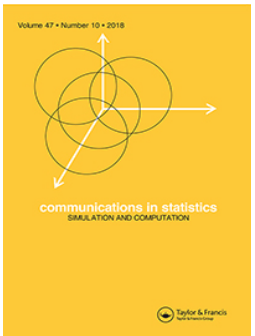


Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Estimation of variance components, heritability and the ridge penalty in high-dimensional generalized linear models

Jurre R. Veerman, Gwenaël G. R. Leday & Mark A. van de Wiel

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1646760

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Estimation of variance components, heritability and the ridge penalty in high-dimensional generalized linear models

Jurre R. Veerman[a,b], Gwenaël G. R. Leday[c], and Mark A. van de Wiel[a,c]

[a]Departement of Epidemiology & Biostatistics, Amsterdam Public Health research institute, Amsterdam University medical centers, Amsterdam, The Netherlands; [b]Mathematical Institute, Leiden University, Leiden, the Netherlands; [c]MRC Biostatistics Unit, Cambridge University, Cambridge, UK

ABSTRACT
For high-dimensional linear regression models, we review and compare several estimators of variances τ^2 and σ^2 of the random slopes and errors, respectively. These variances relate directly to ridge regression penalty λ and heritability index h^2, often used in genetics. Several estimators of these, either based on cross-validation (CV) or maximum marginal likelihood (MML), are also discussed. The comparisons include several cases of the high-dimensional covariate matrix such as multi-collinear covariates and data-derived ones. Moreover, we study robustness against model misspecifications such as sparse instead of dense effects and non-Gaussian errors. An example on weight gain data with genomic covariates confirms the good performance of MML compared to CV. Several extensions are presented. First, to the high-dimensional linear mixed effects model, with REML as an alternative to MML. Second, to the conjugate Bayesian setting, shown to be a good alternative. Third, and most prominently, to generalized linear models for which we derive a computationally efficient MML estimator by re-writing the marginal likelihood as an n-dimensional integral. For Poisson and Binomial ridge regression, we demonstrate the superior accuracy of the resulting MML estimator of λ as compared to CV. Software is provided to enable reproduction of all results.

ARTICLE HISTORY
Received 8 April 2019
Accepted 17 July 2019

KEYWORDS
Random effects; Ridge regression; Empirical Bayes; Marginal likelihood; Cross-validation

1. Introduction

Estimation of hyper-parameters is an essential part of fitting high-dimensional Gaussian random effect regression models, also known as ridge regression. These models are widely applied in genomics and genetics applications, where often the number of variables p is much larger than the number of samples n, i.e. $p \gg n$.

We initially focus on the linear model. The goal is to estimate error variance σ^2 and random effects variance τ^2 or functions thereof, in particular the ridge penalty parameter, $\lambda = \frac{\sigma^2}{\tau^2}$, or heritability index, $h^2 = \frac{p\tau^2}{p\tau^2+\sigma^2}$. Here, the ridge penalty is used in classical ridge

CONTACT Mark A. van de Wiel mark.vdwiel@amsterdamumc.nl Departement of Epidemiology & Biostatistics, Amsterdam Public Health research institute, Amsterdam University medical centers, Amsterdam, The Netherlands.

Supplemental data for this article can be accessed here.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

regression to shrink the regression coefficients towards zero (Hoerl and Kennard 1970), whereas heritability measures the fraction of variation between individuals within a population that is due to their genotypes (Visscher, Hill, and Wray 2008). The estimators of σ^2 and τ^2 can be used to estimate λ or h^2, or for statistical testing (Kang et al. 2008). We review several estimators, based on maximum marginal likelihood (MML), moment equations, (generalized) cross-validation, dimension reduction, and for degrees-of-freedom adjustment. Some of these estimators are classical, while others have recently been introduced.

We systematically review and compare the estimators in a broad variety of high-dimensional settings. For estimation of λ in *low-dimensional* settings, we refer to Muniz and Kibria (2009); Månsson and Shukur (2011); Kibria and Banik (2016). We address the effect of multicollinearity and robustness against model misspecifications, such as sparsity and non-Gaussian errors. The comparisons are extended to the linear mixed effects model, with $q \ll n$ fixed effects added to the model and to Bayesian linear regression. The linear model part is concluded by a genomics data application to weight gain prediction after kidney transplantation.

The observed good performance of MML-estimation in the linear model setting was a stimulus to consider MML for high-dimensional generalized linear models (GLM). MML is more involved here than in the linear model, because of the non-conjugacy of the likelihood and prior. Therefore, approximations are required, such as Laplace ones. While these have been addressed by others (Heisterkam, van Houwelingen, and Downs 1999; Wood 2011), we derive an estimator which is computationally efficient for $p \gg n$ settings. For Poisson and Binomial ridge regression, we demonstrate the superior accuracy of MML estimation of λ as compared to cross-validation.

Our software enables reproduction of all results. In addition, it allows comparisons for one's own high-dimensional data matrix by simulating the response conditional on this matrix, as we do for two cancer genomics examples. Computational shortcuts and considerations are discussed throughout the paper, and detailed at the end, including computing times.

1.1. The model

We initially focus on high-dimensional linear regression with random effects. Variables are denoted by $j = 1, \ldots, p$ and samples by $i = 1, \ldots, n$. Then:

$$\begin{aligned} y_{n\times 1} &= X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \\ \beta_{p\times 1} &\sim \mathcal{N}(0, \tau^2 \mathrm{I}_p) \\ \epsilon_{n\times 1} &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n). \end{aligned} \tag{1}$$

Here, $y = (y_1, \ldots, y_n)$ is the vector of responses, $\beta = (\beta_1, \ldots, \beta_p)^T$ corresponds to the random effects and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is a vector of Gaussian errors. Furthermore, $\boldsymbol{X}$ is a fixed $n \times p$ matrix: $(X_1 \cdots X_n)^T$, with $X_i = (x_{i1}, \ldots, x_{ip})^T$.

1.2. Estimation methods

We distinguish three categories of estimation methods:

1. Estimation of functions of (σ^2, τ^2), in particular: i) $\lambda = \frac{\sigma^2}{\tau^2}$ (Golub, Heath, and Wahba 1979), used in ridge regression to minimize $||y - X\beta||_2^2 + \lambda||\beta||_2^2$; and ii) heritability $h^2 = \frac{p\tau^2}{p\tau^2+\sigma^2}$ (Bonnet, Gassiat, and Lévy-Leduc 2015).

2. Separate estimation of σ^2 (Cule, Vineis, and De Iorio 2011; Cule and De Iorio 2012), possibly followed by plug-in estimation of τ^2.
3. Joint estimation of σ^2 and τ^2.

Below, we discuss several methods for each of these categories. They have several matrices and matrix computations in common, which we therefore introduce first.

1.3. Notation and matrix computations

Throughout the paper, we will use the following notation:

$$\begin{aligned}&\hat{\beta}=\hat{\beta}_\lambda=\mathrm{C}_\lambda y=\left(X^TX+\lambda I_{p\times p}\right)^{-1}X^Ty \text{ i.e. the linear ridge estimator}\\&H=H_\lambda=X\mathrm{C}_\lambda=X\left(X^TX+\lambda I_{p\times p}\right)^{-1}X^T \text{ i.e. the hat matrix.}\end{aligned}\tag{2}$$

Many of the estimators below require calculations on potentially very large matrices. The following two well-known equalities can highly alleviate the computational burden.

