Fast Marginal Likelihood Estimation of Penalties for Group-Adaptive Elastic Net

Mirrelijn M. van Nee, Tim van de Brug, and Mark A. van de Wiel
Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

ABSTRACT
Elastic net penalization is widely used in high-dimensional prediction and variable selection settings. Auxiliary information on the variables, for example, groups of variables, is often available. Group-adaptive elastic net penalization exploits this information to potentially improve performance by estimating group penalties, thereby penalizing important groups of variables less than other groups. Estimating these group penalties is, however, hard due to the high dimension of the data. Existing methods are computationally expensive or not generic in the type of response. Here we present a fast method for estimation of group-adaptive elastic net penalties for generalized linear models. We first derive a low-dimensional representation of the Taylor approximation of the marginal likelihood for group-adaptive ridge penalties, to efficiently estimate these penalties. Then we show by using asymptotic normality of the linear predictors that this marginal likelihood approximates that of elastic net models. The ridge group penalties are then transformed to elastic net group penalties by matching the ridge prior variance to the elastic net prior variance as function of the group penalties. The method allows for overlapping groups and unpenalized variables, and is easily extended to other penalties. For a model-based simulation study and two cancer genomics applications we demonstrate a substantially decreased computation time and improved or matching performance compared to other methods. Supplementary materials for this article are available online.

1. Introduction
A popular approach to prediction and variable selection is elastic net penalization (Zou and Hastie 2005), as it simultaneously estimates and directly selects variables. Prediction and variable selection is hard for high-dimensional data, in which the number of variables far exceeds the number of samples. In addition to the main data, providing information on the samples, so-called co-data (van de Wiel et al. 2016) may be available, providing complementary information on the variables, for example, potentially overlapping groups of variables. These co-data may be exploited to improve the prediction and variable selection.

Early methods like group lasso (Meier, van de Geer, and Bühlmann 2008) or variations thereof (e.g., see Huang, Breheny, and Ma 2012 for a review) account for group structure by selecting entire groups with the level of regularization governed by one global penalty parameter. While this may be useful when there are many small groups, its limited flexibility may render inferior predictive performance for settings with a limited number of groups differing largely in prediction strength (Münch et al. 2019). More recent work includes co-data by allowing for differential group penalty parameters, thereby flexibly penalizing informative groups of variables less than non-informative ones. The full Bayesian analogue of this would be to impose group-specific shrinkage priors with a hyperprior on the group prior parameters. The latter, however, shows already for the basic elastic net a high sensitivity of the results to the specification of the hyperprior parameters (Li and Lin 2010). Instead, empirical Bayes approaches estimate group penalties from the data and show competitive results to full Bayesian approaches for high-dimensional data (van de Wiel, Te Beest, and Münch 2019). The empirical Bayes EM algorithm proposed in Li and Lin (2010) does not allow for multiple groups of variables and -like the full Bayes alternative-does not scale in the number of variables, as the Gibbs sampler is computationally expensive for high-dimensional data. The method ipflattering (Boulesteix et al. 2017) uses cross-validation to select the best group penalties from a proposed grid of possible values. In general, however, it is unclear what values should be proposed. Moreover, the method does not scale in the number of groups as the number of possible grid configurations increases exponentially. The method grap sewage (Münch et al. 2019), for group-adaptive elastic net, uses a variational Bayes algorithm to compute empirical Bayes estimates for the group penalties in logistic regression. The method is group-adaptive and scalable in the number of groups, but not in the number of variables. Moreover, it is only implemented for binary response. The method fweelnet (Tay et al. 2020) extends use of grouped co-data to continuous co-data (termed features of features). The method may be used for grouped co-data as well, for which it is shown to correspond to an elastic net penalty on the group level, with the amount of penalization governed by one global penalty parameter. Though fast, this method may, therefore, not be flexible enough. Finally, other group-adaptive methods have been developed specifically for priors other than the elastic net, for example, for spike-and-slab (Tang et al. 2018;
Here we propose a fast and efficient method for marginal likelihood estimation of group elastic net penalties. The method is group-adaptive and may be used for elastic net penalized generalized linear models in high-dimensional data. Notably, its use is limited to high-dimensional data, as its applicability to low-dimensional data (with a small number of variables) has not been proven. We include details for linear and logistic regression. Groups may be overlapping and the method allows for unpenalized variables, such as an intercept or low-dimensional baseline variables like age and sex. First, we derive a low-dimensional representation of the marginal likelihood for ridge models, a special case of elastic net that does not select variables. Then we show by using the asymptotic multivariate normality of the linear predictor that this marginal likelihood also approximates the marginal likelihood of elastic net models well. We include a practical check for this multivariate normality assumption. Finally, we show how the ridge group penalties are transformed to find optimal marginal likelihood group penalties for elastic net models that select variables, to accommodate medium sparse to dense models. Besides its main contribution of efficiently estimating group-adaptive elastic net penalties, our method has the additional advantage that it is easily extended for estimation of group-adaptive penalties other than the elastic net ones.

The outline of the article is as follows. Section 2 includes details of the method. Section 3 compares performance and computation times of several methods in a linear regression model-based simulation study. Then, it illustrates the benefit of learning from co-data for variable selection. Finally, the method is illustrated in the logistic regression setting on two cancer genomics data examples. Section 4 discusses the method and results.

2. Method

Let the response be given by \( Y \in \mathbb{R}^n \) and the observed high-dimensional data, \( p > n \), by \( X \in \mathbb{R}^{n \times p} \). Suppose that the co-data provide \( G < n \) groups of variables. First assume that groups are not overlapping. Section 2.4 discusses how the method may be used for partly overlapping groups too. Let the observed data for group \( g \) be given by \( X_g \). We model the response with a generalized linear model (glm) with corresponding canonical link function \( g(\cdot) \), mean \( \mu_i := E(Y_i | X_i, \beta, \phi) = g^{-1}(X_i \beta) \) and variance \( \text{var}(Y_i) = \phi V(\mu_i) \) for some specific variance function \( V(\cdot) \) and scale parameter \( \phi \) (McCullagh and Nelder 1989). Besides, we impose a conditional group-regularized elastic net prior on the regression coefficients \( \beta \in \mathbb{R}^p \), with group penalties \( \lambda \in \mathbb{R}^G \):

\[
\pi(Y_i | X_i, \beta, \phi) \overset{\text{ind}}{=} \exp \left( \frac{1}{\phi} (Y_i X_i \beta - b(X_i, \beta)) + c(Y_i, \phi) \right), \quad i = 1, \ldots, n,
\]

\[
\pi(\beta_k | \alpha, \lambda, \phi) \overset{\text{ind}}{=} \exp \left( -\frac{\lambda_{g_k}}{2\phi} |\beta_k| + (1 - \alpha)\beta_k^2 \right), \quad k = 1, \ldots, p,
\]

with \( b(\cdot) \) and \( c(\cdot) \) some specific functions corresponding to the type of generalized linear model used, and \( \lambda_{g_k} \) the group-specific penalty of group \( g \) to which variable \( k \) belongs. The group-specific penalties are scaled by \( \phi \) as in Li and Lin (2010) such that the penalized maximum likelihood estimate for \( \beta \) does not depend on \( \phi \) (Equation (3)). Note that the group-adaptive prior only depends on \( \phi \) and \( \lambda \) via the scaled inverse penalties defined as \( \tau^2 := \phi/\lambda \). Below, we consider \( X \) and \( \alpha \) fixed. Selection of \( \alpha \) is discussed in Section 2.3.1.

