Selection properties of Type II maximum likelihood (empirical bayes) in linear models with individual variance components for predictors

Tahira Jamil, Cajo J. F. ter Braak*

Selection properties of Type II maximum likelihood (empirical bayes) in linear models with individual variance components for predictors

Tahira Jamil, Cajo J. F. ter Braak*

Biometris, Wageningen University and Research Centre, Box 100, 6700 AC Wageningen, the Netherlands.

* corresponding author. Tel.:+31 317 480803 Fax +31 317 483554 E-mail address: cajo.terbraak@wur.nl

This is a postprint, an Accepted Author Manuscript (AAM). Please cite the official version:

Jamil, T. and C. J. F. ter Braak. (2012). Selection properties of Type II maximum likelihood (empirical bayes) in linear models with individual variance components for predictors. Pattern Recognition Letters, 33, 1205–1212. doi: 10.1016/j.patrec.2012.01.004. http://dx.doi.org/10.1016/j.patrec.2012.01.004

Abstract. Maximum Likelihood (ML) in the linear model overfits when the number of predictors (*M*) exceeds the number of objects (*N*). One of the possible solution is the Relevance vector machine (RVM) which is a form of automatic relevance detection and has gained popularity in the pattern recognition machine learning community by the famous textbook of Bishop (2006). RVM assigns individual precisions to weights of predictors which are then estimated by maximizing the marginal likelihood (type II ML or empirical Bayes). We investigated the selection properties of RVM both analytically and by experiments in a regression setting.

We show analytically that RVM selects predictors when the absolute z-ratio (|least squares estimate|/standard error) exceeds 1 in the case of orthogonal predictors and, for $M = 2$, that this still holds true for correlated predictors when the other z-ratio is large. RVM selects the stronger of two highly correlated predictors. In experiments with real and simulated data, RVM is outcompeted by other popular regularization methods (LASSO and/or PLS) in terms of the prediction performance. We conclude that Type II ML is not the general answer in high dimensional prediction problems.

In extensions of RVM to obtain stronger selection, improper priors (based on the inverse gamma family) have been assigned to the inverse precisions (variances) with parameters estimated by penalized marginal likelihood. We critically assess this approach and suggest a proper variance prior related to the Beta distribution which gives similar selection and shrinkage properties and allows a fully Bayesian treatment.

Keywords: Automatic relevance detection; Empirical Bayes; Lasso; Sparse model; Type II maximum likelihood; Relevance vector machine

1. Introduction

Maximum likelihood (ML) or least squares (LS) can lead to severe over-fitting and poor estimation, when the number of predictors or basis functions (M) is large as compared to data size (N) *i.e.*, $M \geq$ N. Regularization or shrinkage estimation can improve an estimate and regularize an ill-posed problem (Bishop, 2006). This involves adding a penalty term to the error function in order to discourage parameters from reaching large values. In a linear model the modified error function takes the form

$$\text{RSS} + \lambda \sum_{m=1}^{M} |w_m|^q \text{ for } q \geq 0, \quad (1)$$

where RSS is the residual sum of squares, $\mathbf{w} = (w_1,.., w_M)^T$ is the parameter vector containing the weights (regression coefficients) for the predictors, and $\lambda \geq 0$ is a complexity parameter that controls the amount of regularization. For q=2 we have ridge regression (RR) (Hoerl and Kennard, 1970) which proportionally shrinks estimates of $\{w_m\}$ to zero, but does not produce a sparse solution. In neural networks this is known as weight decay. For q=1 we have the LASSO (least absolute shrinkage and selection operator) (Tibshirani, 1996) which also shrinks the coefficients towards zero but also puts some coefficients exactly to zero, and therefore performs variable selection (Efron et al., 2004; Tibshirani, 1996). The optimal choice for λ in penalized likelihood is often based on cross validation.

Most regularization methods have a Bayesian interpretation as giving the maximum a posterior (MAP) mode for a given prior distribution for the parameters. The prior in RR is Gaussian and in LASSO it is double exponential. The equivalence of MAP with the shrinkage estimate does not mean that the Bayesian framework is simply a re-interpretation of classical methods. The distinguishing element of Bayesian inference is marginalization. By marginalizing over $\mathbf{w}$ we obtain a marginal likelihood, also known as the type II likelihood or the evidence function (Bishop, 2006). The parameter λ can then be obtained by maximizing this function, *i.e.* by type II maximum likelihood, and then $\mathbf{w}$ is obtained for this value of λ. This procedure is also known as empirical Bayes and automatic relevance determination (MacKay, 1992; Neal, 1996). A fully Bayesian approach would also require a prior for the hyperparameter λ and marginalization over λ.