First, $\mathrm{C}=C_\lambda$, and hence also $\hat{\beta}$ and $\boldsymbol{H}$, can be efficiently computed by using singular value decomposition (SVD). Decompose $X=U_{n\times n}D_{n\times n}(V_{p\times n})^T$ by SVD, and denote $\boldsymbol{\Lambda}_q=\lambda\mathrm{I}_q$. Then,

$$C=\left(X^TX+\boldsymbol{\Lambda}_p\right)^{-1}X^T=V\left(D^2+\boldsymbol{\Lambda}_n\right)^{-1}DU^T.\tag{3}$$

The latter requires inversion of an $n\times n$ matrix only. Second, the following efficient trace computation for matrix products applies to $\mathrm{tr}(H)=\mathrm{tr}(XC_\lambda)$:

$$\mathrm{tr}(A_{p\times n}B_{n\times p})=\sum_{i=1}^{n}\sum_{j=1}^{p}\left[A^\circ B^T\right]_{ij}.\tag{4}$$

2. Methods

2.1. Estimating functions of σ^2 and τ^2

2.1.1. Estimating λ by K-fold CV

A benchmark method that is used extensively to estimate $\lambda=\sigma^2/\tau^2$ is cross-validation. Here, we use K-fold CV, as implemented in the popular R-package glmnet (Friedman, Hastie, and Tibshirani 2010). Let $f(i)$ denote the set of samples left out for testing at the same fold as sample i. Then, CV-based estimation of λ pertains to minimizing the cross-validated prediction error:

$$\lambda_{cv}=\arg\min_\lambda\left\{\sum_{i=1}^{n}\left(y_i-X_i\hat{\beta}_\lambda^{-f(i)}\right)^2\right\},\tag{5}$$

where $\hat{\beta}_\lambda^{-f(i)}$ denotes the estimate of β based on training samples $\{1,\dots,n\}\setminus f(i)$ and penalty λ. Note that for leave-one-out-cross-validation (n-fold CV) the analytical solution of (5) is the PRESS statistic (Allen 1974).

2.1.2. Estimating λ by generalized cross validation

Generalized Cross Validation (GCV) is a rotation-invariant form of the PRESS statistic. It is more robust than the latter to (near-diagonal) hat matrices H_λ (Golub, Heath, and Wahba 1979). For the linear model, the criterion is (Hastie, Tibshirani, and Friedman 2008):

$$\mathrm{GCV}(\lambda) = \sum_{i=1}^{n} \left(\frac{y_i - X_i^T \hat{\beta}_\lambda}{n - \mathrm{tr}(H_\lambda)} \right)^2 \tag{6}$$

where the trace of H_λ can be computed efficiently by (4). Then, $\lambda_{\mathrm{gcv}} = \arg\min_\lambda \mathrm{GCV}(\lambda)$.

2.1.3. Estimating heritability by HiLMM

Heritability is defined by $h^2 = \frac{p\tau^2}{p\tau^2+\sigma^2}$. A recent method which estimates heritability directly using maximum likelihood is proposed by Bonnet, Gassiat, and Lévy-Leduc (2015). Analogously to Eq. (12), it is based on writing:

$$y \sim \mathcal{N}\left(0, h^2\sigma^{*2}\mathrm{R} + (1-h^2)\sigma^{*2}\mathrm{I}_n\right), \tag{7}$$

where $\sigma^{*2} = p\tau^2 + \sigma^2$ and $R = XX^T/p$. Now, apply an eigen-decomposition to $\boldsymbol{R}$: $R = QLQ^T$. Then, heritability is estimated by Bonnet, Gassiat, and Lévy-Leduc (2015):

$$h^2 = \underset{h^2}{\arg\max} \left(-\log\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{y}_i^2}{h^2(\ell_i - 1) + 1} \right) - \frac{1}{n} \sum_{i=1}^{n} \left(\log\left(h^2(\ell_i - 1) + 1 \right) \right) \right), \tag{8}$$

with ℓ_i and $\tilde{y}_i$ the ith element of $\boldsymbol{L}$ and $\tilde{y} = Q^T y$, respectively. The authors provide rigorous consistency results for their estimator, as well as theoretical confidence bounds, also for mixed models and sparse settings.

2.2. Estimation of σ^2

The two methods below rely on an estimate $\hat{\beta} = \hat{\beta}_\lambda$, where $\lambda = \sigma^2/\tau^2$ is estimated by (G)CV. Then σ^2 is estimated conditional on $\hat{\beta}$. If desired, τ^2 may then be estimated by $\hat{\tau}^2 = \hat{\sigma}^2/\hat{\lambda}$.

2.2.1. Basic estimate

A basic estimate of σ^2, and often used in practice, is given by (Hastie and Tibshirani 1990):

$$\hat{\sigma}^2 = \frac{\left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right)}{\nu} \tag{9}$$

which is the residual mean square error. Here, the residual effective degrees of freedom (Hastie and Tibshirani 1990) equals $\nu = n - \mathrm{tr}(2H - HH^T)$, with $\boldsymbol{H}$ as in (2). We also considered (9) with $\nu = n - \mathrm{tr}(H)$, as in Hellton and Hjort (2018), which rendered similar, slightly inferior results.

2.2.2. PCR-based estimate

The estimator for σ^2 may also be based on Principal Component Regression (PCR). PCR is based on the eigen-decomposition $X^TX = \tilde{Q}D^2\tilde{Q}^T$. Denoting $Z = X\tilde{Q}$ and $\alpha = \tilde{Q}^T\beta$, we have $y = Z\alpha + \epsilon$. Then, $\boldsymbol{Z}$ is reduced from p columns to $r \leq \min(n, p)$ principal components, a crucial step (Cule and De Iorio 2012). Using the reduced model, σ^2 is estimated by the residual mean square error (Cule and De Iorio 2012):

$$\hat{\sigma}_r^2 = \frac{(y - Z_r\hat{\alpha}_r)^T(y - Z_r\hat{\alpha}_r)}{n - r}. \tag{10}$$

2.3. Joint estimation of σ^2 and τ^2

2.3.1. MML

An Empirical Bayes estimate of σ^2 and τ^2 is obtained by maximizing the marginal likelihood (MML), also referred to as model evidence in machine learning (Murphy 2012). This corresponds to:

$$\arg\max_{\sigma^2,\tau^2} P(y) = \arg\max_{\sigma^2,\tau^2} \int_\beta \ell(y; \beta, \sigma^2)\pi(\beta; \tau^2)d\beta. \tag{11}$$

Since $y = X\beta + \epsilon, P(y)$ is simply derived from the convolution of Gaussian random variables, implying $E[y] = E[X\beta] + E[\epsilon] = 0$, and $V[y] = V[X\beta] + V[\epsilon] = XX^T\tau^2 + \sigma^2 I_n$, so

$$P(y) = \mathcal{N}(y; \mu = 0, \boldsymbol{\Sigma} = XX^T\tau^2 + \sigma^2 I_n). \tag{12}$$

This is easily maximized over σ^2 and τ^2. Note that after computing XX^T (12) requires operations on $n \times n$ matrices only.

2.3.2. Method of moments (MoM)

An alternative to MML is to match the empirical second moments of $\boldsymbol{y}$ to their theoretical counterparts. From (12) we observe that the covariances depend on τ^2 only. Hence, we obtain an estimator of τ^2 by equating the sum of y_iy_k to that of the theoretical covariances, $\boldsymbol{\Sigma}_{ik} = \mathbb{E}[y_iy_k]$, with $\boldsymbol{\Sigma}$ as in (12). Then, with $\boldsymbol{\Sigma}^X = XX^T$, an estimator for σ^2 is obtained by substituting $\hat{\tau}^2$ and equating the sum of y_i^2 to the sum of theoretical variances, $\boldsymbol{\Sigma}_{ii} = \mathbb{E}[y_i^2]$:

$$\begin{aligned}\hat{\tau}^2 &= \frac{\sum_{i \neq k}^{n,n} y_iy_k}{\sum_{i \neq k}^{n,n} \boldsymbol{\Sigma}_{ik}^X} \\ \hat{\sigma}^2 &= n^{-1}\sum_{i=1}^{n}(y_i^2 - \hat{\tau}^2\boldsymbol{\Sigma}_{ii}^X).\end{aligned} \tag{13}$$

These equations also hold for non-Gaussian error terms, which could be an advantage over MML. Moreover, no optimization over σ^2 and τ^2 is required, so MoM is computationally very attractive.

3. Comparisons

For the linear random effects model (ridge regression) we study the following settings:

- β and ϵ generated from model (1), independent $\boldsymbol{X}$
- β or ϵ generated from non-Gaussian distributions, independent $\boldsymbol{X}$
- β and ϵ from model (1), multicollinear $\boldsymbol{X}$
- β and ϵ from model (1), data-based $\boldsymbol{X}$.

As is common for real data, the variables, i.e. the rows of $\boldsymbol{X}$, were always standardized for the L_2-penalty to have the same effect on all variables. All the results are based on 100 simulated data sets. Cross-validation is applied on 10 folds. Results from n-fold CV (leave-one-out) were generally fairly similar. We focus on the high-dimensional setting with $n = 100, p = 1000$, with excursions to larger data sets and dimensions of real data. In all visualizations below the red dotted lines indicate true values. Moreover, values larger than 20 times the true value were truncated and slightly jittered. Discussion of all results is postponed to Sec. 3.4.

