them and on the data generating processes, different methods might be preferred for different scenarios. Even more so, the statistical aspect might even embrace the idea that there are several "competing" methods and one would typically gain information by inspecting consensus or using aggregation.

Both BPvP and HTT papers demonstrate a careful and responsible comparison of subset selection methods, and we complement with a few additional empirical results in Section 3. With respect to statistical performance, HTT point out some key takeaways from their extensive empirical study: there is no clear winner, and depending on the size of the SNR, different methods perform favorably. In particular, they suggest that ℓ_0-based best subset suffers both computationally and statistically in the low-SNR regime. To address some of the instability concerns raised by HTT and inspired by the Elastic Net Zou and Hastie (2005), BPvP include an ℓ_2 penalty to their MIO formulation. With this new estimator, they demonstrate good performance across the SNR spectrum. We applaud both BPvP and HTT put the No Cherry Picking (NCP) guidelines (Bühlmann and van de Geer, 2018) in action.

2.3.1 *Relaxed and adaptive Lasso*: *A compromise between ℓ_0- and ℓ_1 regularization*. HTT find as an overall recommendation that the relaxed Lasso (Meinshausen, 2007), or a version of it, performs "overall best." In fact, when Nicolai Meinshausen came up independently with the idea around 2005, we learned that Hui Zou has invented the adaptive Lasso (Zou, 2006). Both proposals aim to reduce the bias of Lasso and push the solution more towards ℓ_0 penalization. So we wonder why HTT did not include the adaptive Lasso into their study as well (by using the plain Lasso as initial estimator, see also Bühlmann and van de Geer, 2011, Chapter 2.8). We would expect a similar behavior as for the relaxed Lasso.

The relaxed Lasso and adaptive Lasso above are "pushing" the convex ℓ_1 regularization towards the nonconvex ℓ_0 to benefit from the both regularization. In the same spirit, the $\ell_0 + \ell_2$ approach from BPvP combines the ℓ_0 regularization with the convex ℓ_2 regularization to increase stability in low SNR regime. If one is willing to accept the conceptual connection between the $\ell_0 + \ell_2$ approach from BPvP and the relaxed Lasso, it is no longer surprising to see that the $\ell_0 + \ell_2$ approach from BPvP performs better than best subset and Lasso from HTT in many simulation settings.

2.3.2 *Statistical theory.* The beauty of statistical theory is to characterize mathematically under which assumptions a certain method exhibits performance guarantees such as minimax optimality, rate of statistical convergence, and consistency. Such guarantees are typically provided for prediction performance or feature selection performance of a procedure.

Regarding the ℓ_0 and ℓ_1 regularization in high-dimensional sparse linear models, some statistical guarantees are known and some are still unknown (at least to us). For prediction, that is, estimating the regression surface $X\beta$ – ℓ_0 regularization is very powerful as it leads to minimax optimality under *no* assumption on the design matrix X (Barron, Birgé and Massart, 1999). For the Lasso, this is not true and the fast convergence rate requires conditions on the design such as the restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009) or the compatibility condition (van de Geer and Bühlmann, 2009, Bühlmann and van de Geer, 2011, Bellec, 2018). On the other hand, for feature selection accuracy, that is, estimation of the parameter vector β^*—one necessarily needs a condition on the fixed design matrix X, since β^* is not identifiable in general if $p > n$. Specifically, since the least squares loss function is the quadratic form $\|Y - Xb\|_2^2/n \approx (\beta^* - b)^T X^T X/n(\beta^* - b) + \sigma_\varepsilon^2$ with β^* denoting the true regression parameter and σ_ε^2 the error variance, we might necessarily need a restrictive eigenvalue/compatibility type condition to estimate β^* (these are in fact the weakest assumptions known for accurate parameter estimation or feature selection consistency with the Lasso). Whether ℓ_0 minimization has an advantage over the Lasso for parameter estimation (e.g., requiring weaker assumptions or yielding more accurate estimators) is unclear to us, and we believe investigating such relationships is an interesting direction for future research. On the fine scale, methods which improve upon the bias of the Lasso as the adaptive Lasso (Zou, 2006), the relaxed Lasso (Meinshausen, 2007) or thresholding after Lasso, are indeed a bit better than the plain Lasso, under some assumptions (van de Geer, Bühlmann and Zhou, 2011). We wonder whether these Lasso variants also have similar advantages when introduced with ℓ_0 regularization.