Tipping (2001) created the relevance vector machine (RVM) as a sparse kernel technique build upon a linear model with $M = N$. In RVM each weight w_m is assigned an independent Gaussian prior with an individual precision, resulting in M hyperparameters which are all precisions (or their inverse, variances). Tipping (2001) considered assigning a Gamma prior to the precisions, but eventually focussed on a uniform prior for which maximization of the posterior reduces to maximization of the marginal likelihood, also called the type II likelihood (Tipping, 2001; Bishop 2006). By maximizing the type II likelihood with respect to all M hyperparameters many precisions go to infinity (Faul and Tipping, 2002), so creating a sparse model as each infinite precision effectively eliminates the corresponding predictor from the model. Tipping and Faul (2003) developed a fast sequential algorithm for this. RVM has found wide-spread application with 705 citations in the Web of Science as of July 2011, also outside the kernel world (Li et al., 2002; Rogers and Girolami, 2005) and found general exposure through the exposition in Bishop (2006). However, little is known about the properties of RVM. With (hyper)parameters on the edge of the permissible region, general asymptotic theory for maximum (marginal) likelihood does not apply.

This paper studies the selection and shrinkage properties of RVM in the un-kernelised regression setting (Bishop 2006: section 7.2). We found it easier to work with variance rather than precision, because a predictor drops from the model when its variance component is zero, which is easier to work with than with infinite precision. As Bishop (2006), we phrase and study RVM outside its kernel context as a type II maximum likelihood approach to the linear model with individual variance components for the predictors. We first state the model and rewrite the marginal likelihood in a form that uses inner product matrices of size $M \times M$ rather than $N \times N$. We then obtain an analytical expression for the selection and shrinkage properties of RVM in the special case that the predictors are orthonormal and the error variance is known. The main result here is that RVM drops a predictor from the model if and only if its z-ratio (least-squares estimate of the weight divided by its standard error of estimate) is less than 1 in absolute value. RVM is thus very tolerant in allowing predictors to stay in the model. In practice, particularly in a kernel context and always when $M > \mathrm{N}$, predictors are not orthogonal and regularization methods tend to behave very different in the presence of correlation. For example, if the two predictors are highly correlated, LASSO selects one, whereas ridge, elastic net (Zou and Hastie, 2005) and PLS (Frank and Friedman, 1993) select both. Tibshirani (1996) gave analytical expressions for the two correlated predictors case for the LASSO. In section 4 we attempt similarly for RVM and arrive at analytical expressions for when RVM selects neither, one or both predictors. The main conclusion from these expressions is again that RVM is very tolerant in allowing predictors to stay in the model. In section 5 we compare RVM on simulated and real data for a range of M/N ratios with LASSO and Partial Least Squares (PLS), which is a shrinkage method based on latent variables that is very popular in chemometrics (Wold et al., 2001). We conclude with a discussion of the RVM and its extensions in relation to fully Bayesian approaches.

2. RVM as sparse Bayesian linear regression

RVM for regression is a linear model with a prior that results in a sparse solution (Bishop, 2006). The model for real-valued target variable t, given an input vector $\mathbf{x}$, takes the form

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{2}$$

where $\boldsymbol{w}$ is a vector of M parameters and ϵ is a white noise term that is Gaussian distributed with zero mean and variance σ^2, which we will assume known. The regression function $\boldsymbol{y}(\mathbf{x}, \mathbf{w})$ is then defined as the linear model

$$y(\mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) \tag{3}$$

with fixed nonlinear basis functions $\phi_m(\mathbf{x})$. For ease of presentation we ignore the constant term representing bias as it can be dealt with by centring the target variable and basis functions. For a given set of N independent observations of the target t and input vector $\mathbf{x}$, the data likelihood function of the target vector $\mathbf{t} = (t_1, \ldots, t_N)^T$ for given input vectors $\{\mathbf{x}_i\}_{i=1,\ldots,N}$ is

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi)^{-N/2} \sigma^{-N} \prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}\left(t_i - y(\mathbf{x}_i, \mathbf{w})\right)^2\right). \tag{4}$$

To make it a Bayesian model we need to specify a prior for the parameter $\boldsymbol{w}$. In RVM, each parameter w_m is an independent zero mean Gaussian with a separate variance parameter α_m, giving

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = (2\pi)^{-M/2} \prod_{m=1}^{M} \alpha_m^{-1/2} \exp\left(-\frac{w_m^2}{2\alpha_m}\right) \quad (5)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^T$ is the vector of hyperparameters, which are in our notation *not* precisions but variances. These M independent hyperparameters control the strength of the prior over its associated weight and this form of prior is responsible for the sparsity properties of the model (Tipping, 2001).