3.1. Independent X

In correspondence to model (1) we sample:

$$\begin{aligned} y_{n\times 1} &= X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \qquad & \epsilon_i &\overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2) \\ x_{ij} &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) & \beta_j &\overset{\text{iid}}{\sim} \mathcal{N}(0,\tau^2). \end{aligned} \tag{14}$$

Figure 1a and b display the results for $n = 100, p = 1000, \tau^2 = 0.01, \sigma^2 = 10$ and for a large data setting $n = 1000, p = 15000, \tau^2 = 0.01, \sigma^2 = 150$ (which both imply $h^2 = 0.5$).

3.2. Departures from a normal effect size distribution

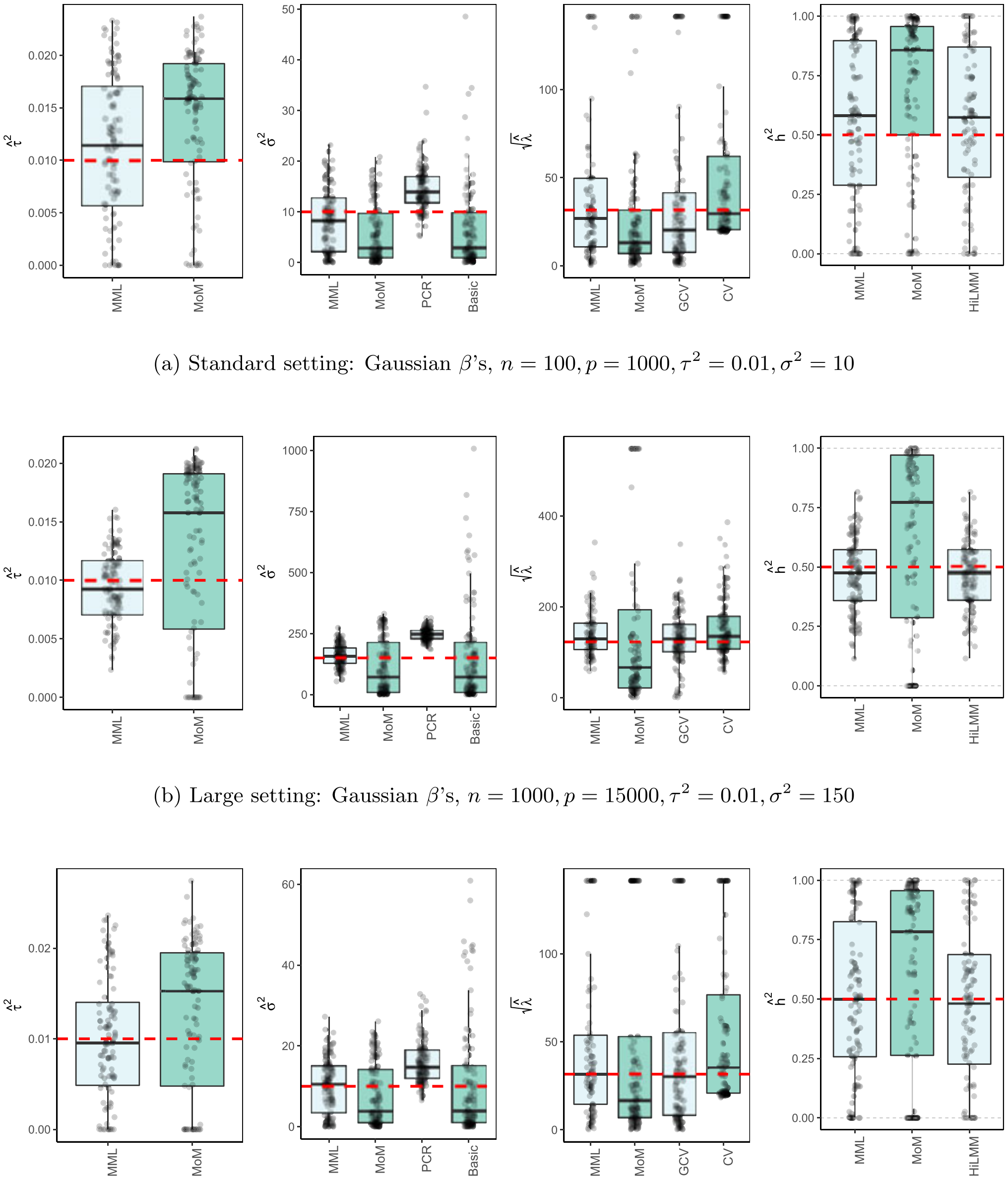
We study the robustness of the methods against (sparse) non-Gaussian effect size distribution or error distribution. In sparse settings, many variables do not have an effect. To mimic this, we simulated the β's from a mixture distribution with a 'spike' and a Gaussian 'slab':

$$\beta_j \overset{\text{iid}}{\sim} p_0\delta_0 + (1-p_0)\mathcal{N}(0,\tau_0^2). \tag{15}$$

Here, we set $p_0 = 0.9, \tau_0^2 = 0.1$, which implies $\tau^2 = \mathbb{V}(\beta_j) = \mathbb{E}(\beta_j^2) - \mathbb{E}(\beta_j)^2 = (1-p_0)\tau_0^2 = 0.01$, as in the Gaussian β_j setting. Moreover, we also considered:

$$\beta_j \overset{\text{iid}}{\sim} \text{ Laplace}(\mu = 0, b = 0.0707) \quad \text{and} \quad \beta_j \overset{\text{iid}}{\sim} \text{ Uniform}(a = -0.17, b = 0.17)$$

where again the parameters are chosen such that $E(\beta_j) = 0$ and $\tau^2 = \mathbb{V}(\beta_j) = 0.01$. Apart from β all other quantities are simulated as in (14). Results are displayed for $\sigma^2 = 10, \tau^2 = 0.01, n = 100, p = 1000$ in Figure 1c for the Laplace (= lasso) effect size distribution and in Supplementary Figure 3 for the spike-and-slab and uniform effect size distribution.



(a) Standard setting: Gaussian β's, $n = 100, p = 1000, \tau^2 = 0.01, \sigma^2 = 10$

(b) Large setting: Gaussian β's, $n = 1000, p = 15000, \tau^2 = 0.01, \sigma^2 = 150$