3. A FEW ADDITIONAL THOUGHTS

BPvP and HTT provide many empirical illustrations on prediction, parameter estimation, cross-validation or degrees of freedom. We complement this with empirical studies on the following points: (i) distributional robustness and (ii) feature selection stability. The first point (i) is very briefly mentioned by BPvP in Sections 2.1.1 and 2.3.1 but not further considered in their empirical results. Further, we highlight a few problem settings—with more complicated decision spaces than linear subset selection—where developing optimization tools for ℓ_0 regularization may be fruitful.

Implementation details. In all our experiments, Lasso is solved via the R package *glmnet* (Friedman, Hastie and Tibshirani, 2010) and best subset is solved via the R package *bestsubset*, with maximum computing time limit of

30 minutes following HTT. The convex relaxation of the $\ell_0 + \ell_2$ approach SS (introduced by BPvP) is solved via the Julia package *SubsetSelection*, with maximum computing iteration limit of 200 iterations following BPvP. While CIO is also an important method in BPvP, we did not include it here because we had a hard time executing the CIO code by BPvP. The difficulty is due to the version incompatibilities of installing both SS and CIO under the same Julia version in the current *SparseRegression* package provided by BPvP. We would like to encourage the authors to update the *SparseRegression* package to make their software contributions more accessible.

The code to generate all the figures here are written in R with the integration of Julia code via R package *JuliaCall*. Our code is released publicly in the Github repository *STSDiscussion_SparseRegression*.[1]

3.1 Distributional Robustness

It is well known that the Lasso has a robustness property with respect to "measurement error" as the robust optimization problem (Xu, Caramanis and Mannor, 2009):

$$\operatorname*{argmin}_b \max_{\Delta \in \mathcal{U}(\lambda)} \|Y - (X + \Delta)b\|_2 \tag{3.1}$$

with the perturbation set $\mathcal{U}(\lambda) = \{\Delta = (\delta_1, \ldots, \delta_p), \|\delta_j\|_2 \le \lambda \ \forall j\}$ is equivalent to the square root Lasso

$$\operatorname*{argmin}_b \|Y - Xb\|_2 + \lambda \|b\|_1.$$

We note that the square root Lasso and the Lasso have the same solution path when varying λ from 0 to ∞. Thus, the Lasso is robust under small covariate perturbations; "small" since the optimal choice of λ to guarantee the optimal statistical risk in the Lasso is typically of order $\sqrt{\log(p)/n}$.

More generally, Bertsimas and Copenhaver (2018) demonstrate that the robust optimization problem (3.1) with a norm $h(\cdot)$ and a perturbation set $\mathcal{U}(\lambda) = \{\Delta : \max_{b \in \mathbb{R}^p} \frac{\|\Delta b\|_2}{h(b)} \le \lambda\}$ is equivalent to $\operatorname{argmin} \|Y - Xb\|_2 + \lambda h(b)$, matching the previous result of Xu, Caramanis and Mannor (2009) in the ℓ_1 norm regularized regression setting. These results suggest a duality between norm-based regularization and robustness. However, as ℓ_0 is not a norm, the connection with robustness is not immediately transparent. We believe that theoretically characterizing the type of perturbations that the ℓ_0 regularization is robust to would be an interesting and important future research direction.

Due to the lack of theoretical comparisons between the distributional robustness of the ℓ_0 and ℓ_1 regularization, in the following, we investigate the distributional robustness of these two regularization techniques empirically. We evaluate the distributional robustness with the following DR metric which takes a regression parameter or its estimate as input and outputs its distributional robustness:

$$\mathrm{DR} : \beta \mapsto \max_{\Delta \in \mathcal{U}(\eta)} \|Y - (X + \Delta)\beta\|_2, \tag{3.2}$$

with (X, Y) generated independently identically distributed with respect to the training data that yielded the estimate for β. DR computes the "worst-case" prediction performance of β when the covariates X have been perturbed inside the set $\mathcal{U}(\eta)$. Here, we take the perturbation set $\mathcal{U}(\eta) = \{\Delta = (\delta_1, \ldots, \delta_p), \|\delta_j\|_2 \le \eta \ \forall j\}$, parametrized by the perturbation magnitude $\eta > 0$. To obtain a better understanding of DR, it helps to consider the distributional robustness difference (DRD)

$$\mathrm{DRD} : \beta \mapsto \max_{\Delta \in \mathcal{U}(\eta)} \|Y - (X + \Delta)\beta\|_2 - \|Y - X\beta\|_2. \tag{3.3}$$

