Web-based Supplementary Materials for Log-ratio Lasso: Scalable, Sparse Estimation for Log-ratio Models by

Stephen Bates and Robert Tibshirani

October 18, 2018

1 Proof of Main Results

The following propositions together give proof of the theorem. Let $b : \mathbb{R}^{\binom{p}{2}} \to \mathbb{R}^p$ be the map that takes a $\boldsymbol{\theta}$ from a log-ratio lasso feature space to the corresponding $\boldsymbol{\beta}$ in the standard feature space:

$$b(\boldsymbol{\theta})_k = \sum_{j=1}^{k-1} -\theta_{j,k} + \sum_{j=k+1}^{p} \theta_{k,j}.$$

Proposition 1. *For $\boldsymbol{\beta} = b(\boldsymbol{\theta})$ we have $\sum_{k=1}^{p} \beta_k = 0$.*

Proof.

$$\begin{aligned}
\sum_{k=1}^{p} \beta_k &= \sum_{k=1}^{p} [\sum_{j=1}^{k-1} -\theta_{j,k} + \sum_{j=k+1}^{p} \theta_{k,j}] \\
&= \sum_{1 \le j < k \le p} -\theta_{j,k} + \sum_{1 \le j < k \le p} \theta_{j,k} \\
&= 0
\end{aligned}$$

□

Proposition 2. *The model corresponding to $\boldsymbol{\beta} = b(\boldsymbol{\theta})$ and the model corresponding to $\boldsymbol{\theta}$ have the same sum of squared residuals.*

Proof.

$$\begin{aligned}\sum_{k=1}^{p} \beta_k \log(x_{i,k}) &= \sum_{k=1}^{p} [\sum_{j=1}^{k-1} -\theta_{j,k} + \sum_{j=k+1}^{p} \theta_{j,k}] \log(x_{i,k}) \\ &= \sum_{1 \le j < k \le p} -\theta_{j,k} \log(x_{i,k}) + \theta_{j,k} \log(x_{i,j}) \\ &= \sum_{1 \le j < k \le p} \theta_{j,k} \log \frac{x_{i,j}}{x_{i,k}}\end{aligned}$$

Thus the two models have the same fitted value for each observation $l = 1, ..., n$, and hence the same sum of squared residuals. □

Proposition 3. *For any $\boldsymbol{\beta}$ such that $\sum_{k=1}^{p} \beta_k = 0$, there exists a $\boldsymbol{\theta}$ such that $\boldsymbol{\beta} = b(\boldsymbol{\theta})$ with the property that $\|\boldsymbol{\beta}\|_1 = 2 \|\boldsymbol{\theta}\|_1$.*

Proof. Without loss of generality, suppose $\beta_1, ..., \beta_{p^+} \ge 0$ and $\beta_{p^+ + 1}, ..., \beta_p < 0$. Let $\theta_{i,j} = 0$ if $i, j \le p^+$ or if $i, j > p^+$. For $1 \le i \le p^+ < j \le p$, let $\theta_{i,j} = \frac{2|\beta_i||\beta_j|}{\|\boldsymbol{\beta}\|_1}$.

Now $\beta_k = \sum_{i=1}^{k-1} -\theta_{i,k} + \sum_{i=k+1}^{p} \theta_{k,i}$ so for $k \le p^+$ we have:

$$\begin{aligned}b(\theta)_k &= \sum_{i=1}^{k-1} -\theta_{i,k} + \sum_{i=k+1}^{p} \theta_{k,i} \\ &= \sum_{i=p^+ + 1}^{p} \theta_{k,i} \\ &= \sum_{i=p^+ + 1}^{p} \frac{2|\beta_i||\beta_k|}{\|\boldsymbol{\beta}\|_1} \\ &= 2\beta_k \sum_{i=p^+ + 1}^{p} \frac{|\beta_i|}{\|\boldsymbol{\beta}\|_1} \\ &= \beta_k.\end{aligned}$$

The last equality follows from the fact that $\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^{p} |\beta_j| = \sum_{j=1}^{p+} \beta_j - \sum_{j=p^+ + 1}^{p} \beta_j$ and

$\sum_{j=1}^{p} \beta_j = 0$. The analogous computation holds for $k > p^+$, so $\boldsymbol{\beta} = b(\boldsymbol{\theta})$.