In type-II maximum likelihood (Berger, 1985), also known as empirical Bayes or evidence approximation (MacKay, 1992), an estimate $\widehat{\boldsymbol{\alpha}}$ is obtained by maximizing the marginal likelihood $p(\boldsymbol{t}|\boldsymbol{\alpha}, \sigma^2)$ over $\boldsymbol{\alpha}$, which is then plugged into posteriori density $p(\mathbf{w}|\mathbf{t}, \hat{\alpha}, \sigma^2)$, which is a multivariate normal, the mean of which is taken as the shrinkage estimate $\widetilde{\boldsymbol{w}}$. The marginal likelihood requires integration over $\boldsymbol{w}$, giving the multivariate normal density (Bishop, 2006)

$$L(\boldsymbol{\alpha}) = p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = \int_{-\infty}^{\infty} p(\mathbf{t}|\mathbf{w}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\mathbf{w} = (2\pi)^{-N/2} |\mathbf{C}|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{t}^T \mathbf{C}^{-1} \mathbf{t}\right), \quad (6)$$

where $\mathbf{C} = \sigma^2 \mathrm{I} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T$ with $\boldsymbol{\Phi}$ the $N \times M$ design matrix, of which the ith row is $(\varphi_1(\mathbf{x}_i), \ldots, \varphi_M(\mathbf{x}_i))^T$ and $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$.

Our goal is now to maximize (6) with respect to the hyperparameters $\boldsymbol{\alpha}$. At this point we deviate from Bishop (2006) and convert the inverse and determinant of the $N \times N$ matrix $\mathbf{C}$ using the matrix inversion and determinant lemma or Woodbury formula (Golub and van Loan, 1989) into forms using $M \times M$ matrices. On deleting terms that do not depend on $\boldsymbol{\alpha}$, we obtain (Appendix A)

$$L(\boldsymbol{\alpha}) \propto |\mathbf{I} + \sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{A}|^{-1/2} \exp\left(\frac{1}{2\sigma^2}\mathbf{t}^T\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sigma^2\mathbf{A}^{-1})^{-1}\boldsymbol{\Phi}^T\mathbf{t}\right). \quad (7)$$

This marginal likelihood has a form equivalent to the posterior distribution of the variance component in a hierarchical linear model or random model (O'Hagan and Forster, 2004; ter Braak, 2006). The study of the selection properties of RVM is equivalent to the study of the conditions under which hyperparameters (α-values) become zero. We do this by setting the derivative of (7) with respect to $\boldsymbol{\alpha}$ to zero, solving the resulting equation for $\boldsymbol{\alpha}$, checking that this represents a maximum and checking whether the obtained $\widehat{\boldsymbol{\alpha}}$ has some zero elements.

3. Orthonormal predictors

In this section we study the selection properties of RVM in the special case that the predictors are orthogonal, *i.e.* $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is a diagonal matrix. In this case the marginal likelihood (7) decomposes as a product of individual likelihoods $L(\alpha_m)$ with (Appendix B)

$$L(\alpha_m) \propto (1 + v_m^{-1}\alpha_m)^{-1/2} \exp\left(\frac{\widehat{w}_m^2 v_m^{-2}\alpha_m}{2(1+v_m^{-1}\alpha_m)}\right) \quad (8)$$

where $\widehat{w}_m = \boldsymbol{\phi}_m^T\mathbf{t}/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the least-squares estimate, $v_m = \sigma^2/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the variance of $\widehat{w}_m$, and $\boldsymbol{\phi}_m$ is the mth column of $\boldsymbol{\Phi}$. The variance component α_m that maximizes (8) is

$$\hat{\alpha}_m = (\widehat{w}_m^2 - v_m)_+, \quad (9)$$

where $(.)_+$ is the positive part operator, defined as $(a)_+ = a$ if $a > 0$ and 0 otherwise. In the orthogonal predictor case, RVM thus leads to soft thresholding (Donoho and Johnstone, 1994; Donoho, 1995) of the variance component, whereas LASSO does this for the weights (Tibshirani, 1996). Also observe that $\hat{\alpha}_m = 0$, iff $\widehat{w}_m^2 \leq v_m$ or, equivalently, |z-ratio| $\equiv$

$|\hat{w}_m/se(\hat{w}_m)| \leq 1$ where $se(.)$ is the standard error of estimate. The elements of the shrinkage estimate $\tilde{\boldsymbol{w}}$, for which the z-ratio in absolute value is smaller than 1, are thus zero. The corresponding predictors can thus be pruned. Fig. 1 displays the result for the case of two uncorrelated predictors.