We use an empirical Bayes approach and estimate the group penalties and scale parameter for a given \( \alpha \) to arrive at the group-adaptive regularized elastic net estimate for the regression coefficients:

\[
(\hat{\lambda}, \hat{\phi}) = \arg \max_{\lambda, \phi} \pi(Y|\alpha, \lambda, \phi)
\]

\[
= \arg \max_{\lambda, \phi} \int_{\beta} \pi(Y|\beta, \phi) \pi(\beta|\alpha, \lambda, \phi) d\beta,
\]

\[
\hat{\beta} = \arg \max_{\beta} \left\{ \log(\pi(Y|\beta, \hat{\phi})) + \log(\pi(\beta|\alpha, \hat{\lambda}, \hat{\phi})) \right\}
\]

\[
= \arg \max_{\beta} \left\{ \sum_{i=1}^n [Y_i X_i \beta - b(X_i, \beta)]
\right.

\[
+ \sum_{k=1}^p \left[ -\frac{\hat{\lambda}_{g_k}}{2} (\alpha |\beta_k| + (1 - \alpha)\beta_k^2) \right] \right\}.
\]

The data \( X \) possibly contain some unpenalized variables next to the penalized variables, for example, for an intercept, which we will refer to by \( X_{\text{unpen}} \in \mathbb{R}^{n \times p_1} \) and \( X_{\text{pen}} \in \mathbb{R}^{n \times p_2} \), respectively, \( p = p_1 + p_2 \). In that case, only the penalized variables are integrated out in Equation (2). We refer to the unpenalized and penalized regression coefficients by \( \beta_{\text{unpen}} \) and \( \beta_{\text{pen}} \).

2.1. Fast Marginal Likelihood Estimation for Group-Adaptive Ridge

First, consider group-regularized ridge models (\( \alpha = 0 \)), as we need those results for the general elastic net setting. Wood (2011) derives a Laplace approximation for the marginal likelihood for semiparametric generalized linear models when the prior is of the form:

\[
\beta | \lambda, \phi \sim N(0, \phi S^2), \quad S = \sum_g \lambda_g S_g,
\]

with \( S^-1 \) a generalized inverse and \( S \) the ridge penalty matrix equal to the weighted sum over \( G \) different known, possibly full penalty matrices \( S_g \). The R-package mcp cv implementing this approximation is developed for low-dimensional data and does not allow for high-dimensional data. In theory, the results include the special case of group-adaptive ridge models in high-dimensional data as well, that is, substitute \( S_g \) by \( \bar{I}_g \in \mathbb{R}^{p \times p} \), with \( \bar{I}_g \) the diagonal matrix with the \( k \)th diagonal element equal to 1 if variable \( k \) belongs to group \( g \) and 0 otherwise, and substitute \( S \) by \( \Lambda := \sum g \lambda_g \bar{I}_g \) as the corresponding group-adaptive ridge, diagonal penalty matrix. In practice, however, the Laplace approximation may be inaccurate for high-dimensional integrals (van de Wiel, Te Beest, and Münch 2019) and is computationally expensive due to the large dimension of \( p \).
For ridge models, we may rewrite the high, \( p \)-dimensional integral as a lower, \( n \)-dimensional integral by observing that the likelihood only depends on \( \beta \) via the linear predictors \( \eta = X\beta \) (Veerman, Leday, and van de Wiel 2019). The resulting prior likelihood only depends on \( n \) likelihood for group-regularized ridge models is dependence on \( \lambda \) group to the hat matrix by \( \hat{\eta} \) that subsequently, we usetheshorthandnotation \( \hat{\eta} \) and drop the dependence on \( \lambda \).

The Laplace approximation of the negative log marginal likelihood for group-regularized ridge models is

\[
\ell(\lambda, \phi) \approx -\ell(\hat{\eta}, \phi) + \frac{1}{2\phi} (y - \mu)^T\hat{\eta} - \frac{1}{2} \log(|I_{nxn} - WH_{pen}|),
\]

with \( \ell(\cdot) \) denoting the log likelihood given parameters \( c, I_{nxn} \in \mathbb{R}^{nxn} \) the identity matrix, \( W \in \mathbb{R}^{nxn} \) the weight matrix in IWLS, and \( H_{pen} \) the hat matrix for the penalized variables:

\[
H_{pen} = X_{pen}(X_{pen}^T W_{pen} X_{pen} + \Lambda_{pen})^{-1} X_{pen}^T.
\]

We state the derivative in terms of \( \rho_g = \log(\lambda_g) \). First, denote the full hat matrix by \( H_g = X(X^T W + \Lambda)^{-1} X^T \), with the efficient lower-dimensional representation given in Section A, Supplementary materials. The partial derivatives of the Laplace approximation of the log marginal likelihood to the group parameters are:

\[
\frac{\partial (-\ell(\rho, \phi))}{\partial \rho_g} = \frac{1}{\phi} (y - \mu)^T V^{-1} g^{-1} [I_{nxn} - HW] H_{g}^T (y - \mu + W\hat{\eta})
\]

\[-\frac{1}{2\phi} (g^{-1}\hat{\eta} + y - \mu)^T [I_{nxn} - HW] H_{g}^T (y - \mu + W\hat{\eta}) + \frac{1}{2} \text{tr} \left( \frac{\partial W}{\partial g} \right)
\]

\[-\frac{1}{2} \text{tr} \left( \exp(-\rho_g)X_gX_g^T \left( W^{-1} + \sum_{g=1}^{G} \exp(-\rho_g)X_gX_g^T \right)^{-1} \right).
\]

where \( V \in \mathbb{R}^{n \times n} \) is the diagonal matrix with diagonal elements \( V(\mu_i) \), \( G \) is the diagonal matrix with diagonal elements \( G(\eta_k) \), and the partial derivative \( \frac{\partial \ell(\cdot)}{\partial \rho_g} \) readily obtained from Wood (2011). The chain rule implies that the partial derivatives for \( \lambda_g \) are obtained by multiplying the derivative to \( \rho_g \) by \( \frac{1}{\lambda_g} \).