Now notice:

$$\begin{aligned}
\|\boldsymbol{\beta}\|_1 &= \sum_{k=1}^{p} |\beta_k| \\
&= \sum_{k=1}^{p} |\sum_{i=1}^{k-1} -\theta_{i,k} + \sum_{i=k+1}^{p} \theta_{k,i}| \\
&= \sum_{k=1}^{p^+} |\sum_{i=p^+ +1}^{p} \theta_{k,i}| + \sum_{k=p^+ +1}^{p} |\sum_{i=1}^{p^+} -\theta_{i,k}| \\
&= \sum_{1 \le k < i \le p} |\theta_{k,i}| + \sum_{1 \le i < k \le p} |-\theta_{i,k}| \\
&= 2 \|\boldsymbol{\theta}\|_1 .
\end{aligned}$$

□

Remark 1. *We can see from this last proof that the log-ratio lasso fit is not identifiable: many different values of $\boldsymbol{\theta}$ correspond to both the same fit and the same 1-norm penalty.*

Remark 2. *The construction in this proposition gives us a way to find a solution to the log-ratio lasso optimization problem (main text equation 2) from a solution to the linearly constrained optimization problem (main text equation 3). We take a solution $\boldsymbol{\beta}^*$ of the linearly constrained lasso optimization and construct the corresponding $\boldsymbol{\theta}^*$. By the preceding proposition, these two solutions have the same loss plus penalty, so by theorem 1, $\boldsymbol{\theta}^*$ is a solution to the log-ratio lasso optimization problem (main text equation 2).*

Proposition 4. *Suppose $\boldsymbol{\theta}$ is a solution to the log-ratio lasso optimization problem, and let $\boldsymbol{\beta} = b(\boldsymbol{\theta})$. Then $\|\boldsymbol{\beta}\|_1 = 2\|\boldsymbol{\theta}\|_1$.*

Proof. Suppose not. Then $|\beta_k| \neq \sum_{i<k} |-\theta_{i,k}| + \sum_{k>i} |\theta_{k,i}|$ for some k. We will now show that the 1-norm of $\boldsymbol{\theta}$ can be reduced without changing the fitted values, which is a contradiction. Suppose without loss of generality that there exist $i < j < k$ such that $\theta_{i,k} < 0 < \theta_{j,k}$ with

$|\theta_{i,k}| > |\theta_{j,k}|$. Then consider a new fit $\widetilde{\boldsymbol{\theta}}$ with $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$ except for the following:

$$\begin{aligned}\widetilde{\theta}_{i,k} &= \theta_{i,k} + \theta_{j,k} \\ \widetilde{\theta}_{j,k} &= 0 \\ \widetilde{\theta}_{i,j} &= \theta_{i,j} + \theta_{j,k}.\end{aligned}$$

$\widetilde{\theta}$ results in the same fit as θ but has a 1-norm reduced by $\theta_{j,k}$. □

Combining these four propositions proves the main theorem.

Corollary 1. *The minimizer of the constrained lasso is unique if the matrix W has full rank.*

Proof. From the KKT conditions of the standard lasso optimization, one can show that all solutions $\boldsymbol{\theta}$ of the log-ratio lasso have equivalent fitted values $Z\boldsymbol{\theta}$ [see e.g. Tibshirani, Ryan J. et al. (2013)]. Since W has full rank, there is a unique $\boldsymbol{\beta}$ such that $Z\boldsymbol{\theta} = W\boldsymbol{\beta}$ and the result follows. □

2 Including unpaired logarithm terms

In some circumstances, it may be desirable to search for a model that includes both log-ratios and unpaired log terms:

$$y_i = \sum_{j=1}^{p} \beta_j \log(x_j) + \sum_{j<k}^{p} \theta_{j,k} \log(x_{i,j}/x_{i,k}).$$