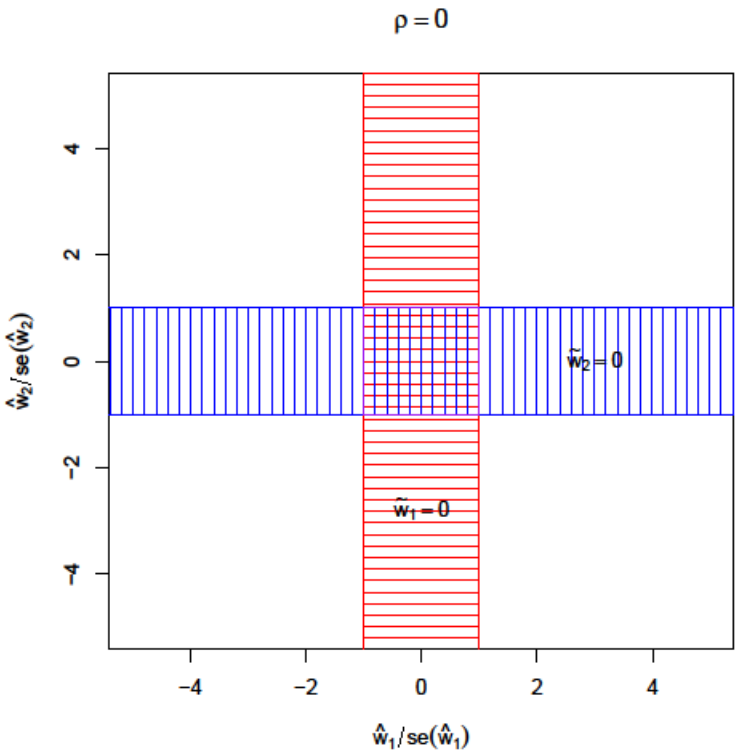

Fig. 1. Two uncorrelated predictor case: regions in terms of the z-ratio (estimate/standard error) where the RVM estimate of the weights and associated variance components are exactly zero. In these regions the corresponding predictor(s) can be pruned from the model.

4. Two correlated predictors

We now consider the case with two correlated predictors and assume they are rescaled such that $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2^T\boldsymbol{\phi}_2 = 1$ and $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_2 = \boldsymbol{\phi}_2^T\boldsymbol{\phi}_1 = \rho$. In this case, $\mathbf{A}$ is a 2×2 diagonal matrix with diagonal elements α_1 and α_2 which are linearly changed by the rescaling. The dependence of the marginal likelihood $L(\boldsymbol{\alpha})$ on σ^2 can be removed by transformation to variance ratios $\boldsymbol{\gamma} = \boldsymbol{\alpha}/\sigma^2$ and by defining $c = \boldsymbol{\phi}_1^T\mathbf{t}/\sigma$, $d = \boldsymbol{\phi}_2^T\mathbf{t}/\sigma$. The maximum is invariant under these transformations. Note that c is the simple z-ratio, that is the z-ratio in least-

squares regression with single predictor $\boldsymbol{\phi}_1$, and the same holds for d and $\boldsymbol{\phi}_2$. On using Mathematica, differentiating $L(\alpha_1, \alpha_2)$ with respect to γ_1 and γ_2 and setting the derivatives equal to zero gives (Appendix C)

$$\begin{bmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{bmatrix} = \begin{bmatrix}\dfrac{\left(c+\hat{\gamma}_2(c-\rho d)\right)^2-(1+\hat{\gamma}_2)\left(1+\hat{\gamma}_2(1-\rho^2)\right)}{(1+\hat{\gamma}_2(1-\rho^2))^2}\\ \dfrac{\left(d+\hat{\gamma}_1(d-\rho c)\right)^2-(1+\hat{\gamma}_1)\left(1+\hat{\gamma}_1(1-\rho^2)\right)}{\left(1+\hat{\gamma}_1(1-\rho^2)\right)^2}\end{bmatrix}. \tag{10}$$

This represents a maximum if $\hat{\gamma}_1 \geq 0$ and $\hat{\gamma}_2 \geq 0$. From (10), if $\hat{\gamma}_1 = 0$, then $\hat{\gamma}_2 = (d^2-1)_+$ and it should hold that

$$\left(c+\hat{\gamma}_2(c-\rho d)\right)^2 \leq (1+\hat{\gamma}_2)\left(1+\hat{\gamma}_2(1-\rho^2)\right). \tag{11}$$