We optimize \( \phi \) jointly with the group parameters. As \( \hat{\eta} \) does not depend on \( \phi \), the partial derivative is given by:

\[
\frac{\partial (-\ell(\rho, \phi))}{\partial \phi} = \frac{\partial (-\ell(\hat{\eta}, \phi))}{\partial \phi} = -\frac{1}{2\phi^2} (y - \mu)^T\hat{\eta}.
\]

The partial derivative of the log likelihood depends on the type of glm considered and is known for canonical models. We include details for linear regression in Section A, supplementary materials. For logistic regression, \( \phi = 1 \) is constant and the partial derivative is 0.

### 2.2. Asymptotic Normality of the Linear Predictors

Next, we show that the linear predictors are asymptotically multivariate normal, corresponding to a group-adaptive ridge prior. The marginal likelihood for elastic net models is then shown to be approximately equal to a reparameterization of the ridge marginal likelihood for which we just showed how to compute (and optimize) it very efficiently.

Without loss of generality, consider the penalized variables only and leave out subscripts \( \text{pen} \) for notational convenience. Denote the ridge penalty parameters, penalty matrix and scaled ridge prior variances by \( \lambda_R, \lambda_g \) and \( \tau_{g,R} = \rho/\lambda_R \) respectively. Define the variance function of the elastic net prior distribution for a given \( \alpha \in [0, 1] \) as \( h(\cdot) \):

\[
h(\tau_{g,R}^2) = h(\alpha \lambda_g) := \text{var}_{\beta|\alpha, \lambda, \phi}(\beta_k).
\]

Note that for each \( \tau_{g,R}^2, g \in \{1, \ldots, G\} \), there exist \( \tau_{g,R}^2 \) such that \( h(\tau_{g,R}^2) = \tau_{g,R}^2 \).

The prior distribution for \( \eta = X\beta \) is not analytical for \( \alpha > 0 \). An important observation is that we can exploit the high-dimensionality of the data to approximate the prior for \( \eta \) with a multivariate normal distribution. The following theorem follows from the multivariate central limit theorem for linear random vector forms in Eicker (1966):

**Theorem 1.** Suppose \( \beta_j \sim \pi_{G}(\beta) \) for \( j = 1, \ldots, p \) and group-specific prior \( \pi_{G}(\cdot) \), with \( E(\beta) = 0 \) and \( \text{var}(\beta) = \tau_{g,R}^2 I_{nxn} \in (0, \infty) \). For \( g = 1, \ldots, G \), let \( G \) be the group size and let \( X_g \in \mathbb{R}^{nxG} \) be the observed data corresponding to group \( g \). Let \( x_{s,j} \) denote the \( j \)-th column of \( X \in \mathbb{R}^{nxp} \). Suppose \( x_{s,j} \neq 0 \in \mathbb{R}^{n} \) for all \( j \), \( \text{rank}(X) = n \) for all \( p \), and for \( p \to \infty \),

\[
\max_{j=1,\ldots,p} x_{s,j}^T (XX^T)^{-1} x_{s,j} \to 0.
\]

Then, for fixed \( G, x_{s,j}, n, p \to \infty \),

\[
\left( \sum_{g=1}^{G} \tau_{g,R}^2 X_gX_g^T \right)^{-1/2} X \beta \overset{d}{\to} N(0, I_{nxn}),
\]

where \( I_{nxn} \) is the \( (n \times n) \)-dimensional identity matrix and \( \left( \sum_{g=1}^{G} \tau_{g,R}^2 X_gX_g^T \right)^{-1/2} \) is the inverse of the unique positive definite square root of \( \sum_{g=1}^{G} \tau_{g,R}^2 X_gX_g^T \).
If \( n = 1 \) then condition (11) is equivalent to \( \max_{j=1,\ldots,p} x_{ij}^2 / \sum_{j=1}^p x_{ij}^2 \to 0, \) for \( p \to \infty. \) Informally, condition (11) can be interpreted as each variable being asymptotically negligible in size compared to the full dataset, illustrated in Figure 1. In practice, this condition will be reasonable for most high-dimensional data, especially as these data are often standardized, but counter examples may exist such as data with extreme outliers.

Hence, the prior on \( \eta \) under an elastic net prior on \( \beta \) may be approximated by the following multivariate normal distribution for \( \tau_{RG}^2 \) such that \( \tau_{RG}^2 = h(\tau_g^2) \) for each \( g: \)

\[
\pi(\eta|\alpha, \lambda, \phi) \approx N \left( 0, \sum_g \tau_{RG}^2 X_g X_g^T \right) = N \left( 0, \sum_g \tau_{RG}^2 X_g X_g^T \right) = \pi(\eta|\lambda_R, \phi). \tag{13}
\]

Then, for \( \lambda_{RG} \) corresponding to \( \tau_{RG}^2 \) such that \( \tau_{RG}^2 = h(\tau_g^2) \), the marginal likelihood may be approximated as follows:

\[
\pi(Y|\alpha, \lambda, \phi) = \int_{\eta} \pi(Y|\eta) \pi(\eta|\alpha, \lambda, \phi) d\eta \\
\approx \int_{\eta} \pi(Y|\eta, \phi) \pi(\eta|\lambda_R, \phi) d\eta = \pi(Y|\lambda_R, \phi), \tag{14}
\]

where the latter expression is efficiently computed using Equation (6). Hence, the marginal likelihood of group-regularized elastic net models is approximately the same as that of ridge models, but parameterized differently. The partial derivatives from the elastic net parameterization may then be obtained from the partial derivatives for the ridge parameterization as given in Equations (8) and (9) by using the chain rule and a change of variables.

2.3. Fast Marginal Likelihood Estimation for Group-Adaptive Elastic Net

As we desire variable selection, we would like to obtain group-adaptive elastic net penalties for other cases (0 < \( \alpha \leq 1 \)) than the ridge model (\( \alpha = 0 \)). Given the approximation for the marginal likelihood in Equation (14), one could again use a general purpose optimizer to maximize the marginal likelihood for the elastic net parameterization. We, however, propose to use an indirect approach and simply transform the marginal likelihood estimates for the ridge parameterization to the elastic net parameterization using the known variance function. As the marginal likelihood of the ridge and elastic net model are approximately equal, the maximal value, obtained in the transformed maximizer, is also approximately equal. So, the elastic net estimates are given by

\[
\hat{\tau}^2 = h^{-1}(\hat{\tau}_R^2), \hat{\tau}_R = \hat{\tau}_g^2, g = 1, \ldots, G, \tag{15}
\]

where \( h^{-1}() \) is applied element-wise. The proposed approach has the advantage that, once the optimal ridge penalties are obtained, the optimal elastic net penalties are quickly obtained for a whole range of possible \( \alpha \) values. Figure 2 illustrates this.