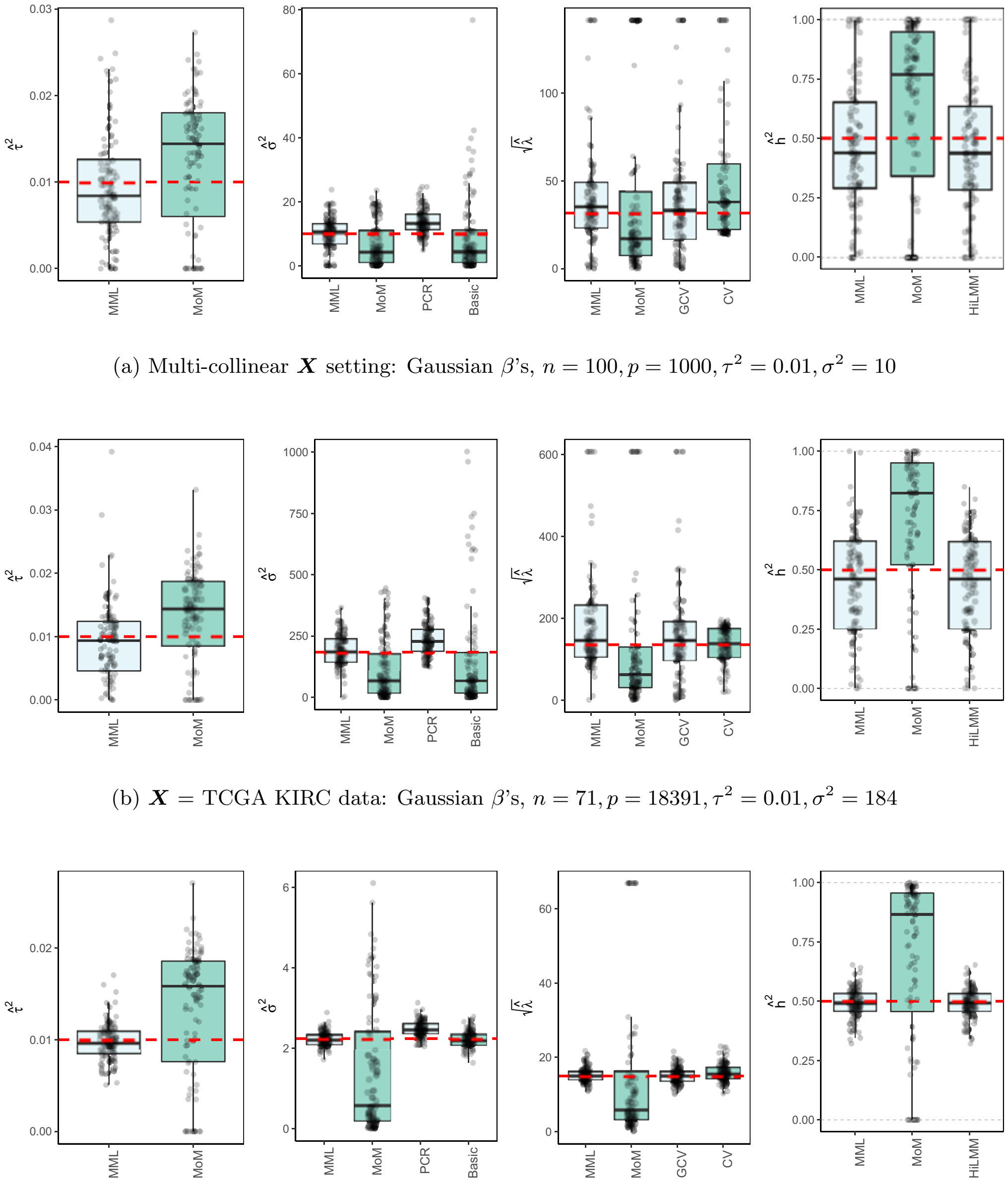
(c) Lasso setting: Laplace β's, $n = 100, p = 1000, \tau^2 = 0.01, \sigma^2 = 10$

Figure 1. Results for independent.

Moreover, we considered heavy-tailed errors by sampling

$$\epsilon_i' \overset{\text{iid}}{\sim} \mathrm{t}_4 \quad \epsilon_i = (10/2)^{1/2}\epsilon_i'$$

where the scalar is chosen such that $\sigma^2 = \mathbb{V}(\epsilon_i) = 10$, as in the Gaussian error setting. Apart from ϵ, all other quantities are simulated as in (14). Results are displayed in Supplementary Figure 3c.



(a) Multi-collinear $\boldsymbol{X}$ setting: Gaussian β's, $n = 100, p = 1000, \tau^2 = 0.01, \sigma^2 = 10$

(b) $\boldsymbol{X}$ = TCGA KIRC data: Gaussian β's, $n = 71, p = 18391, \tau^2 = 0.01, \sigma^2 = 184$

(c) $\boldsymbol{X}$ = TCPA OV data: Gaussian β's, $n = 408, p = 224, \tau^2 = 0.01, \sigma^2 = 2.24$

Figure 2. Results for multi-collinear and real.

3.3. Multicollinear X

3.3.1. Simulated X

Next, the design matrix $\boldsymbol{X}$ is sampled using block-wise correlation. We replace the sampling of $\boldsymbol{X}$ in simulation model (14) by:

$$X_{n\times p} \sim \mathcal{N}(0, \Xi), \tag{16}$$

where $\boldsymbol{\Xi}$ is a unit variance covariance matrix with blocks of size $p^* \ll p$ with correlations ρ on the off-diagonal. Figure 2a shows the results for $\rho = 0.5,\ p^* = 10, n = 100, p = 1000$.

3.3.2. Real data X

Finally, we consider the estimation of τ^2 and σ^2 in a high- and medium-dimensional setting where $\boldsymbol{X}$ are real data, with likely collinear columns. The first data set (TCGA KIRC) concerns gene expression data of $p = 18,\ 391$ genes for $n = 71$ kidney tumors. The second data set (TCPA OV) holds expression data of $p = 224$ proteins for $n = 408$ ovarian tumor samples. Details on both data sets are supplied in the Supplementary Information. To generate response $\boldsymbol{y}$ we use model (14) with $\boldsymbol{X}$ given by the data. Here, $\tau^2 = 0.01$ and σ^2 is set such that $h^2 = 0.5$. Figure 2b and c show the results.

3.4. Discussion of results

3.4.1. MML vs MoM, basic and PCR

Figures 1 and 2 and Supplementary Figure 3 clearly show superior performance of MML compared to MoM: both the bias and variability are much smaller for MML. Generally, MML also outperforms the Basic and PCR estimators of σ^2. The PCR estimator approaches the performance of MML for the KIRC and TCPA data (Figure 2b and c), and the Basic estimator performs reasonably well for the latter ($p < n$) data set. For other settings, the Basic estimator performs equally inferior as MoM. The results highlight the importance of joint estimation of σ^2 and τ^2 in high-dimensional settings, because of their delicate interplay.

3.4.2. MML vs GCV and CV

For the estimation of λ MML seems slightly superior to GCV and CV. GCV shows more estimates that deviate towards too small values of λ (e.g. Figures 1b and 2b, i.e. the large p settings), whereas CV tends to render somewhat more skewed results, either to the right (Figures 1a and c, 2a), or to the left (Figure 2b). For the spike-and-slab and uniform effects sizes and the t_4 errors the right-skewness of the CV-results is more pronounced (Supplementary Figure 3), indicating that minimization of the cross-validated prediction error (5) is more vulnerable to non-Gaussian $\boldsymbol{y}$ than MML and GCV. Note that the Laplace setting (Figure 1c) relates directly to the lasso prior with scale parameter $1/\lambda_1$ (Tibshirani 1996). The results indicate that MML with Gaussian prior could be useful to find the lasso penalty, or serve as a fast initial estimate by simply setting the lasso penalty $\lambda_1 = \sqrt{2}/\hat{\tau}$, which follows from the variance of the lasso prior.

3.4.3. MML vs HiLMM

For the estimation of heritability h^2 Figures 1 and 2 and Supplementary Figure 3 show very comparable performance of MML and HiLMM. This similar performance is not surprising given that both methods are likelihood-based. Hence, while reparametrizing the likelihood (7) is certainly useful to study it as function of h^2 (Bonnet, Gassiat, and Lévy-Leduc 2015), the reparametrization seems not beneficial for the purpose of estimating h^2. In addition, unlike HiLMM, MML also returns estimates of τ^2 and σ^2. Finally, comparing Figure 1a and b we observe that both MML and HiLMM clearly benefit from the larger n and p.

4. Data example

We re-analyse the weight gain data, recently discussed in Hellton and Hjort (2018). Details on the data are presented there, we provide a summary. The data consists of expression profiles of $n = 26$ individuals with kidney transplants, where profiles consists of 28,869 genes as measured by Affymetrix Human Gene 1.0 ST arrays. The data is available in the EMBL-EBI ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-GEOD-33070. It is known that kidney transplantation may lead to weight gain, and the study by Cashion et al. (2013) investigates whether gene expression can be used to predict this. Such a prediction can be used to decide upon additional measures to prevent excessive weight gains. We reproduced the analysis by Hellton and Hjort (2018) as much as possible, including their prior selection of 1000 genes. Details on minor discrepancies, and an alternative analysis that accounts for the gene selection are discussed in the Supplementary Material. These did not affect the comparison qualitatively.

In Hellton and Hjort (2018), the authors illustrate their focused ridge (fridge) method and compare it with conventional ridge. In short, fridge estimates *sample-specific* ridge penalties, based on minimizing a per sample mean squared error (MSE) criterion on the level of the linear predictor $X_i\beta$. Since β is not known, it is replaced by an initial ridge estimate, $\hat{\beta}_\lambda$. Their sample specific penalty then depends on X_i, and also on both $\hat{\lambda}$ and $\hat{\sigma}^2$. The authors use GCV (6) to obtain λ, and a slight variation of (9) to estimate σ^2. They show that fridge improves upon GCV-based ridge estimation. We wish to investigate whether i) MML estimation of $\lambda = \sigma^2/\tau^2$ also improves the performance of GCV-based ridge regression; and ii) whether MML estimation further boosts the performance of the fridge estimator. Here, predictive performance is measured by the mean squared prediction error (MSPE) using leave-one-out cross-validation (loocv).