Inserting $\hat{\gamma}_2 = (d^2-1)_+$ in (11) and solving for c gives upper and lower bounds

$$c = \rho d(1-d^{-2}) \pm \sqrt{1-\rho^2(1-d^{-2})}. \tag{12}$$

Subject to $\hat{\gamma}_2 > 0$, values of c within the bounds of (12) give $\hat{\gamma}_1$=0, and consequently $\widetilde{w}_1 = 0$. Interchanging the roles of $\hat{\gamma}_1$and $\hat{\gamma}_2$ and thus c and d yields similar for bounds of d such that $\hat{\gamma}_2$=0, and consequently $\widetilde{w}_2 = 0$, now subject to $\hat{\gamma}_1 > 0$. If $\hat{\gamma}_1 = 0$, then $\hat{\gamma}_2 = (d^2-1)_+$, and if $\hat{\gamma}_2$=0, then $\hat{\gamma}_1 = (c^2-1)_+$, so that both variance ratios are zero if both $|c|$ and $|d|$ are smaller than 1. The bounds are a function of *c, d* and ρ^2.

Figs 2a,b shows these bounds for ρ = 0.5 and 0.9 in the (*c*,*d*)-plane and the resulting regions where none, one or both weights are exactly zero. Figures for the corresponding negative values of ρ differ only in rotation over 90° and shading.

Whereas for ρ = 0, $\widetilde{w}_1 = 0$ if the simple z-ratio *c* is less than 1 in absolute value ($|c| < 1$), no such simple rule exists for ρ ≠ 0. The interval of *c*-values for which the first weight is exactly zero depends on *d*, as shown in Figs 2a,b. For example, for ρ = 0.9 then still $\widetilde{w}_1 = 0$ for $|c| < 1$ if *d* = 1, but if *d* = 4, then $\widetilde{w}_1 = 0$ if $2.88 < c < 3.87$ (Fig. 2b). The simple z-ratio alone thus says little about the nullity of the first weight estimate. We need both *c* and *d*.

Figs 2c,d shows the same bounds in terms of the (multiple) z-ratio's, $\widehat{w}_m/se(\widehat{w}_m)$, *m* = 1,2, *i.e.* $\widehat{\boldsymbol{w}} = (\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\boldsymbol{T}}\mathbf{y}$ and $se(\widehat{w}_m)^2$ is a diagonal element of $\sigma^2(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi})^{-1}$, so that, in the two predictor case, $z_1 = (c-\rho d)/\sqrt{1-\rho^2}$ and $z_2 = (d-\rho c)/\sqrt{1-\rho^2}$. For ρ = 0, this is the identity transformation and the result is the same as Fig. 1. With $d \to \pm\infty$ in (12), $c \to \rho d \pm \sqrt{1-\rho^2}$, so that $z_1 \to \pm 1$ (Figs 2c,d); the associated values of z_2 are $d\sqrt{1-\rho^2} \pm \rho$. For small and intermediate values of *d* the bounds are less simple. The same holds for $c \to \pm\infty$ so that $z_2 \to \pm 1$ with $z_1 = c\sqrt{1-\rho^2} \pm \rho$.

Some more insight into Fig. 2 is obtained by noting that the corners of the unit rectangle in Fig. 2a transform to the corners of the approximate trapezium in Fig. 2c; the (1,1) corner becomes (0.58,0.58), the (1,-1) corner becomes (1.73,-1.73) for ρ = 0.5, and the opposite corners (-1,-1) and (-1,1) follow by mirroring. This means that with the z-ratio pair (0.6,0.6) both variables stay in the model. So it is not even necessary that the z-ratio exceeds 1 for obtaining a non-zero type II ML estimate. For ρ = 0.9, the corners become (0.23, 0.23) and (4.36,-4.36). So, for example, the z-ratio pair (0.3,0.3) gives two non-zero type II ML estimates, but the pair (4,-4) yields two zero estimates in type II ML so pruning both predictors from the model, despite the fact that for this ρ the chi-square test-statistic of the latter point is about 9 times that of the former. This is a remarkable property of type II ML.

Note that the white upper-right and lower-left corners in Figs 2c,d come from the small white wedges in Figs 2a,b in the same corners. In Fig. 2b the wedge is very small: if the correlation among predictors is high, both predictors to stay in the model when *c* and *d* are very close or very different, or both should be very large.