The variance function for the elastic net prior for a given \( \alpha \) is

\[
\begin{align*}
\hat{\tau}_g^2 &= \var\beta_{\alpha,\lambda_g,\phi}(\beta_k) = \tau_g^2(1-\alpha)^{-1} \\
&= \frac{\alpha \tau_g}{2(1-\alpha)^{3/2}} \left( \frac{\alpha}{2\tau_g(1-\alpha)^{1/2}} \right) + \frac{\alpha^2}{4(1-\alpha)^2},
\end{align*}
\]

with \( \Phi(\cdot) \) and \( \varphi(\cdot) \) the standard normal cumulative density function and probability density function respectively.
The expression simplifies for lasso ($\alpha = 1$) to $h(\tau^2_g) = 8\tau^4_g$, and is easily solved analytically. For $0 < \alpha < 1$, we use the root-finding algorithm of the R-function uniroot to find the roots of $h(\tau^2_g) - \tau^2_g R_g = 0$. This suffices for values of $10^{-6} < \tau^2_g < 10^6$. The evaluation of the variance function is numerically unstable for more extreme values of $\tau^2_g$. Therefore, we truncate the values to either the ridge or lasso estimate while maintaining the monotonicity for more extreme $\tau^2_g$, as illustrated in Figure 2. We expect that this fix will suffice as the precise absolute value has less impact in the extremes.

2.3.1. Practical Selection of $\alpha$
Joint optimization of $\alpha$ and one elastic net penalty $\lambda$ is hard as the marginal likelihood shows a flat optimal region (van de Wiel, Te Beest, and Münch 2019). For our approach specifically, the approximated marginal likelihood is even identical for all $\alpha$ and transformed $\lambda$ (Equation (14)), making a jointly optimal $\alpha$ unidentifiable. Stacked elastic net (Rauschenberger, Glaab, and van de Wiel 2020) may improve elastic net performance by combining multiple values of $\alpha$ instead of selecting one value, but does not produce sparse models. In practice, we advise to compare the cross-validated likelihood or another performance criterion for few, say four or five, fixed $\alpha$-values ranging from 0 to 1. One may then select the model with the best performance or the sparsest model with similar performance. In our application in Section 3.2, for example, we compared the cross-validated AUC performance of models for $\alpha \in \{0.01, 0.3, 0.8, 1\}$.

2.4. Overlapping Groups
So far, we have assumed the groups to be nonoverlapping. A simple way to allow for partly overlapping groups is to make artificial, nonoverlapping groups, similarly as proposed as naive implementation of the latent overlapping group lasso (Jacob, Obozinski, and Vert 2009). Full details are included in Section A.5, supplementary materials. In short, columns of $X$ of variables that belong to multiple groups are first duplicated and scaled such that the corresponding prior on the linear predictors is identical to the one corresponding to the original data, to account for multiplicity. This results in the extended data matrix $\tilde{X}$ with artificial nonoverlapping groups, on which our method may be used to obtain the group-adaptive penalty estimates. The dimension of $\tilde{X}$ increases rapidly for largely overlapping groups. In practice, however, it is not necessary to store this matrix as the computations for the marginal likelihood require the high-dimensional computation of $G \times n \times n$-dimensional matrices $X^T X_g$ only once. Finally, given the penalty estimates, the original data matrix $X$ is used to obtain estimates for the regression coefficients.

2.5. Recalibration by Cross-validation
In particular for logistic regression, we experienced that the components $\lambda_g$ of $\lambda$ are estimated well in a relative sense, but less so in the absolute sense: penalties are overestimated, leading to overpenalized regression coefficients and sub-optimal prediction performances (exemplified in Section B.1.1 in the supplementary materials). While in practice the Laplace approximation is accurate enough to be applied to a wide range of distributions (Rue, Martino, and Chopin 2009), the problem of overpenalization is known to arise when the marginal likelihood is approximated for “difficult” binary responses (Kuss, Rasmussen, and Herbrich 2005; Ferkingsdal and Rue 2015). The overpenalization does not show for Poisson distributed response or binomial response with more than one repeated observation (Veerman, Leday, and van de Wiel 2019). Initial errors in the ridge penalty maximizers may be amplified when transforming to elastic net penalties (see Figure 6).

Though more accurate, but more complex approximations may improve the estimates (Ferkingsdal and Rue 2015; Kuss, Rasmussen, and Herbrich 2005), here we propose to counter the overpenalization with a simple approach of recalibrating the group penalties with one cross-validated global rescaling penalty, $\lambda_g$: $\lambda = \lambda_0 \lambda_g$. So, after estimating the group penalties $\lambda$ by maximizing the approximate marginal likelihood, we estimate $\lambda_0$ by cross-validation of the likelihood. As $\lambda_0$ is a scalar, this can efficiently and easily be done, for example, by using glmnet with penalty factors $\lambda$. Whether or not the recalibration step is needed for generalized linear models other than for linear and logistic response may in practice be checked by comparing the prediction performance of the models with and without recalibration.

2.6. Normality Check
Our hyperparameter estimation relies on the assumption of multivariate normality of $\eta = X\beta$, where $\beta$ is generated from the prior. A posteriori, this assumption is easily checked by repeatedly sampling $\beta$ from its estimated (elastic net) prior, rendering $\beta^{(k)}$, and computing $\eta^{(k)} = X\beta^{(k)}$, for $k = 1, \ldots, K$. Then, compute the Mahalanobis distances $d^{(k)} = (\eta^{(k)} - \mu)^T \Sigma^{-1} (\eta^{(k)} - \mu)$, where $\mu = 0$, because the prior has mean zero, and covariance matrix $\Sigma = \phi \sum g \lambda^{-1}_g X_g X^T_g$ as in Equation (5). Finally, we propose the standard visual check for multivariate normality by a QQ-plot: the empirical quantiles of $d$ are plotted against the theoretical ones of the $\chi^2_g$ distribution. Figure A1, supplementary materials shows some examples for different priors. The normality check is accommodated by the software.