The estimates of MML differ markedly from those of GCV: $(\hat{\lambda}_{\text{MML}}, \hat{\sigma}^2_{\text{MML}}) = (0.77, 0.59)$, while $(\hat{\lambda}_{\text{GCV}}, \hat{\sigma}^2_{\text{GCV}}) = (20.92, 8.08)$. Using $\hat{\lambda}_{\text{MML}}$ instead of $\hat{\lambda}_{\text{GCV}}$ for the estimation of β substantially reduced the mean squared prediction error: $\text{MSPE}_{\text{MML}} = 14.40$, while $\text{MSPE}_{\text{GCV}} = 16.38$, a relative decrease of 12.1%. Using $\hat{\lambda}_{\text{GCV}}$, as in Hellton and Hjort (2018), fridge also reduced the MSPE, but to a lesser extent: $\text{MSPE}_{\text{fridge}} = 15.80$, a relative decrease of 3.5% with respect to MSPE_{GCV}. Application of fridge using $\hat{\lambda}_{\text{MML}}$ did not further decrease MSPE_{MML}, nor did it increase it. Possibly, the already fairly small value of $\hat{\lambda}_{\text{MML}}$ left little room for improvement. Figure 3 displays absolute prediction errors per sample and illustrates the improved prediction by ridge using λ_{MML} (and to a lesser extent by fridge) with respect to ridge using λ_{GCV}.

5. Extensions

5.1. Extension 1: mixed effects model

A natural extension of the high-dimensional random effects model (1) is the mixed effects model:

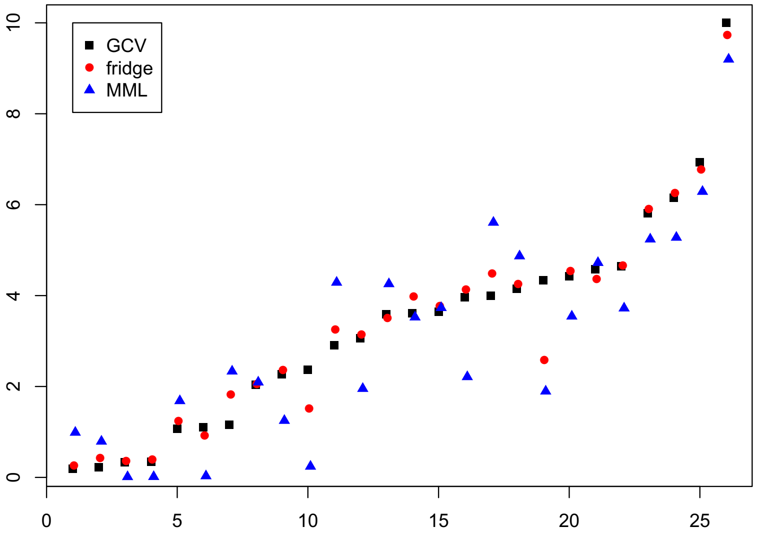


Figure 3. Absolute prediction errors (obtained by loocv; y-axis) for ridge using λ_{GCV}, for fridge and for ridge using λ_{MML}. Sample indices (x-axis) are sorted by GCV results.

$$y = X_{\text{f}}\alpha + X_{\text{r}}\beta + \epsilon, \tag{17}$$

where we assume that the $n \times m$ design matrix for the fixed effects, X_{f}, is of low-rank, so $m \ll n$, as opposed to the random effects design matrix X_{r}. Restricted maximum likelihood (REML) deals with the fixed effects by contrasting them out. For the error contrast vector $y - X_{\text{f}}\hat{\alpha}^{\text{OLS}} = \text{A}^T y$, with $\text{A} = \text{I}_n - X_{\text{f}}(X_{\text{f}}^T X_{\text{f}})^{-1} X_{\text{f}}^T$, the marginal likelihood for the variance components equals (see e.g. Zhang 2015):

$$P\left(\text{A}^T y\right) = \mathcal{N}\left(y; \mu = 0, \boldsymbol{\Sigma} = \text{A}^T \boldsymbol{\Sigma}_r \text{A}\right) \tag{18}$$

with $\boldsymbol{\Sigma}_r = X_r X_r^T \tau^2 + \sigma^2 \text{I}_n$. In addition to maximizing (18) as a function of (σ^2, τ^2), we attempted solving the set of two estimation equations suggested by Jiang (2007), but this rendered instable results inferior to maximizing (18) directly.

Alternatively, MML may be used, but it has to be adjusted to also estimate the fixed effects in the model. This implies replacing $\mathbf{0}$ in Gaussian likelihood (11) by $X_{\text{f}}\alpha$, and optimizing (11) with respect to $2 + m$ parameters, where m is the number of fixed parameters. The mixed model simulation setting is as follows:

$$\begin{array}{ll} y_{n\times 1} = X_{\text{f},n\times m}\alpha_{m\times 1} + X_{\text{r},n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} & \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \\ x_{f,ik} \overset{\text{iid}}{\sim} \mathcal{N}(0,1) & x_{r,ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\ \alpha_k \overset{\text{iid}}{\sim} p_{0,f}\delta_0 + (1-p_{0,f})\mathcal{N}\left(0, \tau_{0,f}^2\right) & \beta_j \overset{\text{iid}}{\sim} p_0\delta_0 + (1-p_0)\mathcal{N}\left(0, \tau_0^2\right), \end{array} \tag{19}$$

where $n = 100, p = 1000, m = 10, p_0 = 0.9, \tau_0^2 = 0.1$ (implying variance $\tau^2 = (1-p_0)\tau_0^2 = 0.01$ for generating random effects) and $p_{0,f} = 0.5, \tau_{0,f}^2 = 0.20$ (implying variance $\tau_f^2 = 0.1$ for generating fixed effects). Note that we focused on a fairly sparse setting for the random effects and larger prior variance of fixed effects than of random effects, which enables a stronger impact of the small number of fixed effects. Figure 4

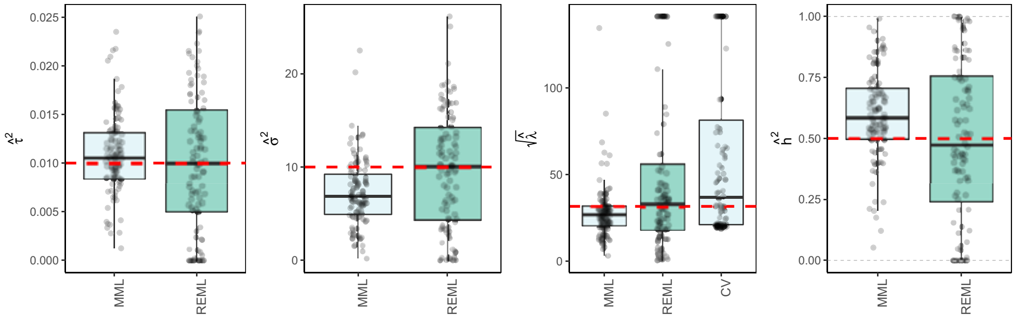


Figure 4. Estimates for mixed effects model, $\tau^2 = 0.01, \sigma^2 = 10, n = 100, m = 10, p = 1000$.

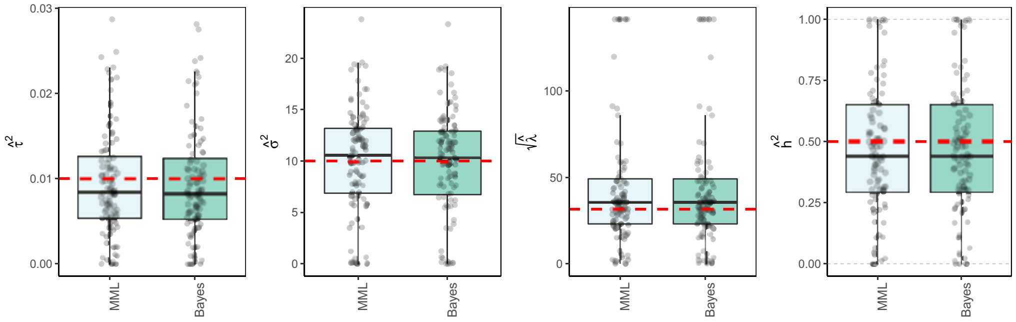


Figure 5. Bayes and MML (12) estimates for multi-collinear X, with $\tau^2 = 0.01, \sigma^2 = 10, n = 100, p = 1000$.

shows the results of REML, MML and CV (by glmnet, using penalty factor 0 for the fixed effects) for the estimation of $\tau^2, \sigma^2, \lambda$ and h^2.