The span of this model is again contained in the span of model (4), but in this case the analyst wishes to favor the selection log-ratio terms wherever possible. Fitting this model requires only a simple modification of the log-ratio lasso. We augment the feature matrix with the vector of ones: $x_{p+1} = 1$. Since $\log(x_i/x_{p+1}) = \log(x_i)$, the log-ratio lasso with the augmented feature matrix can select unpaired terms whenever it is beneficial to do so. This model fit gives a compromise between the standard lasso on the logarithmically transformed features and the

log-ratio lasso. The coefficient of x_{p+1} need not be zero, and the size of this coefficient is the amount of deviation from a pure all-pairs log-ratio model. When including unpaired terms, this is no longer a model for compositional data, so this should only be used in a setting where the magnitude of the individual features is believed to carry information about the response.

3 Including shrinkage in the two-stage procedure

As an alternative, one can fit the sparse regression procedure in step 3 using the predicted values $\widehat{y}$ from step 1 instead of the observed values y. This will result in a final model that is more similar to the single-stage log-ratio lasso fit, but the terms will be paired into ratios for easier interpretation. We will refer to this variant as *conservative two-stage log-ratio lasso*, since by maintaining the shrinkage from the first stage it will usually have less variance (at the price of more bias) than the standard two-stage procedure. As with all sparse regression procedures, these methods are most appropriate when the underlying signal is sparse, whereas other methods such as ridge regression are more appropriate for denser signals. We empirically study the performance of both versions of the log-ratio lasso estimator in the expanded simulation experiments in section 6 of the supplementary material.

4 Post-Selective Inference Technical Details

We will state the relevant technical results from Lee et al. (2016) and then customize them for our setting. Let $\widehat{M}$ and $\widehat{s}$ be the support set and signs selected by the lasso. Tibshirani, Ryan J. et al. (2016) and Lee et al. (2016) establish that the event $\{\widehat{M} = M\}$ can be expressed as a polyhedron:

$$\{\widehat{M} = M\} = \{A(M, s)y \leq b(M, s)\}. \tag{1}$$

The matrices $A(M, s)$ and vectors $b(M, s)$ are given by the following:

$$A(M,s) := \begin{bmatrix} \frac{1}{\lambda} X_{-M}^\top (I - P_M) \\ \frac{-1}{\lambda} X_{-M}^\top (I - P_M) \\ -\mathrm{diag}(s)(X_M^\top X_M)^{-1} X_M^\top \end{bmatrix}$$

$$b(M,s) := \begin{bmatrix} 1 - X_{-M}^\top (X_M^\top X_M)^{-1} X_M^\top s \\ 1 + X_{-M}^\top (X_M^\top X_M)^{-1} X_M^\top s \\ -\lambda \mathrm{diag}(s)(X_M^\top X_M)^{-1} s \end{bmatrix}$$

where P_M denotes the orthogonal projection onto the column span of X_M. Using this result, Lee et al. (2016) compute the conditional distribution of $\eta_M^\top y$ given $\{\widehat{M} = M\}$ for any vector η_M. That work explicitly treats the case where η_M is chosen to test hypotheses about the partial regression coefficients $\beta_j^{(M)}$, which is often interest. For our setting, we instead use these results to test whether a log-ratio model is consistent with the observed lasso fit, which we formulated as a formal hypothesis in (8). Taking $\eta_M = 1_M^\top (X_M^\top X_M)^{-1} X_M^\top$, we have that

$$\begin{aligned} \eta_M^\top \mathbb{E}[y] &= 1_M^\top (X_M^\top X_M)^{-1} X_M^\top \mathbb{E}[y] \\ &= 1_M^\top \mathbb{E}[(X_M^\top X_M)^{-1} X_M^\top y] \\ &= 1_M^\top \boldsymbol{\beta}_M. \end{aligned}$$

Thus, this choice of η_M corresponds to testing the hypothesis in (8). From here, an application of the Lee et al. (2016) machinery yields a pivotal quantity for $\eta_M^\top y$ after conditioning on $\{\widehat{M} = M\}$, which we encapsulate in the following proposition.