2.7. Extension to other Sparse Priors
The method may easily be extended to other group-adaptive sparse penalties corresponding to group-specific priors with finite variance. For any such prior, we only need to know its variance function and invert this (numerically) to find the marginal likelihood estimates for the group-specific prior parameters as in Equation (15). We provide examples in Section A.7, supplementary materials for two other well-known priors; a spike-and-slab prior and the generalized normal prior, corresponding to the bridge penalty.

3. Data Examples
The method is termed `squeezy` as it squeezes out some sparsity from dense group-adaptive ridge models by transforming ridge penalties to sparse penalties. We conduct a model-based linear regression simulation study and illustrate the method on...
Table 1. Overview of the methods compared in Section 3.

| Method     | glmnet | fwelnet | ipflasso | gren | multiridge* | ecpc | squeezy |
|------------|--------|---------|----------|------|-------------|------|---------|
| Group-adaptive estimation | x      | x       | x/V^*    | v    | v           | v    | v       |
| Group parameter estimation   | NA     | Joint optimisation | CV       | Variational Bayes EM | Fast CV | Marginal moment | Marginal likelihood |
| Elastic net                   | v      | v       | v        | v    | x (only ridge) | x (only ridge) | v       |
| GLM                        | v      | v       | v        | v    | v           | v    | v       |

NOTE: *Multiridge (van de Wiel, van Nee, and Rauschenberger 2021) is not included in the data examples, but some of its computational shortcuts are used in squeezy as explained in Section 2.1. **ipflasso is able to adapt group penalties, but for a finite set of proposed values only.

two cancer genomics examples in the logistic regression setting. We include the following elastic net models, summarized in Table 1, to compare performance and computation time:

1. EN (glmnet, Friedman, Hastie, and Tibshirani 2010): a co-data agnostic elastic net penalty, with global penalty parameter obtained by cross-validation (CV);
2. fwEN and fwEN (continuous) (fwelnet, Tay et al. 2020): a globally adaptive elastic net penalty on group level for grouped co-data (fwEN) or elastic net penalty with weights a function of continuous co-data (fwEN (continuous)). Note that we include fwelnet (continuous) only in the first cancer genomics data example for which continuous co-data is available;
3. ipf and ipf2 (ipflasso, Boulesteix et al. 2017): a group-adaptive elastic net penalty, which uses cross-validation to select the group penalty factors from a grid of possible values. We take the grid where each penalty factor is in [1, 2] (ipf) or in [1, 2, 4] (ipf2). Note that we only include the computationally expensive ipf2 on the whole dataset in the model-based simulated data example for comparison of computing times;
4. gren (gren, Münch et al. 2019): a group-adaptive elastic net penalty, with the group penalty factors obtained by an approximate empirical-variational Bayes framework. Note that gren is not included in the first linear regression example, as it is not implemented for the linear model;
5. ecpcEN squeezy (ecpc, van Nee, Wessels, and van de Wiel 2021 and squeezy): a group-adaptive elastic net penalty. The method ecpc provides empirical Bayes moment estimates for group-adaptive ridge penalties. We use squeezy to transform these to elastic net penalties combined with a recalibrated global rescaling penalty;
6. squeezy (single) and squeezy (single+reCV): a co-data agnostic elastic net penalty, where the global penalty is obtained by transforming the marginal likelihood estimate for the ridge penalty to an elastic net penalty, without or with recalibration of a global rescaling penalty. While glmnet internally standardizes linear response, squeezy (single+reCV) does not;
7. squeezy (multi) and squeezy (multi+reCV): a group-adaptive elastic net penalty using the proposed method, without or with recalibration of a global rescaling penalty.

3.1. Model-based Simulation Study

We perform a model-based simulation study to compare prediction performance and computation time of several methods, and to compare group parameter estimates with known sample estimates. We show that our method improves upon co-data agnostic methods by learning from informative co-data, that is, estimating differential group priors, and scales better than other group-adaptive methods. Finally, we illustrate the benefit of learning from co-data for the purpose of variable selection in two other model-based simulated scenarios.

We simulate training and test sets of observed data \( X \) in 10 correlated blocks of size \( p/10 \), regression coefficients from a Laplace distribution \( (\alpha = 1) \) with mean zero and group-specific scale parameter \( b_{\beta k} \), \( \beta_k \in \{1, \ldots, 5\} \), for five equally sized groups, and response from a normal distribution:

\[
\begin{align*}
\mathbf{x}_{i b} & \sim N \left( 0, I(p/10) \times (p/10) + \frac{\rho^2}{(1 - \rho)^2} I(p/10 \times p/10) \right), \\
\mathbf{Y} & \sim N(\mathbf{X} \mathbf{\beta}, \sigma^2 \mathbf{I}_{n \times n}),
\end{align*}
\]

with \( \rho = 0.2 \) and \( \sigma^2 = 2 \). A similar simulation study for binary response is included in Section B.1.1 in the supplementary materials. We consider three settings of group parameters: (i) no groups: the group scale parameters \( b \) are the same and chosen such that the variance under each group prior is equal to \((0.1, 0.1, 0.1, 0.1, 0.1) \cdot \frac{1200}{p}\); (ii) weakly informative groups: the group scale parameters are different, but less so than in the third setting. The variances match \((0.141, 0.161, 0.186, 0.336, 0.536) \cdot \frac{1200}{p}\); (iii) informative groups: groups are different, with variances matching \((0.01, 0.05, 0.1, 0.4, 0.8) \cdot \frac{1200}{p}\).

3.1.1. Prediction Performance Comparison

First, we fit the methods listed above for \( \alpha = 0.3 \) on 100 independent training and test sets, with \( n = 150, p = 1200 \) and the co-data providing group membership of the \( G = 5 \) groups, to compare performance. Our method squeezy easily obtains optimal elastic net group penalties for a range of \( \alpha \). We observe that there is usually little benefit from recalibration by cross-validation (reCV) in the linear case (Figure 3).

The MSE performance on the test sets for all methods is shown for informative groups (Figure 3) and weakly informative or no groups (Figure B2, supplementary materials). Our proposed method performs well. The method squeezy (single) (either with or without reCV) outperforms the other methods in the single group setting. The difference in performance between squeezy (single+reCV) and EN is likely due to the different methods for determining the penalty, marginal likelihood versus CV. More importantly, squeezy (multi) (either with or without reCV) outperforms the other methods including squeezy (single) in the setting with informative groups, illustrating the benefit of informative co-data. The performance of squeezy (multi) and
squeezy (single) is more alike in the setting with weakly informative groups, and superior to the other methods. The method performs better when maximum marginal likelihood estimates for ridge penalties are transformed (squeezy) than when moment estimates are transformed (ecpcEN squeezy). Note that the performance of ipf might improve by using a larger grid, but only at a very substantial computational cost as shown in Figure 4. While the difference in performance between squeezy and the other methods is smaller for $\alpha = 0.8$, squeezy (multi) still outperforms the other methods when the co-data is informative (Figure B3 in the supplementary materials).