The metrics DR and DRD are related trivially by the relation $\mathrm{DR}(\beta) = \mathrm{DRD}(\beta) + \|Y - X\beta\|_2$, for a fixed β. Further, due to the ℓ_1 norm and robustness duality obtained in Xu, Caramanis and Mannor (2009), $\mathrm{DRD}(\beta) = \eta\|\beta\|_1$. As a consequence, the DRD definition is independent of how the data (X, Y) is generated. We also deduce that $\mathrm{DR}(\beta) = \eta\|\beta\|_1 + \|Y - X\beta\|_2$. In other words, the distributional robustness of an estimator β (as measured by $\mathrm{DR}(\beta)$) trade-offs between the ℓ_1 norm of its solution and the prediction performance on unperturbed data.

3.1.1 *Empirical illustration*: *DRD vs regularization parameter.* We empirically explore distributional robustness difference, DRD, as a function of the regularization parameter for both the Lasso and best subset (i.e., the ℓ_0 regularization solver originally developed in Bertsimas, King and Mazumder, 2016).

In this simulation, we consider the stylized setting where $n = 100$, $p = 30$, $s = 5$, SNR $= 2.0$, and the design matrix is sampled from a Gaussian distribution with identity covariance (i.e., correlation $\rho = 0$). We further fix the perturbation magnitude $\eta = 1.0$. Figure 1 shows DRD of the Lasso and best subset estimates as a function of the regularization parameter λ for the Lasso and as a function of the number of features selected k for best subset. Due to the equality $\mathrm{DRD}(\beta) = \eta\|\beta\|_1$, the behavior of DRD mimics the ℓ_1 norm of the estimates in the regularization solution path. As expected, choosing larger λ leads to smaller DRD for the Lasso; we observe a similar behavior with best subset for smaller k.

For both problems, we label two choices for the regularization parameters: 1. the choice that leads to the best prediction performance on a separate validation set of size n; 2. the choice that leads to the best prediction performance on the same validation set among all regularization values that yield support size less than or equal to $s = 5$. Choice 1 prioritizes the test prediction performance, while Choice

[1] https://github.com/yuachen/STSDiscussion_SparseRegression

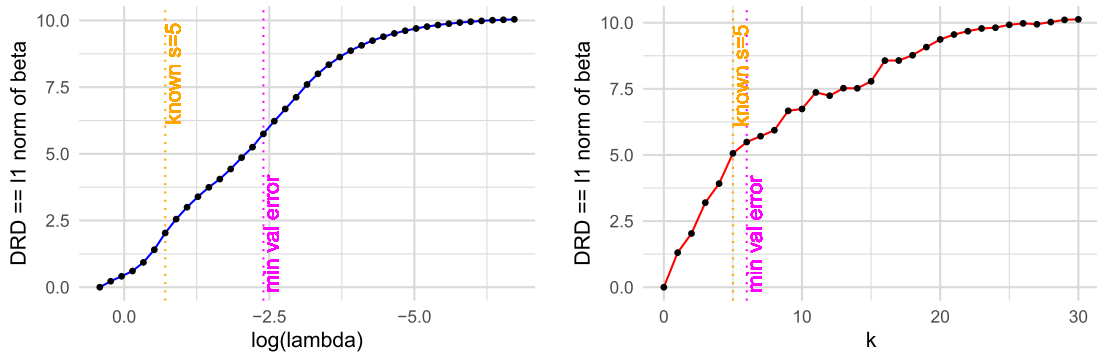


FIG. 1. *Left—Distributional robustness difference (DRD) of the Lasso as a function of the regularization parameter λ, right—DRD of best subset as a function of the regularization parameter k.*

2 prioritizes selecting the correct number of features. The left plot of Figure 1 shows that for the Lasso, the two choices lead to very different values for λ, and as a result, different DRD. Theoretically, the Lasso can be used to achieve good prediction performance and good feature selection performance, but the choice of λ for each goal is different (Wainwright, 2009, Zhao and Yu, 2006). In particular, optimizing with respect to prediction performance requires a smaller value of λ (Choice 1) and optimizing for model selection accuracy requires a larger values of λ (Choice 2). The right plot of Figure 1 shows that the two choices for best subset are similar. Comparing the left and right plots in Figure 1, we observe that Choice 1 yields approximately the same DRD for the Lasso and best subset. Choice 2, however, chooses a large λ for the Lasso, leading to a slightly better DRD for the Lasso than for best subset.