Proposition 1 (Post-selective test of the log-ratio model, detailed version)**.** *Let $F_{\mu,\sigma^2}^{[a,b]}$ be the CDF of a $N(\mu, \sigma^2)$ random variable truncated to the set $[a, b]$. Let $z := (I - P_{\eta_M})y$ be the residual of the projection of y onto η_M, which is independent of $\eta_M^\top y$, and let $c := \frac{\eta_M}{\|\eta_M\|^2}$.*

Define:

$$V^+_{M,s}(z) := \max_{j:(A(M,s)c)_j<0} \frac{b(M,s)_j - (A(M,s)z)_j}{(A(M,s)c)_j}$$
$$V^-_{M,s}(z) := \min_{j:(A(M,s)c)_j>0} \frac{b(M,s)_j - (A(M,s)z)_j}{(A(M,s)c)_j}.$$

Then the following holds:

$$F^{[V^-_{M,s}(z),V^+_{M,s}(z)]}_{1^\top \boldsymbol{\beta}^{(M)},\|\eta_M\|^2}(\eta_M^\top y)|\{\widehat{M} = M\} \sim \mathit{Unif}(0,1).$$

Proof. This is an application of Theorem 5.3 of Lee et al. (2016) with $\eta_M = 1^\top_M (X^\top_M X_M)^{-1} X_M$ using the characterization of the lasso selection event in (1). □

5 Solving the Constrained Lasso Optimization Problem

The constrained lasso optimization problem given in equation 3 is a convex optimization problem in p variables. It can be cast as an optimization problem with a quadratic objective function in $2p$ variables with only linear inequality constraints and a single linear equality constraint:

$$\begin{aligned}
\underset{\beta^+_1,\beta^-_1,...,\beta^+_p,\beta^-_p}{\text{minimize}} \quad & \frac{1}{2}\sum_{i=1}^{n}[\sum_{j=1}^{p}(y_i - \beta^+_j \log(x_{i,j}) + \beta^-_j \log(x_i,j))^2] + \lambda(\sum_{j=1}^{p}\beta^+_j + \beta^-_j) \\
\text{subject to} \quad & \beta^+_j \geq 0 \ \text{ for } \ j = 1,...,p \\
& \beta^-_j \geq 0 \ \text{ for } \ j = 1,...,p \\
& \sum_{j=1}^{p} \beta^+_j - \beta^-_j = 0.
\end{aligned}$$

Such an optimization problem can be efficiently solved with standard optimization libraries such as the popular open-sourced `CVX`[Grant & Boyd (2014)] for MATLAB or `CVXPY`[Diamond & Boyd (2016)] for python.

The constrained lasso optimization problem can also be solved efficiently using lasso solvers such as `glmnet` which allow for weighted observations. One simply augments the data with an additional data point with all features equal to 1, and response value zero. By assigning this value a large weight, the resulting solution $\boldsymbol{\beta}$ will have $\sum_{j=1}^{p} \beta_j \approx 0$. The value of $\sum_{j=1}^{p} \beta_j$ can be made arbitrarily small with large values of the weight. Similarly, for the logistic regression analog of the constrained lasso, one simply augments the feature matrix with two entries of large equal weight. One entry is assigned value 1 to all features and value 1 to the response. The other entry is assigned value 1 to all features and value 0 to the response. Because dedicated lasso solvers use specialized tricks to improve performance, this approach will typically be much faster than using a general-purpose convex optimization solver. Lin et al. (2014) give a detailed analysis of a coordinate descent algorithm for the constrained lasso that is similar to this proposal in the special case where the lasso solver is using coordinate descent [Friedman et al. (2010)].

6 Expanded Simulation Experiment

In this section we present an expanded version of the simulation experiment presented in the main text. We include the shrinkage version of the two-stage procedure explained in the supplamentary material as well as ridge regression.