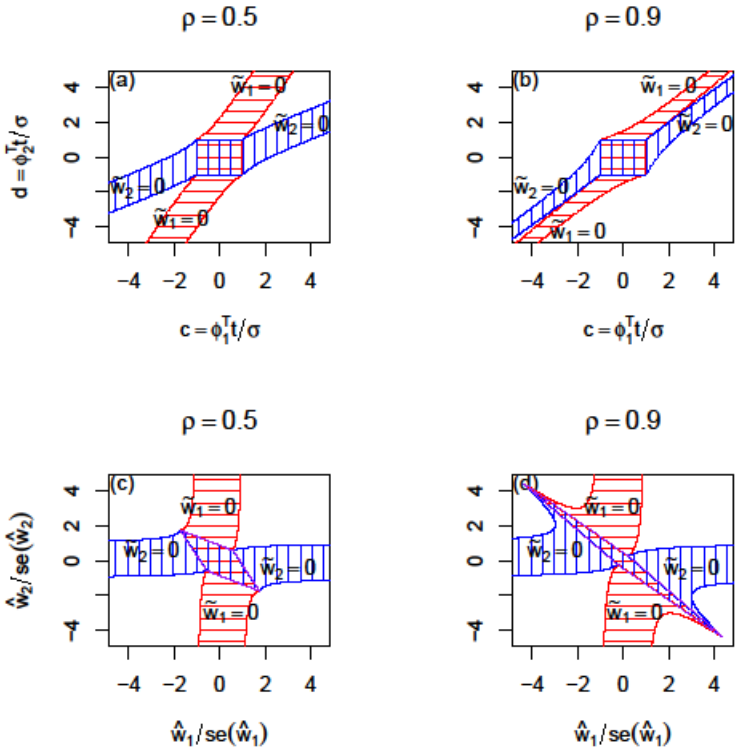

Fig. 2. Two correlated predictor case: regions in terms of the simple z-ratio (a,b) and multiple z-ratio (c,d) where the RVM estimate of the weights and associated variance ratios are exactly zero. In these regions RVM prunes the corresponding predictor(s) from the model. The simple z-ratio is based on least-squares with a single predictor, the (multiple) z-ratio on least-squares with two predictors.

In conclusion, if the (estimated least squares) effect of one predictor is very strong, the bound for the additional correlated predictor comes close to the bound for the uncorrelated case (|z-ratio| > 1 for a predictor to stay in the model). If, by contrast, neither predictors has a large effect, then type II ML prunes the one with the smallest effect. If, for positively correlated predictors, they have virtually identical estimated effects, then both predictors stay in the model, even if their z-ratio is as small as 0.6 and 0.3 for ρ = 0.5 and 0.9, respectively. However, if for positively correlated predictors, the estimated least squares effects are of opposite sign, type II ML excludes both predictors, except when the z-ratios are large.

5. Experiments

In the following we compare the performance of RVM with LASSO and PLS on simulated and real data. Computation was carried out in R (R Development Core Team, 2010) using the packages lme4 (Bates et al., 2011), glmnet (Friedman et al., 2010) and pls (Wehrens and Mevik, 2006). The kernelized version of RVM was carried out with the function rvm in the kernlab package (Karatzoglou et al., 2004). Results are for two types of kernels: RVM_{rbf} (Gaussian radial basis kernel) and RVM_{lin} (the linear or dot product kernel). A prototype statement to carry out un-kernelized RVM (by Type II ML) in lme4 with $M = 2$ is

lmer(t ~ (0 + x1 | v) + (0 + x2 | v), data=train, REML=FALSE)

where t is the target, x1 and x2 predictors, v is an all ones N-vector and train is a data frame containing these vectors. The argument REML shows that RVM could also be fitted using Residual Maximum Likelihood (Searle et al., 2008). REML estimates of variance components are generally less biased than ML estimated. In the experiments we show results from both Type II ML and REML. In RVM context the differences are expected to be small.

5.1 Simulation study

We first checked that lmer follows our theoretical analysis that RVM with orthogonal predictors sets the variance of predictors to zero if their |z-ratio| ≤ 1. For this, we generated data sets with R-package mvrnorm with M=6 orthogonal predictors and target $\mathbf{t}$ such that the z-ratio's in a least squares fit were 0.90, 0.94, 0.98, 1.02, 1.06, and 1.10. For large N (*e.g.* N = 100 and 1000), lmer followed the theory in all such data sets. For small N, the two small differences between our theory and lmer play a role. First we could not fix the error variance to 1 as we did in our theory and, secondly, we could not omit the intercept. The REML- and ML- estimates for the error variance by lmer were biased downward with, as expected, less bias in REML than in ML, and with less bias for larger N. For small N (we tested with N = 8 and 20), our theory still turned out to work for the *estimated* z-ratio, that is, the z-ratio in which the estimated error variance $\hat{\sigma}^2$ is inserted for σ^2. The variance estimates by both REML and ML were in accordance with equation (9) with $v_m = \hat{\sigma}^2/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, except in occasional cases of non-convergence. So, lmer followed the theory for orthogonal predictors also quantitatively.