3.1.2 *Empirical illustration*: *Robustness versus signal-to-noise ratio and ambient dimension.* In this experiment, we explore the distributional robustness of the Lasso, best subset, as well as the new method SS developed by BPvP.

Following the simulation settings in HTT, we consider the following four data generation settings:

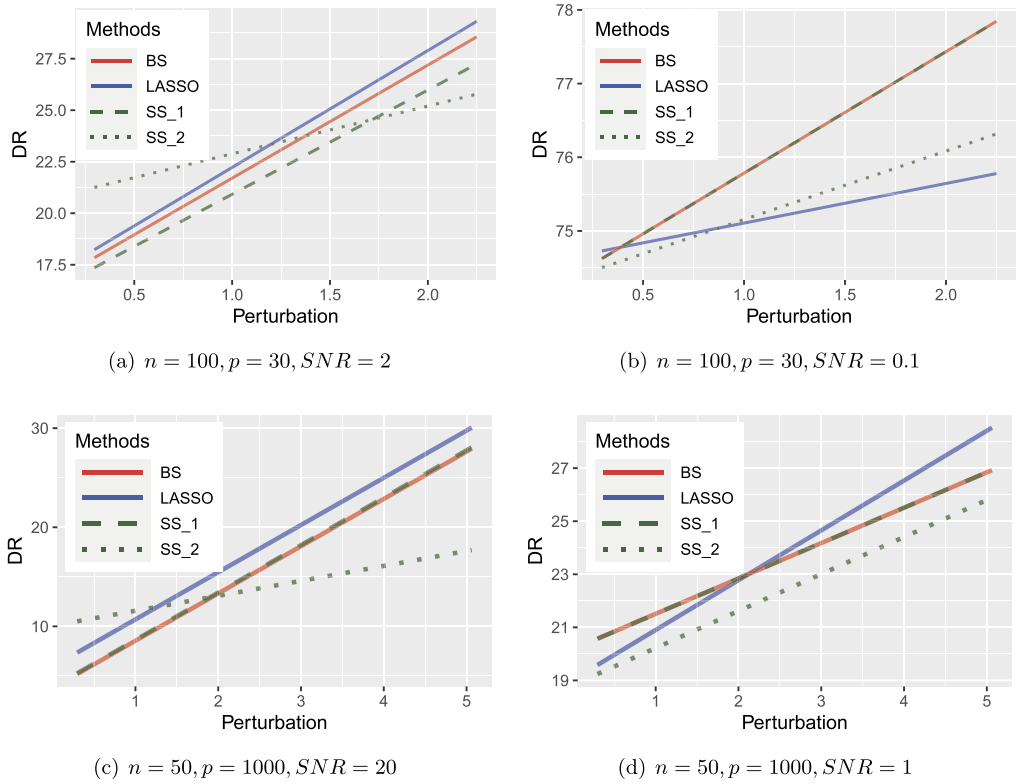
1. $n = 100$, $p = 30$, $s = 5$, SNR $= 2.0$, $\rho = 0$
2. $n = 100$, $p = 30$, $s = 5$, SNR $= 0.1$, $\rho = 0$
3. $n = 50$, $p = 1000$, $s = 5$, SNR $= 20.0$, $\rho = 0$
4. $n = 50$, $p = 1000$, $s = 5$, SNR $= 1.0$, $\rho = 0$,

and the design matrix is sampled from a Gaussian distribution with identity covariance (i.e., correlation $\rho = 0$). For the Lasso and best subset, we select their regularization parameters based on prediction performance on a separate validation set of size n. Since SS has two hyperparamter choices, we fix the ℓ_2 penalty regularization (denoted as $1/\gamma$ by BPvP) to take one of the two values $\gamma = \{1000, 0.01\}$ and tune the ℓ_0 regularization parameter based on validation performance. The two levels of the ℓ_2 penalty lead to two versions of SS: SS_1 ($\gamma = 1000$) with small ℓ_2 penalty and SS_2 with large ℓ_2 penalty ($\gamma = 0.01$). For each methods, after appropriately selecting the regularization parameters and computing the corresponding estimates, we evaluate the metric DR in equation (3.2) as a function of the perturbation magnitude η. Figure 2 compares the performance of the four methods. Given the relation $\text{DR}(\beta) = \eta\|\beta\| + \|Y - X\beta\|_2$, the Y-intercept of the linear curves in Figure 2 represent the prediction performance with unperturbed data, and the slopes are precisely the ℓ_1 norm of the β.