We now examine the performance of our proposed methods for fitting log-ratio models with simulation experiments. We examine the following seven methods:

1. *Approximate forward stepwise* (approx-fs): the approximate forward stepwise procedure described in algorithm 2.

2. *Forward stepwise selection* (fs): forward stepwise selection applied on the logarithmically transformed features.

3. *Ridge regression* (ridge): ridge regression applied to the logarithmically transformed features.

4. *Single-stage log-ratio lasso* (single-stage): the method described in (2), which we showed is equivalent to the constrained lasso method of Lin et al. (2014).

5. *Two-stage log-ratio lasso* (two-stage): algorithm 1 using forward stepwise selection for the pruning stage.

6. *Two-stage log-ratio lasso* (two-stage-conservative): the variation of algorithm 1 described at the end of subsection 3.4, again using forward stepwise selection for the pruning stage.

7. *Lasso* (vanilla-lasso): the usual lasso estimator on the logarithmically transformed raw features.

All tuning parameters are chosen by cross-validation. Methods 2,3, and 7 are fitting linear models in the logarithmically transformed feature space of the form (4) whereas methods 1,4,5, and 6 are fitting log-ratio models of the form (1).

6.1 Experiment 1: Two Log-ratio Signals

We first examine the performance of our estimator when the data is generated from a log-ratio model. We consider the following model, consisting of two log-ratio terms of different amplitudes:

$$y_i = 2s\log(\frac{x_{i,1}}{x_{i,2}}) + s\log(\frac{x_{i,3}}{x_{i,4}}) + \epsilon_i \text{ for } i = 1, ..., n.$$

We take $\epsilon_i \overset{iid}{\sim} N(0,1)$. In the following simulations we use $n = 100$, $p = 30$, $X_{i,j} \overset{iid}{\sim} |N(0,1)|$. The signal strength s is taken across a grid of values from 0 to 3. We present the result in figure 1.

MSE, bias, and variance

We find that there is a large regime of signal strengths where the two-step procedure preforms significantly outperforms the original lasso and ridge regression. For coefficients from .5 to about 3, there is a MSE reduction of about 40% relative to the lasso. The two step procedure

has very low bias in the sparse setting, because it sets many coefficients to zero. It has slightly more variance than lasso, due to its discontinuous nature. We note that the conservative two-stage procedure has lower variance and higher bias than the two-stage procedure, as expected. We also note that the lasso and the single-stage log-ratio lasso are quite close, with the single-stage log-ratio lasso performing slightly better. This is expected; the single-stage log-ratio lasso has one extra piece of true information built into the procedure, the fact that the sum of the coefficients must be zero. Approximate forward stepwise selection has competitive MSE to the two-stage procedures. The approximate forward stepwise has superior performance to standard forward stepwise in this case, because approximate forward stepwise is picking out log-ratios, whereas forward stepwise is choosing single predictors. Overall, the two-stage procedures have the best performance in terms of MSE.

Support recovery

We next consider the support recovery properties of these procedures. Ridge regression is fitting a dense model, so it is omitted from the following discussion. The lasso and single-stage procedure recovers the signals slightly more often than the two-step procedure. This is expected, because the two-step procedure is a pruning of the single stage procedure. The two-step procedure selects very few null variables. This explains why this procedure has much better MSE and is an appealing aspect of this procedure in scientific contexts. The approximate forward stepwise procedure selects slightly more nulls than the two-stage procedures, and recovers the true signals slightly less frequently. Forward stepwise selection selects roughly the same number of non-nulls as the two-stage procedures, but selects the true signals slightly less often. Overall, the two-stage procedures have the best support recovery properties; these procedures recover the true signals very frequently and rarely select null variables.

6.2 Experiment 2: Robustness to Model Misspecification

We now generate data from a model that does not consist only of log-ratio terms. The data generating process is now:

$$y = 2s\log(\frac{x_1}{x_2}) + s\log(\frac{x_3}{x_4}) + .3\log(x_5) + \epsilon.$$

Notice the inclusion of an unpaired raw term $.3\log(x_5)$, so in this case we say the log-ratio model is misspecified. The amplitude of additional term is chosen to be large enough that it can be detected with high probability by the standard cross-validated lasso. We present the results in figure 2.