Next, we simulated data where all predictors were assumed to be independent and Gaussian distributed, but not necessarily orthogonal in each particular data set due to randomness or $N < M$. We simulated data from the true model $\mathbf{t} = \boldsymbol{\Phi}\boldsymbol{w} + \sigma\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0,1)$, $\sigma = 1$ with $w_1 = 3$ and $w_m = 0$ for $m > 1$ where $m = 1, \ldots, M$, and M = 1, 5, 10, 20 and 100 predictors. The examples thus differ in the number of weights equal to zero (noise predictors). The lme4 implementation of type II ML did not allow much higher M. We set N = 20 to still get a wide range of M/N. For each example, 100 datasets were generated. For computing mean-squared error of prediction of the target (MSEP), each dataset was split into training data of N = 20 observations and test data of 1000 observations and MSEP was calculated from the test data using the weights estimated from the training data.

Table 1 and Fig. 3 summarize the results. Type II ML and REML behaved similar and almost identically to LASSO for M = 1 and 5 but behaved worse for $M \geq N$. PLS had the worst performance in all examples.

The next three examples are similar to those in Zou and Hastie (2005), where simulated 100 data sets are simulated from the model $\mathbf{t} = \boldsymbol{\Phi w} + \sigma\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(0,1)$. These examples are:

In example 1, $N = 20$, $M = 8$, $\boldsymbol{w} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the predictors are Gaussian with $\text{corr}(\phi_n, \phi_m) = \rho^{|n-m|}$ with $\rho = 0.5$. We set $\sigma = 3$ and this implies SNR ≈ 1.5.

Example 2 is the same as example 1, except that $w_m = 0.85 \ \forall \ m$ (SNR ≈ 1.3).

Table 1. Median mean-squared prediction errors for the simulations with independent predictors for different methods (100 replications). In parentheses are the corresponding standard errors (of the medians) computed via 1000 bootstrap resamples of the 100 mean squared errors. The null model used the mean for prediction.

				Type II	
M	**null model**	**LASSO**	**PLS**	**ML**	**REML**
1	10.33 (0.062)	1.09 (0.011)		1.09 (0.010)	1.09 (0.010)
5	10.31 (0.086)	1.17 (0.024)	1.32 (0.032)	1.14 (0.031)	1.14 (0.027)
10	10.45 (0.087)	1.27 (0.021)	2.04 (0.104)	1.32 (0.035)	1.31 (0.035)
20	10.32 (0.079)	1.45 (0.045)	4.97 (0.275)	2.44 (0.111)	2.38 (0.101)
100	10.39 (0.063)	1.52 (0.045)	8.81 (0.123)	2.21 (0.080)	2.14 (0.053)

In example 3, $N = 50$, $M = 40$, $w_m = 3$ for $m = 1, \ldots, 15$ and $w_m = 0$ for $m = 16, \ldots, 40$ and SNR ≈ 1.7. The first 15 predictors are three equally important groups of 5 predictors each, which are generated as follows:

$\phi_m = h_1 + \epsilon_m^{\phi}$ with $h_1 \sim \text{N}(0,1)$, $m = 1, \ldots, 5$

$\phi_m = h_2 + \epsilon_m^{\phi}$ with $h_2 \sim \text{N}(0,1)$, $m = 5, \ldots, 10$

$\phi_m = h_3 + \epsilon_m^{\phi}$ with $h_3 \sim \text{N}(0,1)$, $m = 10, \ldots, 15$