Focusing on the high-SNR regimes in the left column of Figure 2, the Lasso, best subset and SS_1 have comparable distributional robustness for small amounts of perturbation in both low and high dimensions. As a comparison, SS_2 has larger prediction error with unperturbed data but has better distributional robustness (as measured by DR) if the perturbations are large enough; this behavior can be attributed to the large ℓ_2 penalty in SS_2.

Focusing on the low-SNR regimes in the right column of Figure 2, in low dimensional setting 2, the Lasso is more robust than best subset. The robustness of the Lasso is largely due to the ℓ_1 shrinkage property provided by ℓ_1 regularization. In the low-SNR and high dimensional setting 4, best subset turns out to be more robust than the the Lasso. We observe that in the low-SNR and high dimensional setting, best subset selects very small number of features, enhancing its robustness. Once again, we observe that SS_2 yields substantially more robust solutions as compared to the other three methods.

Our experiments, as well as the empirical studies by BPvP, suggest that $\ell_0 + \ell_2$ minimization can substantially improve robustness. It is worth noting that this comes at the cost of choosing an additional regularization parameter. In practice, searching over the two-dimensional grid of regularization parameters could make SS less preferred choice for computational reasons. Nonetheless, the promising results of SS raise interesting questions about its statistical properties, as well its optimization guarantees. On the statistical side, it is interesting for future research to understand theoretically when the nonconvex



(a) $n = 100, p = 30, SNR = 2$ (b) $n = 100, p = 30, SNR = 0.1$

(c) $n = 50, p = 1000, SNR = 20$ (d) $n = 50, p = 1000, SNR = 1$

FIG. 2. *Distributional robustness (DR) of the Lasso, best subset,* SS_1 *and* SS_2 *for four ambient dimensions and SNR settings. For all problems, the data matrix is uncorrelated, that is,* $\rho = 0$.

$\ell_0 + \ell_2$ regularization has similar prediction, model selection or robustness guarantees as the Lasso. On the optimization side, since SS only solves a relaxation of the $\ell_0 + \ell_2$ regularization optimization problem, it is interesting to understand when the SS solution is close to the global optimum.

3.2 Sampling Stability for Feature Selection

Next, we empirically study the stability of best subset and Lasso as feature selection methods. We measure the feature selection stability by the probability of true and null variables being selected across multiple independent identically generated datasets. Ideally, we would like the probability of selecting any true variable to be close to one and any null variable close to zero. As with distributional robustness, the feature selection stability of the Lasso and best subset are strongly dependent on the choice of regularization parameter. We illustrate this phenomena in the supplementary material Section A.1 (see Chen, Taeb and Bühlmann, 2020).

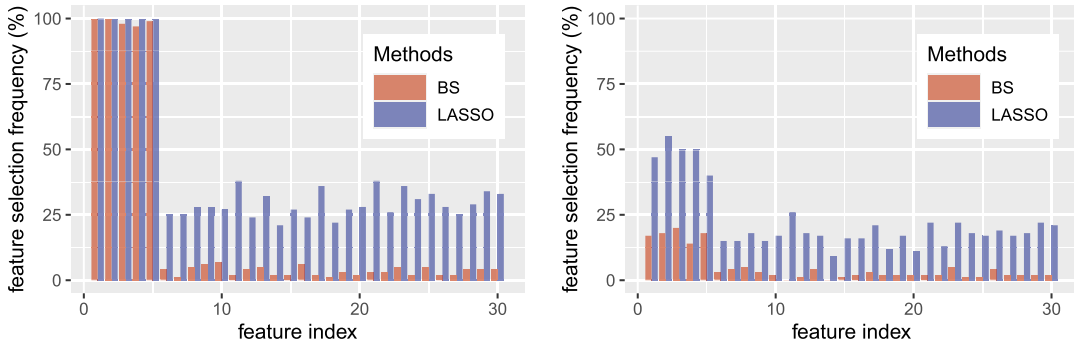
We consider the following stylized setting:

1. $n = 100$, $p = 30$, $s = 5$, SNR $= 2.0$, $\rho = 0.35$
2. $n = 100$, $p = 30$, $s = 5$, SNR $= 0.1$, $\rho = 0.35$

Figure 3 shows the feature stability of Lasso and best subset where the regularization parameters are chosen based on prediction on a validation set. In the high-SNR regime, the feature selection stability of the ℓ_0 approach is better than that of the Lasso. This is perhaps not very surprising as it is known that using validation error to select the regularization parameter for the Lasso yields overly complex models (Wainwright, 2009) (we consider in supplementary material Section A.2 of Chen, Taeb and Bühlmann, 2020 the setting where λ for the Lasso is chosen so that the solution has s nonzeros). In the low-SNR regime (albeit still a bit larger than the extreme SNR $= 0.05$ considered by HTT), we see both methods struggle to tease away true variable from null although we would argue that the Lasso performs favorably here as the true variables appear with much larger probability.