Even in the misspecified setting, we see that the two-step procedure has low MSE, again significantly outperforming both lasso and ridge regression. Forward stepwise selection has slightly better MSE that the two-stage procedures, and the approximate forward stepwise procedure has slightly worse MSE. We again see that the lasso and the single-stage procedure have similar performance. In this case, forward stepwise selection selects the fewest null variables, followed closely by the two-stage procedures and approximate forward stepwise selection. The two stage procedures recover the true signals slightly more often than the forward stepwise and approximate forwards stepwise procedures. This simulation study gives us some confirmation that even in the presence of moderate model misspecification, the log-ratio lasso procedure will retain good performance.

7 Real Data Example: Zero Replacement Details

In our data set, many of the entries are zero, which means the the chemical marker was not detected in the sample. In order to use the log-ratio lasso on this data set, we must first address these zero values. Various strategies for dealing with zeros in compositional data hae been explored in the literature. A useful survey of many methods is given by Martín-Fernández et al. (2003), and a corresponding R implementation called `zCompositions` was developed in

Palarea-Albaladejo & Martín-Fernández (2015).

On our data set, we compared three zero-replacement methods: additive imputation, multiplicative replacement, and imputation via a log-normal fit, the latter two were carried out using the `zCompositions` package. We compared the predictive accuracy of standard lasso and ridge regression (with penalty chosen by cross-validation) after these imputations, the results are presented in table 8. Both methods had significantly higher test set accuracy using the additive imputation (see the table below), so we use this replacement strategy in the paper. We speculate that the additive imputation is more heavily dampening the effects of covariates that are close to 0, which have very large leverage when converted to the log scale. Further exploration would be desirable, and we leave this as a direction for future work.

8 Software

A software implementation of the method and source code for the experiments and figures in the paper are available in an supplement.

References

Diamond, S. & Boyd, S. (2016), 'CVXPY: A Python-embedded modeling language for convex optimization', *Journal of Machine Learning Research* **17**(83), 1–5.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software, Articles* **33**(1), 1–22.

Grant, M. & Boyd, S. (2014), 'CVX: Matlab software for disciplined convex programming, version 2.1', `http://cvxr.com/cvx`.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016), 'Exact post-selection inference, with application to the lasso', *Ann. Statist.* **44**(3), 907–927.

Lin, W., Shi, P., Feng, R. & Li, H. (2014), 'Variable selection in regression with compositional covariates', *Biometrika* **101**(4), 785–797.

Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. (2003), 'Dealing with zeros and missing values in compositional data sets using nonparametric imputation', *Mathematical Geology* **35**(3), 253–278.

Palarea-Albaladejo, J. & Martín-Fernández, J. A. (2015), 'zCompositions - R package for multivariate imputation of left-censored data under a compositional approach', *Chemometrics and Intelligent Laboratory Systems* **143**, 85 – 96.

Tibshirani, Ryan J., Taylor, J., Lockhart, R. & Tibshirani, R. (2016), 'Exact post-selection inference for sequential regression procedures', *Journal of the American Statistical Association* **111**(514), 600–620.

Tibshirani, Ryan J. et al. (2013), 'The lasso problem and uniqueness', *Electronic Journal of Statistics* **7**, 1456–1490.