From Figure 4 we observe that REML indeed improves MML in terms of bias, however at the cost of increased variability. For the estimation of λ, CV is fairly competitive to REML and MML, although it renders markedly more over-penalization.

5.2. Extension 2: Bayesian linear regression

So far, we focused on classical methods. Bayesian methods may be a good alternative. We applied the standard Bayesian linear regression model, i.e. the conjugate model with i.i.d. priors $\pi(\beta_j) = N(0, \sigma^2\tau^2)$, with τ^2 fixed and σ^2 endowed with a vague inverse-gamma prior (see Supplementary Material for details). For this model the maximum marginal likelihood estimator for τ^2 is still analytical (Karabatsos 2018), and so is the posterior mode estimate of σ^2. Figure 5 shows the results in comparison to MML, i.e. maximization of (12), for the random effects case with multi-collinear $\boldsymbol{X}$, as in Sec. 3.3.1. Results for other settings were in essence very similar.

From the results we conclude that the conjugate Bayes estimates are very close to those of MML. This is in line with the fact that both estimators maximize a marginal likelihood and the conjugate model with prior variance $\tau^2 = \sigma^2/\lambda$ is known to render posterior mean estimates of β that equal the λ-penalized ridge regression estimates.

The conjugate Bayesian model is scale-invariant, because the β prior contains the error variance σ^2. Recently, it was criticized for its non-robustness against misspecification of the fixed τ^2 when estimating σ^2 (Moran, Rockova, and George 2018). However, in practice one needs to estimate τ^2 by either empirical Bayes (e.g. maximum marginal likelihood) or full Bayes. We repeated the simulation by Moran, Rockova, and George (2018) (see Supplementary Material). The results show that the estimates of σ^2 are much better when estimating τ^2 by empirical Bayes instead of fixing it, and in fact very competitive to alternatives proposed by Moran, Rockova, and George (2018).

5.3. Extension 3: generalized linear models

5.3.1. Setting

Motivated by the good results for MML in the linear setting, we wish to extend MML estimation to the high-dimensional generalized linear model (GLM) setting, where the likelihood depends on the regression parameter β only via the linear predictor, $X\beta$. Hence, likelihood $\mathcal{L}(\mathrm{Y};\beta,X)$ is defined by a density $f_\mu(\mathrm{Y})$ (e.g. Poisson), where $X\beta$ is mapped to μ by a link function (e.g. log). As before, we a priori assume i.i.d. $\beta_j \sim N(0,\tau^2)$, here equivalent to an L_2 penalty $\lambda = 1/\tau^2$ when estimating β by penalized likelihood. In Heisterkam, van Houwelingen, and Downs (1999) an iterative algorithm to estimate λ is derived which alternates estimation of β by maximization w.r.t. λ, requiring the computation of the trace of a Hessian of a $p \times p$ matrix. Here, the estimation of β itself is much slower than in the linear case, because it is not analytic and requires iterative weighted least squares approximation. Below we show how to substantially alleviate the computational burden in the $p \gg n$ setting by re-parameterizing the marginal likelihood implying computations in $\mathbb{R}^n$ instead of $\mathbb{R}^p$.

5.3.2. Method

We have for the marginal likelihood:

$$\mathrm{ML}(\lambda) = \int_{\beta\in\mathbb{R}^p} \mathcal{L}(\mathrm{Y};\beta,X)\pi_\lambda(\beta)d\beta = \int_{\beta\in\mathbb{R}^p} \mathcal{L}(\mathrm{Y};\beta,X)\phi(\beta_1;0,1/\lambda)\cdots\phi\left(\beta_p;0,1/\lambda\right)d\beta \tag{20}$$

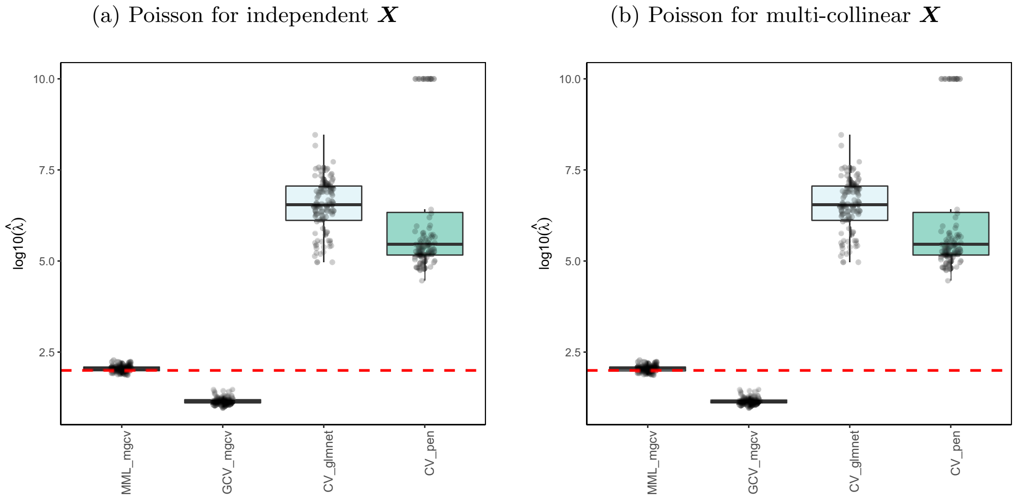
where $\phi(\beta,\mu,\tau^2)$ denotes the normal density with mean μ and variance τ^2. Now a crucial observation is that for GLM:

$$\mathrm{ML}(\lambda) = E_{\pi_\lambda(\beta)}[\mathcal{L}(\mathrm{Y};\beta,X)] = E_{\pi_\lambda(\beta)}[\mathcal{L}(\mathrm{Y};X\beta)] = E_{\pi'_\lambda(X\beta)}[\mathcal{L}(\mathrm{Y};X\beta)] \tag{21}$$

because the likelihood depends on β only via the linear predictor $X\beta$. Here, $\pi'_\lambda(X\beta)$ is the implied n-dimensional prior distribution of $X\beta$. This is a multivariate normal: $\phi(\beta^X;\mu=0,\Sigma_\lambda = XX^T/\lambda)$. Therefore, we have:

$$\begin{aligned}\mathrm{ML}(\lambda) &= \int_{\beta\in\mathbb{R}^p} g_{\mathrm{Y},\lambda}(\beta)d\beta = \int_{\beta\in\mathbb{R}^p}\mathcal{L}(\mathrm{Y};\beta,X)\phi(\beta_1;0,1/\lambda)\cdots\phi\left(\beta_p;0,1/\lambda\right)d\beta \\ &= \int_{\beta^X\in\mathbb{R}^n} h_{\mathrm{Y},\lambda}\left(\beta^X\right)d\beta^X = \int_{\beta^X\in\mathbb{R}^n}\mathcal{L}\left(\mathrm{Y};\beta^X,\mathbf{I}_n\right)\phi\left(\beta^X;0,\Sigma_\lambda\right)d\beta^X.\end{aligned} \tag{22}$$

Hence, the p-dimensional integral may be replaced by an n-dimensional one, with obvious computational advantages when $p \gg n$. Moreover, the use of (22) allows


(a) Poisson for independent $\boldsymbol{X}$

(b) Poisson for multi-collinear $\boldsymbol{X}$

Figure 6. λ estimates for Poisson ridge regression, $\lambda = 1/\tau^2 = 100, n = 100, p = 1000$.

applying implemented Laplace approximations, which tend to be more accurate in lower dimensions. The Laplace approximation requires $\hat{\beta}^X = \arg\max_{\beta^X}\{h_{Y,\lambda}(\beta^X)\}$. We emphasize that this does generally not equal $X\hat{\beta}$, where $\hat{\beta} = \arg\max_{\beta}\{g_{Y,\lambda}(\beta)\}$: the maximum of the commonly used L_2 penalized (log)-likelihood. However, $\hat{\beta}^X$ can be computed by noting that

$$\log h_{Y,\lambda}(\beta^X) \propto \ell(Y;\beta^X,\mathbf{I}_n) - (\beta^X)^T\Sigma_\lambda^{-1}\beta^X. \tag{23}$$

In other words, this is the penalized log-likelihood when regressing $\boldsymbol{Y}$ on the identity design matrix $\mathbf{I}_n$ using an L_2 smoothing penalty matrix $(\beta^X)^T\Sigma_\lambda^{-1}\beta^X = \lambda(\beta^X)^T(XX^T)^{-1}\beta^X$. The latter fits conveniently into the set-up of Wood (2011), as implemented in the R-package mgcv. This also facilitates MML estimation of λ by maximizing ML(λ), with $h_{Y,\lambda}(\beta^X)$ as in (23). If the columns of $\boldsymbol{X}$ are standardized (common in high-dimensional studies), XX^T has rank n – 1 instead of n, implying that $(XX^T)^{-1}$ does not exist and should be replaced by a pseudo-inverse $(XX^T)^+$, such as the Moore-Penrose inverse.