### 3.1.2. Group Parameter Estimates

For $\alpha = 1$, we may compare the empirical Bayes group parameter estimates $\lambda_g$ and $\tau_g$ obtained from ecpcEN squeezy (marginal moment estimator) and squeezy (multi) (marginal likelihood estimator) with the "true" sample estimates obtained from maximizing the likelihood and lasso prior given the sampled $X$, $Y$ and $\beta$ in each dataset:

$$
\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - X_i \beta)^2,
$$

$$
\tau_g^2 = \frac{1}{2(p/G)} \sum_{k \in \{1,...,p; |g_k| > 0\}} |\beta_k| \lambda_g^2 = \frac{\sigma^2}{\tau_g^2}.
$$

Figure B4, supplementary materials shows that squeezy estimates the group parameters well, while ecpcEN squeezy underestimates $\tau_g^2$ and overestimates $\lambda_g$. This difference may be a result from jointly optimizing $\sigma^2$ and $\lambda_R$ in squeezy, while $\lambda_R$ is estimated given a fixed estimate of $\sigma^2$ in ecpcEN squeezy.

### 3.1.3. Computation Time Comparison

Next, we fit the methods on five training sets to compare computation time for varying $n$, $p$ and $G$, for $\alpha = 0.3$. Results are similar for other $\alpha$. The group parameters are set according to the setting with informative groups. We set $(n,p,G) =
(150, 1200, 5). Then, we vary one of \( n, p, G \) while keeping the others fixed. Figure 4 shows the average computation time for fitting the models. Unsurprisingly, the group-adaptive methods are slower than the other methods, but \texttt{squeezy} scales substantially better than \texttt{ipf} in the number of groups.

3.1.4. Benefit of Co-data Learning for Variable Selection

We include two linear regression scenarios to show how variable selection may benefit from having differential group prior penalties. Suppose that \( n = 100 \) and \( p = 400 \), from which \( p^* = 24 \) “true” regression coefficients are nonzero and the rest are zero. With the first scenario we aim to show that co-data can help to discriminate the contribution of a relevant variable from a strongly correlated nonrelevant one. With the second scenario we aim to show how co-data can improve variable selection particularly for variables with weaker signal. We consider four co-data settings: (A) strong; (B) moderately strong; (C) non-informative; (D) no groups. Co-data settings A, B, and C are fit with \texttt{squeezy} and are compared to D, fit with \texttt{glmnet}, all for \( \alpha = 1 \) (lasso). Details for the simulation set up are given in Section B.1.2, supplementary materials.

Figure 5 shows the ROC curves for variable selection in both scenarios, parameterized by the global L1-penalty. Note that the models are not necessarily strictly nested, causing minor non-monotone effects. We observe clearly superior performance for \texttt{squeezy} in case of (moderately) strong co-data, and (near) equal performance to \texttt{glmnet} in case of non-informative co-data. For the first scenario, Figure B7, supplementary materials shows the trace plots for the four co-data settings. True variables come up earlier in both settings A and B, with strong and moderate co-data, respectively. Hence, variable selection performance is clearly better than for setting D, without co-data. In setting C, with nonrelevant co-data, variable selection performance is on par with that of setting D. The traceplots in Figure B8, supplementary materials highlight the 12 pairs of colinear variables for settings A and D. Overall, setting A better discriminates true from false variables. In particular, for variables 9 and 11 (first and third variable on bottom row) we observe that the no co-data setting tends to prefer the false variable, whereas \texttt{squeezy} prefers the true variable. For the second scenario, we observe that true variables come up earlier in both settings A and B, in particular for the weak, true variables (Figure 5). Hence, variable selection performance is clearly better than for setting D, without co-data. In setting C, with nonrelevant co-data, variable selection performance is on par with that of setting D.

So, variable selection improves markedly when accounting for strongly or moderately informative co-data. This improvement is particularly noticeable for true variables that are weaker either due to strong colinearity with a false variable (scenario 1), or due to a relatively small coefficient (scenario 2). It seems that the “weak” are helped by the “strong,” because they share co-data information.
3.2. Application to Predicting Therapy Response

We apply squeezy to predict therapy response (clinical benefit vs. progressed disease) using microRNA (miRNA) data from a study on colorectal cancer (Neerincx et al. 2018), consisting of $p = 2114$ miRNA expression levels for $n = 88$ independent individuals. As co-data, we use local false discovery rates (FDRs) for differential expression in the primary tumor versus adjacent colorectal tissue. The FDRs were obtained in a previous study (Neerincx et al. 2015) from a different set of nonoverlapping samples. These co-data were previously shown to be informative for the prediction, miRNAs that are tumor-specific tend to be more important for the response prediction (van Nee, Wessel, and van de Wiel 2021). Here, we discretise the continuous FDRs in $G = 8$ equally sized groups, such that we have around 10 samples per group parameter.

We fit the methods listed above on 10-fold to compare cross-validated performance (Figure 6) and average computation time (Table 2). A visual check of the QQ-plot shows that the normality assumption of squeezy is reasonable (Figure B9, Supplementary materials). Figure 6 (left) clearly shows the benefit of recalibration by cross-validation (reCV) in this logistic regression setting. Our method squeezy (multi+reCV) performs as well as $\alpha$ (Figure 6 (right), cf 10 vs. 5) but is around 25 times as fast. The method ecpcEN squeezy is twice as fast as squeezy (multi+reCV) and is competitive to squeezy (multi+reCV) and $\alpha$. These three methods all benefit from the informative co-data and outperform the other methods.

3.3. Application to Diagnostic Classification

We apply the method to methylation data with the goal of classifying samples as normal cervical tissue or precursor lesion, which has a high risk of progressing to cancer. The data consist of methylation levels for $p = 40,000$ probes in $n = 37$ samples and are extensively described in van de Wiel et al. (2016). The available co-data are biologically defined groups which are based on the distance to so-called CpG-islands. CpG islands are regions in the DNA that are often close to gene promoter regions and may therefore be more important for classification than other regions. The six groups ranging from close to far from an island are: CpG island, North Shore, South Shore, North Shelf, South Shelf, and Distant.