3.3 More Complicated Decision Spaces

Much of the focus of BPvP and HTT has been on subset selection for linear sparse regression settings. Although BPvP consider logistic regression and hinge loss in their mixed integer framework as well, the theoretical and computational aspects of these approaches appear less understood as compared to the linear setting. In practice, ℓ_1 norm regularization has been widely used beyond the linear setting for generalized linear models, survival models but also for scenarios with nonconvex loss functions such as mixture regression or mixed effect models (cf. Hastie, Tibshirani and Friedman, 2009, Bühlmann and van de Geer, 2011, Hastie, Tibshirani and Wainwright, 2015).



(a) SNR = 2 (b) SNR = 0.1

FIG. 3. *Feature stability of the Lasso and best subset across* 100 *independent datasets with problem parameters* $n = 100$, $p = 30$, $s = 5$, $\rho = 0.35$ *and* $\text{SNR} = \{2, 0.1\}$. *The regularization parameters are chosen based on prediction performance on a validation set.*

As such, we believe that deepening our understanding of ℓ_0-optimization for problems involving nonquadratic loss functions (beyond generalized linear models) is an important direction for future research. Furthermore, we next outline a few problem instances where ℓ_0-based approaches may provide a fresh and interesting perspective.

3.3.1 *Plug-in to squared error loss.* There are extensions of linear models in a plug-in sense, as indicated below, which can deal with important issues around hidden confounding, causality and distributional robustness in the presence of large perturbations.

The trick with such plug-in methodology is as follows. We linearly transform the data to $\tilde{X} = FX$ and $\tilde{Y} = FY$ for a carefully chosen $n \times n$ matrix F. Subsequently, we fit a linear model of $\tilde{Y}$ versus $\tilde{X}$, typically with a regularization term such as ℓ_1 norm or now, due to recent contributions by BPvP, we can also use ℓ_0 regularization. We list here two choices of F:

1. F is a spectral transform which trims the singular values of X. If there is unobserved hidden linear confounding which is *dense* affecting many components of X, then the regularized regression of $\tilde{Y}$ versus $\tilde{X}$ yields the deconfounded regression coefficient. (Cevid, Bühlmann and Meinshausen, 2018).
2. F is a linear transformation involving projection matrices and a robustness tuning parameter. Then Anchor regression is simply regression of $\tilde{Y}$ versus $\tilde{X}$, and we may want to use ℓ_0 or ℓ_1 regularization (Rothenhäusler et al., 2018). The obtained regression coefficient has a causal interpretation and leads to distributional robustness under a class of large perturbations.

3.3.2 *Fitting directed acyclic graphs.* Fitting Gaussian structural equation models with acyclic directed graph (DAG) structure from observational data is a well-studied topic in structure learning. As mentioned in Section 2.1, one should always prefer the ℓ_0 regularization principle as it respects the Markov equivalence property. The optimization of the ℓ_0-regularized log-likelihood function with a DAG constraint is very difficult, since the DAG constraint induces a high degree of non-convexity. The famous proposal of greedy equivalent search (GES) is the most used heuristics, but with proven crude consistency guarantees in low dimensions (Chickering, 2003). It would be wonderful to have more rigorous optimization tools, which could address this highly nonconvex problem.

3.3.3 *Extensions to low-rank matrix estimation.* A common task in data analysis is to obtain a low-rank model from observed data. These problems often aim at solving the optimization problem:

$$\underset{X \in \mathbb{R}^{n \times p}}{\text{argmin}}\, f(X) \quad \text{s.t. } \text{rank}(X) \le k, \tag{3.4}$$