In a full Bayesian linear model setting, dimension reduction is also discussed by Bernardo et al. (2003), where $X\beta$ is substituted by a n-dimensional factor analytic representation, which requires an SVD of $\boldsymbol{X}$. In addition, there it is not used for hyper-parameter estimation by marginal likelihood, but instead for specifying (hierarchical) priors for the factors.

5.3.3. Results

R packages like glmnet (Friedman, Hastie, and Tibshirani 2010) and penalized (Goeman 2010) estimate λ by cross-validation, and also mgcv allows, next to the MML estimation, (generalized) CV estimation (Wood 2011). Figure 6a and b show the results for Poisson ridge regression, with $Y_i \sim \text{Pois}(\lambda_i), \lambda_i = \exp(X_i\beta), \beta$ generated as in (14), and $\boldsymbol{X}$ generated as in (14) and (16), which denote the independent $\boldsymbol{X}$ and multi-collinear $\boldsymbol{X}$ setting, respectively.

Figure 6 clearly shows the superior performance of MML based on (22) over CV. In particular, glmnet and penalized render strongly upward biased values. The mgcv GCV values are still inferior to MML based ones, but much better than the latter two, which may be due to the different regression estimators used (Laplace approximation versus iterative weighted least squares). We should stress that CV does not target for the estimation of λ as such, but merely for minimizing prediction error. Nevertheless, the difference is remarkably larger than in the corresponding linear case (see Figures 1 and 2).

The Supplementary Material shows the results for Binomial ridge regression. While the differences in performance are less dramatic than for the Poisson setting, MML still renders much better estimates of λ than CV-based approaches.

6. Computational aspects and software

All methods and simulations presented here are implemented in a few wrapper R scripts: one for the linear random effects model (which includes the conjugate Bayes estimator), one for the linear mixed effects model, and one for Poisson and Binomial ridge regression. Parallel computations are supported. The scripts allow exact reproduction of the results in this manuscript as well as comparisons for other simulation or user-specific real data $\boldsymbol{X}$ cases. In addition, a script is supplied to produce the box-plots as in this manuscript.

HiLMM, PCR and CV implementations are provided by the R-packages HiLMM, v1.1 (Bonnet, Gassiat, and Lévy-Leduc 2015), ridge, v1.8-16 (Cule and De Iorio 2012) (code slightly adapted for computational efficiency) and glmnet, v2.0-16 (Friedman, Hastie, and Tibshirani 2010). The methods MML, REML, Bayes, MoM, Basic and GCV were implemented by us for the linear random and mixed effects models. For Poisson and Binomial ridge regression we applied mgcv, v1.8-16 (Wood 2011) after our re-parametrization (22) to obtain MML and GCV results, while for CV glmnet and penalized, v0.9-50 (Goeman 2010) were applied. For all methods that required optimization the R routine optim was used, with default settings. CV was based on 10 folds.

Computing times of the various methods largely depend on n and p, much less so on the exact simulation setting. These are displayed for $n = 100$, 500 and $p = 10^3, 10^4, 10^5$ in Table 1, based on computations with one CPU of an Intel®$Xeon^{®} CPUE5-2660v3@2.60\,GHz$ server. For Poisson ridge regression, we only report the computing

Table 1. Computing times for hyper-parameter estimation for linear and Poisson ridge regression.

	$n = 100$			$n = 500$		
	$p = 10^3$	$p = 10^4$	$p = 10^5$	$p = 10^3$	$p = 10^4$	$p = 10^5$
Linear						
MML	0.06	0.15	1.12	2.18	6.07	26.64
Bayes	0.04	0.31	4.38	1.10	7.78	93.25
MoM	0.01	0.08	1.03	0.17	2.32	23.70
PCR	0.05	0.39	5.36	1.39	10.31	116.80
Basic	0.05	0.46	6.56	1.44	12.40	145.18
GCV	0.20	0.46	4.56	12.26	26.41	111.38
CV	0.81	6.57	39.95	2.62	21.69	183.50
HiLMM	0.03	0.17	2.01	0.66	3.14	27.99
Poisson						
MML_mgcv	0.32	0.33	0.31	26.21	40.19	48.17
GCV_mgcv	0.39	0.33	0.62	33.48	41.44	54.01

times of MML and GCV, because, as reported in Figure 6, the performance of CV-based methods was very inferior.

From Table 1 we conclude that MML is also computationally very attractive. Its efficiency is explained by the fact that, unlike many of other methods, it does not require an SVD or other matrix decomposition of $\boldsymbol{X}$. Moreover, the only computation that involves dimension p is the product XX^T.

7. Discussion

We compared several estimators in a large variety of high-dimensional settings. The results showed that plain maximum marginal likelihood works well in many settings. MML is generally superior to methods that aim to separately estimate σ^2 (9, 10). Apparently, the estimates of σ^2 and τ^2 are so intrinsically linked in the high-dimensional setting that separate estimation is sub-optimal. The moment estimator (MoM) is generally not competitive to MML. It may, however, be useful in large systems with multiple hyper-parameters to estimate *relative* penalties, which are less sensitive to scaling issues than the global penalty parameter (Van de Wiel et al. 2016). MoM may also be a useful initial estimator for more complex estimators that are based on optimization, such as MML.

Possibly somewhat surprising is the good performance of MML for estimating λ and h^2, as these are functions of σ^2 and τ^2. For the estimation of λ it is generally better than or competitive to (generalized) CV, an observation also made for the low-dimensional setting (Wood 2011). The inferior performance of the basic estimator of σ^2 (9) implies that alternative estimators of λ that use $\hat{\sigma}^2$ as a plug-in are unlikely to perform well in high-dimensional settings. Such estimators, including the original one by Hoerl and Kennard (1970), are compared by Muniz and Kibria (2009); Kibria and Banik (2016), who show that some do perform well in the *low-dimensional* setting. For Poisson ridge regression, similar estimators of λ are available (Månsson and Shukur 2011), but these rely on an initial maximum likelihood estimator of β, and hence do not apply to the high-dimensional setting. For estimating h^2 it should be noticed that HiLMM (Bonnet, Gassiat, and Lévy-Leduc 2015) aims to compute a confidence interval for h^2 as well. For that purpose their direct estimator (8) is likely more useful than MML on the pair (τ^2, σ^2). We also used Esther (Bonnet et al. 2018), which precedes HiLMM by sure independence screening. It did not improve HiLMM in our (semi-)sparse settings, and requires manual steps. However, it likely improves HiLMM results in very sparse settings (Bonnet et al. 2018).

For mixed effect models with a small number of fixed effects, MML compares fairly well to REML, with a larger bias, but smaller variance. Probably the potential advantage of contrasting out the fixed effects is small when the number of random effects is large. REML may have a larger advantage in very sparse settings (Jiang et al. 2016) or when the number of fixed effects is large with respect to n. Estimates from the conjugate Bayes model are very similar to those by MML. We show that estimating τ^2 along with σ^2 highly improves the σ^2 estimates presented by Moran, Rockova, and George (2018), where a fixed value of τ^2 is used. In the case of many variance components or multiple similar regression equations, Bayesian extensions that shrink the estimates by a common prior are appealing, in particular in combination with efficient posterior approximations such as variational Bayes (Leday et al. 2017).