We fit an elastic net ($\alpha = 0.3$) logistic regression model with or without co-data and the methods described above in ten 10-fold cross-validations. A visual check of the QQ-plot shows that the normality assumption of squeezy is reasonable (Figure B10, supplementary materials). The inverse elastic net group penalties estimated with squeezy are relatively large for the CpG island group and relatively small for the Distant group (Figure 7), corroborating the biological expectations. Note that for some folds, the estimated inverse penalties are close to zero, corresponding to nearly full shrinkage of the regression coefficients. This may be due to the small sample size. Table 3 shows the results for the average computation time, AUC performance measure, number of selected variables and cross-validated log-likelihood (CVLL) for various methods. Recalibration by cross-validation (reCV) is essential for good performance of squeezy. When compared to the co-data agnostic elastic net model, the performance of our proposed method squeezy (multi+reCV) is similar in terms of AUC and slightly better in terms of the cross-validated log-likelihood. Our method squeezy (multi+reCV) is faster than the other group-adaptive methods and more memory-efficient than ipf2 and $\alpha$. The group-adaptive methods ipf2 and $\alpha$ did not run on the full data due to memory issues and have therefore been fit on a smaller, random subsample of variables on which the methods were able to run; a subset of 5000 variables for ipf2 and one of 20,000 for $\alpha$.

**Figure 6.** Results of 10-fold cross-validation in miRNA data example. Left: AUC performance for squeezy (multi) and squeezy (multi+reCV) for 100 evenly spaced values of $\alpha \in [0, 1]$. Right: AUC performance for several values of $\alpha$ for various methods.

**Table 2.** Results of 10-fold CV in miRNA data example.

| Method               | Time (s)     | Method               | Time (sec) |
|----------------------|--------------|----------------------|------------|
| 1. EN                | 1.17(1.06)   | 2. fwEN (groups)     | 12.20(6.84) |
| 7. squeezy (single)  | 5.12(3.29)   | 3. fwEN (continuous) | 13.58(13.59) |
| 6. ecpcEN squeezy    | 5.48(3.01)   | 4. ipf               | 14.11(11.13) |
| 9. squeezy (single+reCV) | 7.45(5.77)   | 5. $\alpha$          | 351.52(96.74) |
| 8. squeezy (multi)   | 10.47(10.50) | 10. squeezy (multi+reCV) | 422.39(259.80) |

NOTE: Mean time per fold and standard deviation over folds, sorted from shortest to longest time.
4. Discussion

The proposed method, termed `squeezy`, computes fast marginal likelihood estimates for group-adaptive elastic net penalties. The method estimates those penalties by deducing them from group-adaptive ridge penalties, for which we derive a low-dimensional representation of the Taylor approximation of the marginal likelihood and its first derivative for generalized linear models, using results from Wood (2011). `Squeezy` implements this and uses BFGS optimization to find the penalties. An extension would be to also use the Hessian (Wood 2011), which may speed up the optimization.

An alternative implementation of our method uses the R-package `mgcv` (Wood 2011) to estimate the ridge penalties from the linear predictors, as detailed in (van de Wiel, van Nee, and Rauschenberger 2021). As `mgcv` does not allow for more variables than samples, this solution cannot include unpenalized variables. A fix to this is to include these as pre-estimated offsets. In the absence of unpenalized variables, we found that the two implementations provided similar results at comparable computing times. The alternative solution opens up for application of `squeezy` to a wider variety of models (provided that the elastic net counter part is also available) such as the penalized Cox model for survival data.

We showed that the marginal likelihood of ridge models also approximates the marginal likelihood of elastic net models, by using the asymptotic multivariate normality of the linear predictor. This result also holds for other priors with finite variance. We provided examples for a spike-and-slab prior and generalized normal prior. Another example is to use a hierarchical prior for shrinkage or selection on group level when the co-data includes many groups. The result does not hold for priors with infinite variance, such as the highly sparse horseshoe prior (Carvalho, Polson, and Scott 2009). Moreover, the practical use of our method for other relatively sparse priors should be tested.

The method transforms the ridge penalties to elastic net penalties by using the variance function of the elastic net prior. We showed in two data examples that it is beneficial to recalculate the transformed elastic net penalties in logistic regression. Furthermore, the simulation study and data examples showed that our method is more scalable and faster than other group-adaptive methods, while it benefits from co-data and performs as well as or better than other (group-adaptive) methods. Aside from prediction, variable selection is an important aim of sparse methods. Therefore, we showed that in combination with informative co-data the use of `squeezy` benefits variable selection.

As we consider settings with high-dimensional data, the groups of variables are usually large. The approximation of

---

**Table 3.** Results of ten 10-fold cross-validations in the methylation data example.

| Method             | Time (sec)  | AUC        | # Variables | CVLL        |
|--------------------|-------------|------------|-------------|-------------|
| EN                 | 2.88±0.14   | 0.938±0.017| 182±27      | −11.34±1.50 |
| fwEN               | 74.66±24.32 | 0.935±0.015| 198±17      | −11.21±1.09 |
| ipf                | 173.12±18.50| 0.937±0.015| 60±29       | −16.38±1.04 |
| ipf2*              | 473.65±9.38 | 0.902±0.039| 59±35       | −17.13±1.80 |
| gren*              | 32.88±1.63  | 0.933±0.011| 167±19      | −11.67±0.86 |
| epcEN squeezy      | 80.77±86.99 | 0.839±0.025| 129±28      | −18.41±1.72 |
| squeezy (single)   | 12.69±9.48  | 0.608±0.039| 36±55       | −22.90±1.01 |
| squeezy (multi)    | 26.89±21.35 | 0.658±0.059| 47±63       | −21.62±1.43 |
| squeezy (single+reCV) | 16.87±8.18  | 0.938±0.016| 203±14      | −10.96±1.03 |
| squeezy (multi+reCV) | 29.30±20.99 | 0.939±0.019| 178±19      | −10.59±1.09 |

NOTE: The mean and standard deviation of the computation time, AUC, number of selected variables and cross-validated log-likelihood (CVLL) are shown for several methods. *The methods ipf2 and gren did not run due to memory issues. Results shown are based on a smaller, random subsample of variables for which the methods were able to run, 5000 for ipf2 and 20,000 for gren, whereas the other methods use all 40,000 variables.

---

**Figure 7.** Methylation data example. Estimated inverse elastic net group penalties in all folds in ten 10-fold cross-validations of the data.
the prior on the linear predictors benefits from this, as the multivariate normal result holds asymptotically in the number of variables. Should squeezy perform inferior to other methods, it may be useful to check whether this might be due to invalidity of the multivariate normal assumption for \( \eta = X \beta \), for which we included a posterior check. This check may be used too for data settings in which \( n \) is of the same order of magnitude as \( p \). Our approach may be used in these settings, but will have less benefit than in a high-dimensional setting, as the likelihood will be relatively more informative compared to the prior. For the low-dimensional setting in which \( n > p \), our approach cannot be used as the assumption on the rank of \( X \) in Theorem 1 is then violated.