for some differentiable function $f : \mathbb{R}^{n \times p} \to \mathbb{R}$ and $k \ll \min\{n, p\}$. The constraint $\text{rank}(X) \le k$ can be viewed as $\|\texttt{singular-values}(X)\|_{\ell_0} \le k$, making equation (3.4) a matrix analog of the subset selection problem analyzed in BPvP and HTT. Due to computational intractability of rank minimization, a large body of work has resorted to convex relaxation in order to obtain computationally feasible solution by replacing the rank constraint $\text{rank}(X)$ with the nuclear norm penalty $\lambda \|X\|_\star$ in the objective (Fazel, 2002), yielding a semidefinite program for certain function classes f. Such relaxations, while having the advantage of convexity, are not scalable to large problem instances. As such, practitioners often resort to solving other nonconvex formulations of equation (3.4) such as the Burer–Monteiro approach (Burer and Monteiro, 2003) and projected gradient descent on the low-rank manifold (Jain, Tewari and Kar, 2014). For a range of problem settings, these nonconvex techniques achieve appealing estimation accuracy (see Chen and Chen, 2018 for a summary), and are able to solve larger-dimensional problems than approaches based on convex relaxation. While these nonconvex approaches are commonly employed, they do not certify optimality. As such, nonlinear

semi-definite optimization techniques have been developed to produce more accurate solutions for some specializations of equation (3.4); see, for example, (Bertsimas, Copenhaver and Mazumder, 2017) in the context of rank-constrained factor analysis. Beyond the setting analyzed in (Bertsimas, Copenhaver and Mazumder, 2017), we wonder whether there are conceptual advancements to the mixed integer framework proposed in BPvP to handle low-rank optimization problems (3.4) involving continuous nonconvex optimization. Developing this connection may enable a fresh and powerful approach to provide—in a computationally efficiently manner—certifiably optimal solutions to equation (3.4).

4. CONCLUSIONS

It is great that practitioners can begin to integrate fast and exact ℓ_0-based solvers in their data analysis pipelines. The methods developed by BPvP can benefit many data analysts who would prefer to focus on the application aspect of the problem rather than the optimization aspect. Aiming for the most parsimonious model fit is a very plausible principle which is easy to communicate in many applications. But this alone does not rule out the attractiveness of ℓ_1 norm regularization and its versions: it builds in additional estimation shrinkage, which may be desirable in high noise settings, and sticking to convexity is yet another plausible principle. In this discussion, we tried to add a few additional thoughts to the excellent and insightful papers by BPvP and HTT, in the hope of encouraging new research in this direction.

ACKNOWLEDGMENTS

Y. Chen, A. Taeb and P. Bühlmann have received funding from the European Research Council under the Grant Agreement No. 786461 (CausalStats—ERC-2017-ADG). They all acknowledge scientific interaction and exchange at "ETH Foundations of Data Science."

SUPPLEMENTARY MATERIAL

Supplement to "A Look at Robustness and Stability of ℓ_1- versus ℓ_0-Regularization: Discussion of Papers by Bertsimas et al. and Hastie et al." (DOI: 10.1214/20-STS809SUPP; .pdf). The supplement contains an empirical analysis of the dependence of feature selection stability on the choice of regularization parameter for the Lasso and best subset. Further, in the context of feature selection stability, the effect of choosing λ for the Lasso to obtain a fixed cardinality is explored.

REFERENCES

ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G. S., DAVIS, A., DEAN, J. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. MR0483125

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028 https://doi.org/10.1007/s004400050210

BELLEC, P. (2018). The noise barrier and the large signal bias of the Lasso and other convex estimators. Preprint. Available at arXiv:1804.01230.

BERTSIMAS, D. and COPENHAVER, M. S. (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European J. Oper. Res.* **270** 931–942. MR3814540 https://doi.org/10.1016/j.ejor.2017.03.051

BERTSIMAS, D., COPENHAVER, M. S. and MAZUMDER, R. (2017). Certifiably optimal low rank factor analysis. *J. Mach. Learn. Res.* **18** Paper No. 29, 53. MR3634896

BERTSIMAS, D. and KING, A. (2017). Logistic regression: From art to science. *Statist. Sci.* **32** 367–384. MR3696001 https://doi.org/10.1214/16-STS602

BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618 https://doi.org/10.1214/15-AOS1388

BERTSIMAS, D. and SHIODA, R. (2009). Algorithm for cardinality-constrained quadratic optimizaiton. *Comput. Optim. Appl.* **43** 1–22. MR2501042 https://doi.org/10.1007/s10589-007-9126-9

BERTSIMAS, D. and VAN PARYS, B. (2020). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *Ann. Statist.* **48** 300–323. MR4065163 https://doi.org/10.1214/18-AOS1804

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720 https://doi.org/10.2307/1269730

BREIMAN, L. (1996a). Bagging predictors. *Mach. Learn.* **24** 123–140.