Our model (1) implies a dense setting, but we have demonstrated that the MML and REML estimators of τ^2 and σ^2 are fairly robust against moderate sparsity, which corroborates the results by Jiang et al. (2016). Nevertheless, true sparse models may be preferable when variable selection is desired, which depends on accurate estimation of β. On the other hand, post-hoc selection procedures can be rather competitive (Bondell and Reich 2012). Moreover, the sparsity assumption is questionable for several applications. For example in genetics, it was suggested that many complex traits (such as height or cholesterol levels) are not even polygenic, but instead 'omnigenic' (Boyle, Li, and Pritchard 2017).

The extension of MML to high-dimensional GLM settings (22) is promising given its computational efficiency and performance for Poisson and Binomial regression. A special case of the latter, logistic regression, requires further research, because the Laplace approximations of the marginal likelihood are less accurate here (Wood 2011). Extension to survival is a promising avenue, because Cox regression is directly linked to Poisson regression (Cai and Betensky 2003). Alternatively, parametric survival models may be pursued. To what extent the estimates of hyper-parameters impact predictions depends on the sensitivity of the likelihood to these parameters. For the linear setting, a re-analysis of the weight-gain data showed that predictions based on $\hat{\lambda}_{\text{MML}}$ improved those based on $\hat{\lambda}_{\text{CV}}$. Karabatsos (2018) shows that MML estimation also performs well compared to GCV for linear power ridge regression, which extends ridge regression by multiplying λ by $(X^TX)^\delta$.

The MML estimator can be extended to estimation of multiple variance components or penalty parameters, which was addressed by iterative likelihood minorization (Zhou et al. 2015) and by parameter-based moment estimation (Van de Wiel et al. 2016). The latter extends to non-Gaussian response such as survival or binary. Further comparison of these methods with multi-parameter MML, both in terms of performance and computational efficiency, is left for future research. Finally, in particular in genetics applications, extensions of estimation of variance components by MML to non-independent individuals can be implemented by use of a well-structured between-individual covariance matrix $\mathbf{\Sigma}$ (Kang et al. 2008).

Although our simulations cover a fairly broad spectrum of settings, many other variations could be of interest. We therefore supply fully annotated R scripts https://github.com/markvdwiel/Hyperpar that allow i) comparison of all algorithms discussed here, also for one's 'own' real covariate set $\boldsymbol{X}$; and ii) reproduction of all results presented here.

Acknowledgment

Gwenaël Leday was supported by the Medical Research Council, grant number MR/M004421. We thank Jiming Jiang and Can Yang for their correspondence, input and software for the MM algorithm. In addition, we thank Kristoffer Hellton for providing the fridge software and data. Finally, Iuliana Ciocănea-Teodorescu is acknowledged for preparing the TCGA KIRC data.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Allen, D. 1974. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* 16 (1):125–7. doi:10.1080/00401706.1974.10489157.

Bernardo, J. M., M. J. Bayarri, J. O. Berger, and A. P. Dawid. 2003. Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian Statistics* 7:733–42.

Bondell, H., and B. Reich. 2012. Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association* 107 (500):1610–24. doi:10.1080/01621459.2012.716344.

Bonnet, A., E. Gassiat, and C. Lévy-Leduc. 2015. Heritability estimation in high dimensional sparse linear mixed models. *Electronic Journal of Statistics* 9 (2):2099–129. doi:10.1214/15-EJS1069.

Bonnet, A., C. Lévy-Leduc, E. Gassiat, R. Toro, and T. Bourgeron. 2018. Improving heritability estimation by a variable selection approach in sparse high dimensional linear mixed models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67 (4):813–39. doi:10.1111/rssc.12261.

Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. An expanded view of complex traits: From polygenic to omnigenic. *Cell* 169 (7):1177–86. doi:10.1016/j.cell.2017.05.038.

Cai, T., and R. Betensky. 2003. Hazard regression for interval-censored data with penalized spline. *Biometrics* 59 (3):570–9. doi:10.1111/1541-0420.00067.

Cashion, A., A. Stanfill, F. Thomas, L. Xu, T. Sutter, J. Eason, M. Ensell, and R. Homayouni. 2013. Expression levels of obesity-related genes are associated with weight change in kidney transplant recipients. *PLoS One* 8 (3):e59962. doi:10.1371/journal.pone.0059962.

Cule, E., and M. De Iorio. 2012. A semi-automatic method to guide the choice of ridge parameter in ridge regression. arXiv preprint, *arXiv:1205.0686*.

Cule, E., P. Vineis, and M. De Iorio. 2011. Significance testing in ridge regression for genetic data. *BMC Bioinformatics* 12:372. doi:10.1186/1471-2105-12-372.

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

Goeman, J. 2010. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 52 (1):70–84. doi:10.1002/bimj.200900028.

Golub, G. H., M. Heath, and G. Wahba. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21 (2):215–23. doi:10.1080/00401706.1979.10489751.

Hastie, T., and R. Tibshirani. 1990. *Generalized additive models*. Boca Raton, FL: CRC Press.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2008. *The elements of statistical learning*. 2nd ed. New York, NY: Springer.

Heisterkam, S. H., J. C. van Houwelingen, and A. M. Downs. 1999. Empirical Bayesian estimators for a Poisson process propagated in time. *Biometrical Journal* 41 (4):385–400. doi:10.1002/(SICI)1521-4036(199907)41:4<385::AID-BIMJ385>3.0.CO;2-Z.

Hellton, K. H., and N. L. Hjort. 2018. Fridge: Focused fine-tuning of ridge regression for personalized predictions. *Statistics in Medicine* 37 (8):1290–303. doi:10.1002/sim.7576.

Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1):55–67. doi:10.1080/00401706.1970.10488634.

Jiang, J. 2007. *Linear and generalized linear mixed models and their applications*. New York, NY: Springer Science & Business Media.

Jiang, J., C. Li, D. Paul, C. Yang, and H. Zhao. 2016. On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics* 44 (5):2127–60. doi:10.1214/15-AOS1421.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178 (3):1709–23. doi:10.1534/genetics.107.080101.

Karabatsos, G. 2018. Marginal maximum likelihood estimation methods for the tuning parameters of ridge, power ridge, and generalized ridge regression. *Communications in Statistics - Simulation and Computation* 47 (6):1632–51. doi:10.1080/03610918.2017.1321119.

Kibria, B., and S. Banik. 2016. Some ridge regression estimators and their performances. *Journal of Modern Applied Statistical Methods* 15:12.

Leday, G. G. R., M. C. M. de Gunst, G. B. Kpogbezan, A. W. van der Vaart, W. N. van Wieringen, and M. A. van de Wiel. 2017. Gene network reconstruction using global-local shrinkage priors. *The Annals of Applied Statistics* 11 (1):41–68. doi:10.1214/16-AOAS990.

Månsson, K., and G. Shukur. 2011. A Poisson ridge regression estimator. *Economic Modelling* 28 (4):1475–81. doi:10.1016/j.econmod.2011.02.030.

Moran, G. E., V. Rockova, and E. I. George. 2018. On variance estimation for Bayesian variable selection. arXiv preprint, arXiv:*1801*.03019.

Muniz, G., and B. M. G. Kibria. 2009. On some ridge regression estimators: An empirical comparisons. *Communications in Statistics - Simulation and Computation* 38 (3):621–30. doi:10.1080/03610910802592838.

Murphy, K. 2012. *Machine learning, a probabilistic perspective*. Cambridge, MA: The MIT Press.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* 58 (1):267–88.

van de Wiel, M. A., T. G. Lien, W. Verlaat, W. N. van Wieringen, and S. M. Wilting. 2016. Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine* 35 (3):368–81. doi:10.1002/sim.6732.

Visscher, P. M., W. G. Hill, and N. R. Wray. 2008. Heritability in the genomics era – Concepts and misconceptions. *Nature Reviews Genetics* 9 (4):255–66. doi:10.1038/nrg2322.

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 (Statistical Methodology)* 73 (1):3–36. doi:10.1111/j.1467-9868.2010.00749.x.

Zhang, X. 2015. A tutorial on restricted maximum likelihood estimation in linear regression and linear mixed-effects model. http://statdb1.uos.ac.kr/teaching/multi-grad/ReML.pdf.

Zhou, H., L. Hu, J. Zhou, and K. Lange. 2015. MM algorithms for variance components models. arXiv preprint, arXiv:*1509.07426*.