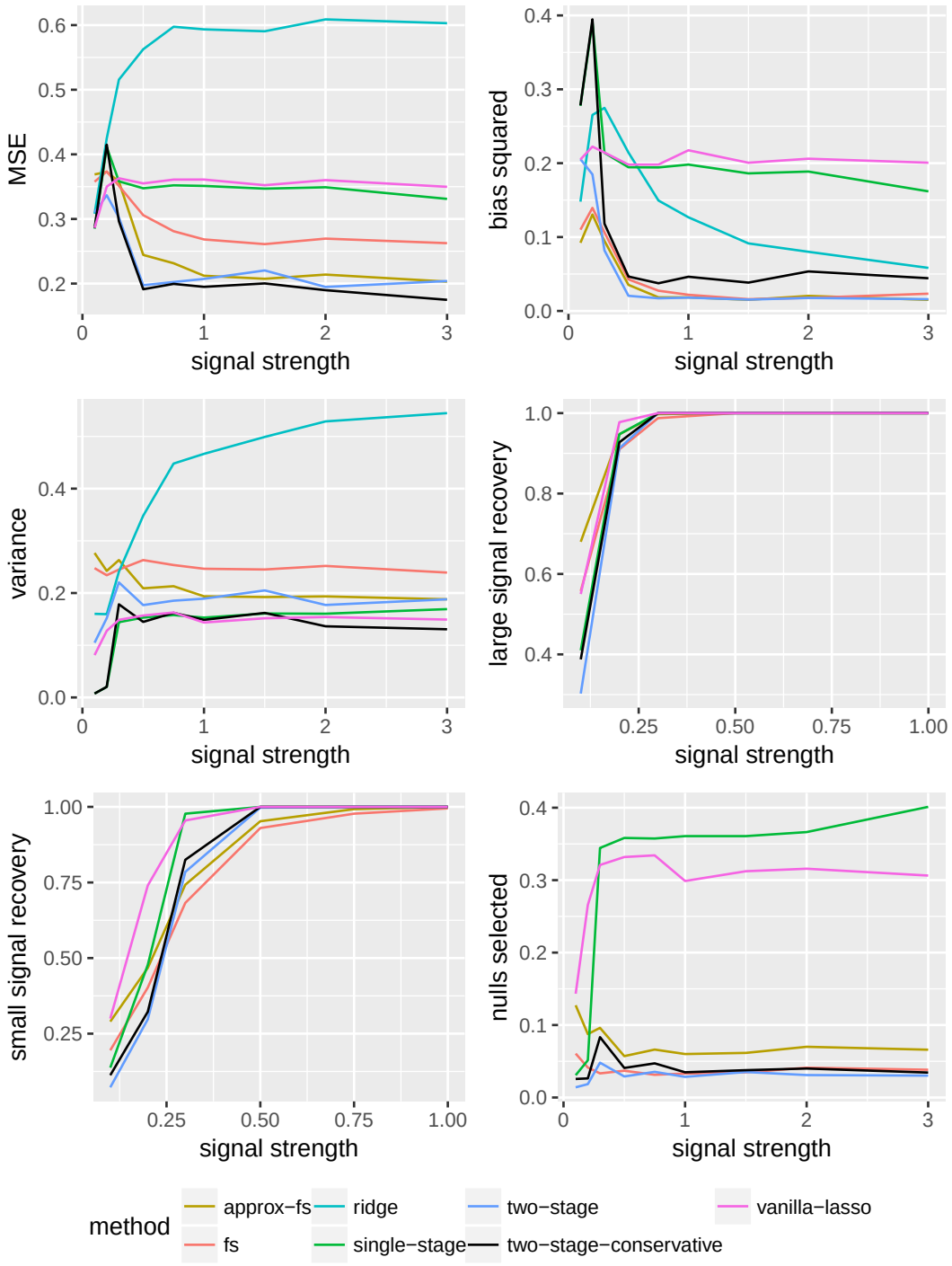

Figure 1: *Results of experiment 1: MSE and support recovery of log-ratio lasso in the sparse log-ratio model. The "large signal recovery" and "small signal recovery" graphs report the proportion of times that the true large signal and true small signal are selected, respectively. The "nulls selected" graph shows the average fraction of null variables that are selected.*