BREIMAN, L. (1996b). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383. MR1425957 https://doi.org/10.1214/aos/1032181158

BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559–583. MR2281878 https://doi.org/10.1214/009053606000000092

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

BÜHLMANN, P. and VAN DE GEER, S. (2018). Statistics for big data: A perspective. *Statist. Probab. Lett.* **136** 37–41. MR3806834 https://doi.org/10.1016/j.spl.2018.02.016

BURER, S. and MONTEIRO, R. D. C. (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.* **95** 329–357. MR1976484 https://doi.org/10.1007/s10107-002-0352-8

CAI, T. T. and WANG, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Inf. Theory* **57** 4680–4688. MR2840484 https://doi.org/10.1109/TIT.2011.2146090

CEVID, D., BÜHLMANN, P. and MEINSHAUSEN, N. (2018). Spectral deconfounding and perturbed sparse linear models. Preprint. Available at arXiv:1811.05352.

CHEN, Y. and CHEN, Y. (2018). Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Process. Mag.* **35** 14–31.

CHEN, S. and DONOHO, D. (1994). Basis pursuit. In *Proceedings of* 1994 28*th Asilomar Conference on Signals, Systems and Computers* **1** 41–44. IEEE, New York.

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43** 129–159. MR1854649 https://doi.org/10.1137/S003614450037906X

CHEN, Y., TAEB, A. and BÜHLMANN, P. (2020). Supplement to "A look at robustness and stability of ℓ_1- versus ℓ_0-regularization: discussion of papers by Bertsimas et al. and Hastie et al." https://doi.org/10.1214/20-STS809SUPP

CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554. MR1991085 https://doi.org/10.1162/153244303321897717

DONOHO, D. L. (2006). For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* **59** 797–829. MR2217606 https://doi.org/10.1002/cpa.20132

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067

FAZEL, M. (2002). Matrix rank minimization with applications. Ph.D. thesis, Stanford, CA.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

FRIEDRICH, J., ZHOU, P. and PANINSKI, L. (2017). Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13** 1–26.

FUCHS, J.-J. (2004). On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Theory* **50** 1341–1344. MR2094894 https://doi.org/10.1109/TIT.2004.828141

GATU, C. and KONTOGHIORGHES, E. J. (2006). Branch-and-bound algorithms for computing the best-subset regression models. *J. Comput. Graph. Statist.* **15** 139–156. MR2269366 https://doi.org/10.1198/106186006X100290

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

HOFMANN, M., GATU, C. and KONTOGHIORGHES, E. J. (2007). Efficient algorithms for computing the best subset regression models for large-scale problems. *Comput. Statist. Data Anal.* **52** 16–29. MR2409961 https://doi.org/10.1016/j.csda.2007.03.017

JAIN, P., TEWARI, A. and KAR, P. (2014). On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems* 685–693.

JEWELL, S. and WITTEN, D. (2018). Exact spike train inference via ℓ_0 optimization. *Ann. Appl. Stat.* **12** 2457–2482. MR3875708 https://doi.org/10.1214/18-AOAS1162

KOTZ, S. and JOHNSON, N. L., eds. (1992). *Breakthroughs in Statistics*. *Vol. I*: *Foundations and Basic Theory*. *Springer Series in Statistics*. *Perspectives in Statistics*. Springer, New York. MR1182794

MALLAT, S. and ZHANG, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41** 3397–3415.

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.* **52** 374–393. MR2409990 https://doi.org/10.1016/j.csda.2006.12.019

MILLER, A. J. (1990). *Subset Selection in Regression*. *Monographs on Statistics and Applied Probability* **40**. CRC Press, London. MR1072361 https://doi.org/10.1007/978-1-4899-2939-6

PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 8026–8037.

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2018). Anchor regression: Heterogeneous data meets causality. Preprint. Available at arXiv:1801.06229. To appear in J. Roy. Statist. Soc. Ser. B.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TROPP, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory* **50** 2231–2242. MR2097044 https://doi.org/10.1109/TIT.2004.834793

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. MR3099113 https://doi.org/10.1214/13-AOS1085

VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **5** 688–749. MR2820636 https://doi.org/10.1214/11-EJS624

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

XU, H., CARAMANIS, C. and MANNOR, S. (2009). Robust regression and Lasso. In *Advances in Neural Information Processing Systems* 1801–1808.

ZHANG, T. (2011). Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Inf. Theory* **57** 6215–6221. MR2857968 https://doi.org/10.1109/TIT.2011.2162263

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x