As our method is based on (marginal) likelihood, it in principle facilitates model selection in terms of the hyper-parameters (e.g., single group vs. multi-group) by the use of information criteria (Greven and Kneib 2010). To what extent which criterion will be useful in our setting is a topic for further research.

The software is available in the R-package squeezy on CRAN. We provide scripts demonstrating the package and reproducing the results on https://github.com/Mirrelijn/squeezy.

Supplementary Materials

Appendix The online Appendix contains Supplemental Section A with details of the method and Supplemental Section B with additional results and figures to the data examples (Appendix_squeezy.pdf, pdf file).

Code This supplemental zip-folder contains the code for reproducing all results, explained in the corresponding README.txt files (Code_squeezy.zip, zip file).

Acknowledgments

The authors would like to thank Lodewyk Wessels and Soufiane Mourragui (Netherlands Cancer Institute) for the many worthwhile discussions.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The author Mirrelijn M. van Nee is supported by ZonMw TOP grant COMPUTE CANCER (40-00812-98-16012).

References

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017), “IPF-LASSO: Integrative-Penalized Regression with Penalty Factors for Prediction based on Multi-Omics Data,” Computational and Mathematical Methods in Medicine, 2017, 7691937. [950,955]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009), “Handling Sparsity via the Horseshoe,” in Artificial Intelligence and Statistics, PMLR. [959]

Eicker, F. (1966), “A Multivariate Central Limit Theorem for Random Linear Vector Forms,” Annals of Mathematical Statistics, 37, 1825–1828. [952]

Ferkingstad, E., and Rue, H. (2015), “Improving the INLA Approach for Approximate Bayesian Inference for Latent Gaussian Models,” Electronic Journal of Statistics, 9, 2706–2731. [954]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization Paths for Generalized Linear Models via Coordinate Descent,” Journal of Statistical Software, 33, 1–22. [955]

Greven, S., and Kneib, T. (2010), “On the Behaviour of Marginal and Conditional AIC in Linear Mixed Models,” Biometrika, 97, 773–789. [960]

Huang, J., Breheny, P., and Ma, S. (2012), “A Selective Review of Group Selection in High-Dimensional Models,” Statistical Science, 27, 481–499. [950]

Jacob, L., Obozinski, G., and Vert, J.-P. (2009), “Group Lasso with Overlap and Graph Lasso,” in ICML, ACM. [954]

Kuss, M., Rasmussen, C. E., and Herbrich, R. (2005), “Assessing Approximate Inference for Binary Gaussian Process Classification,” Journal of Machine Learning Research, 6, 1679–1704. [954]

Li, Q., and Lin, N. (2010), “The Bayesian Elastic Net,” Bayesian Analysis, 5, 151–170. [950,951]

McCullagh, P., and Nelder, J. (1989), Generalized Linear Models II, London: Chapman and Hall. [951]

Meier, L., van de Geer, S., and Bühlmann, P. (2008), “The Group Lasso for Logistic Regression,” Journal of the Royal Statistical Society, Series B, 70, 53–71. [950]

Münch, M. M., Peeters, C. E., van der Vaart, A. W., and van de Wiel, M. A. (2019), “Adaptive Group-Regularized Logistic Elastic Net Regression,” Biostatistics, 22, 723–737. [950,955]

Neerincx, M., Poel, D., Sie, D. L., van Grieken, N. C., Shankararaiah, R. C., Van Der Wolf-De, F. S., van Waesbergh, J. H. T., Burggraaf, J.-D., Eijk, P. P., Verhoef, C., Ylstra, B., Meijer, G. A., van de Wiel, M. A., Buffart, T. E., and Verheul, H. M. W. (2018), “Combination of a Six microRNA Expression Profile with Four Clinicopathological Factors for Response Prediction of Systemic Treatment in Patients with Advanced Colorectal Cancer,” PloS One, 13, e0201809. [958]

Rue, H., Martino, S., and Chopin, N. (2009), “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations,” Journal of the Royal Statistical Society, Series B, 71, 319–392. [954]

Tang, Z., Shen, Y., Li, Y., Zhang, X., Wen, J., Qian, C., Zhuang, W., Shi, X., and Yi, N. (2018), “Group Spike-and-Slab Lasso Generalized Linear Models for Disease Prediction and Associated Genes Detection by Incorporating Pathway Information,” Bioinformatics, 34, 901–910. [950]

Tay, J. K., Aghaeepour, N., Hastie, T., and Tibshirani, R. (2020), “Feature-Weighted Elastic Net: Using ‘Features of Features’ for Better Prediction,” arXiv preprint arXiv:2006.01395. [950,955]

van de Wiel, M., Lien, T., Verlaat, W., van Wieringen, W., and Witting, S. (2016), “Better Prediction by Use of Co-data: Adaptive Group-Regularized Ridge Regression,” Statistics in Medicine, 35, 368–381. [950,958]

van de Wiel, M. A., Te Beest, D. E., and Münch, M. M. (2019), “Learning From a Lot: Empirical Bayes for High-Dimensional Model-based Prediction,” Scandinavian Journal of Statistics, 46, 2–25. [950,951,954]

van de Wiel, M. A., van Nee, M. M., and Rauschenberger, A. (2021), “Fast Cross-Validation for Multi-Penalty High-Dimensional Ridge Regression,” Journal of Computational and Graphical Statistics, 30, 835–847. [951,952,955,959]

van Nee, M. M., Wessels, L. F., and van de Wiel, M. A. (2021), “Flexible Co-data Learning for High-dimensional Prediction,” Statistics in Medicine, 40, 5910–5925. [951,955,958]

Veerman, J. R., Leday, G. G., and van de Wiel, M. A. (2019), “Estimation of Variance Components, Heritability and the Ridge Penalty in High-dimensional Generalized Linear Models,” Communications in Statistics, 51, 116–134. [952,954]

Velten, B., and Huber, W. (2019), “Adaptive Penalization in High-Dimensional Regression and Classification with External Covariates using Variational Bayes,” Biostatistics, 22, 348–364. [951]

Wood, S. N. (2011), “Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models,” Journal of the Royal Statistical Society, Series B, 73, 3–36. [951,952,959]

Zou, H., and Hastie, T. (2005), “Regularization and Variable Selection via the Elastic Net,” Journal of the Royal Statistical Society, Series B, 67, 301–320. [950]