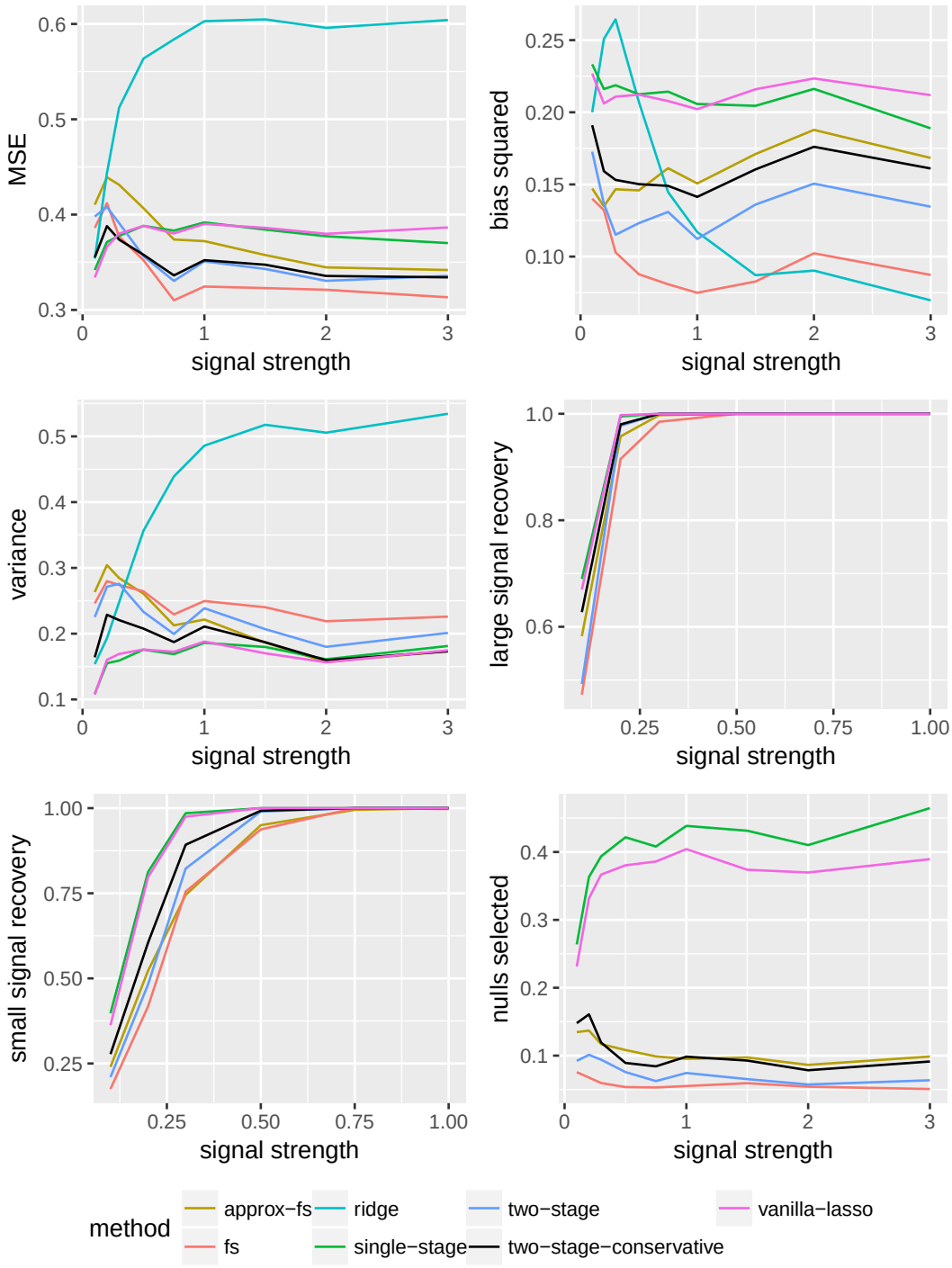

Figure 2: *Results of experiment 2: performance in the presence of model misspecification. The "large signal recovery" and "small signal recovery" graphs report the proportion of times that the true large log-ratio signal and true small log-ratio signal are selected, respectively. The "nulls selected" graph shows the average fraction of null variables that are selected.*

	lasso	ridge regression
additive imputation	0.718	0.708
multiplicative imputation	0.649	0.633
log-normal imputation	0.649	0.653

Table 1: Test set MSE of lasso and ridge regression across imputation schemes.