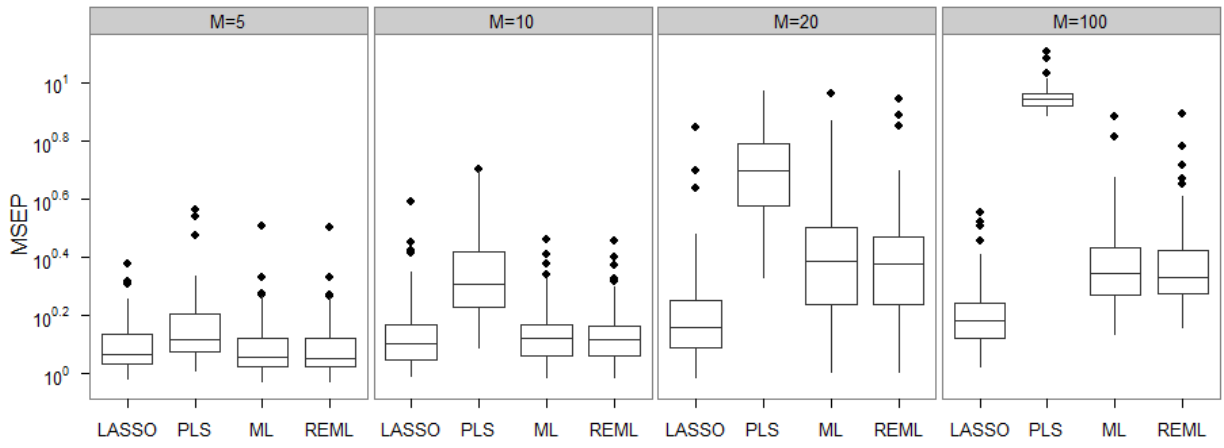

Fig. 3. Box plot of the mean-squared prediction error (MSEP) for LASSO, PLS, ML and REML of the 100 simulations with independent predictors.

and ϵ_m^{ϕ}~N(0,0.16) for $m = 1, \ldots, 15$. In this model, the pairwise correlations within groups are 0.86 and the correlations between groups are 0. The remaining 25 predictors are pure noise features.

The next four examples are from ter Braak (2009) and use a latent variable model. In these examples the target was generated from four independent standard Gaussian latent variables $h_1, \cdots, h_4$ by

$t_n = \sum_{l=1}^{4} \psi_l h_{nl} + \epsilon_n$ with ϵ_n~N$(0, \sigma^2)$

and fixed $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_4)^T$, and the predictors were generated by

$\phi_{nk}^{(l)} = \tau_{lk} h_{nl} + \epsilon_n^{\phi}$ with ϵ_{nk}^{ϕ}~N$(0, 1 - \tau_{lk}^2)$ $(n = 1, \ldots, N, l = 1, \ldots, 4, k = 1, \ldots, m_l)$,

and fixed $\{\tau_{lk}\}$,yielding predictors with unit variance. The following four examples differ in the number of predictors per latent variable (m_l) , the weights ($\boldsymbol{\psi}$, $\{\tau_{lk}\}$), and the number of noise variables added.

In Example 4, $N = 50$, $M = 75$, $\boldsymbol{\psi} = (22.9, 22.9, 22.9, 22.9)^T$ and $\sigma = 15$, so that signal to noise ratio (SNR ≡ sd(E(**t**)/σ) is 3. The first latent variable h_1 generate $m_1 = 5$ predictors with $\tau_{1k} = 0.85 \ \forall \ k$ (*S*NR= 1.6). The second, third and fourth latent variables generate 10, 20 and 40 predictors in the same way by using $q = 2$, 4 and 8 repetitions of the τ coefficients, respectively. In this setup, the population least square weights for the predictors associated with the first latent variable are **w** = (5, 5, 5, 5, 5)^T and the weights for the predictors associated with the other three latent variables are equal to ~5/q, more precisely 2.59, 1.32 and 0.67. The within-group correlations are 0.72. This example has 75 nonzero coefficients and no zero coefficients.

Example 5 is as example 4, except that 75 nuisance predictors are added.

Example 6 has $N = 50$ with M =75 predictors and target generated as in example 4 but with different $\{\tau_{lk}\}$, σ = 15 (SNR ≈3.2). For first latent variable $\{\tau_{1k}\}$ = (20,20,0,– 20,– 20) . The second latent variable generated two block of five coefficients; in each block $\{\tau_{2k}\}$ = (10,10,0,–10,–10) . The third latent variable generated four blocks of five predictors; in each block $\{\tau_{3k}\}$ = (2, 2, 0, 2, 2) and the fourth latent variable generated eight blocks of five

predictors; in each block $\{\tau_{4k}\}$ = (1, 1, 0, 1, 1). In this setup, contrasts of correlated predictors derived from the first and second latent variable are important for precise prediction (ter Braak, 2009).

Example 7 is as example 6, except that 75 pure noise features are added.

Table 2 summarizes the results based on 100 simulations of the examples. Note that the numbers in Table 2 are σ^2 higher than those in ter Braak (2009). Type II ML and REML perform comparably to LASSO and PLS in examples 1-2, but do poorly in examples 3-7. RVM_{rbf} does better than either Type II ML or REML, except in examples 1 and 2, and better than RVM_{lin}, except in examples 4 and 6. The performance of PLS is the best in all examples, except in examples 1 and 6 where LASSO dominates all.

5.2 Real data example

In this example we reconsider the barley dataset from the North American Barley Genome Mapping project to illustrate the performance of Type II maximum likelihood and LASSO (Xu, 2007). The data consists of N = 145 doubled haploid population lines of barley. The target **t** was average kernel weight. The input vector **x** was the genotype of the line, consisting of M = 127 markers. Each marker was coded as $\phi_m(\mathbf{x}) = 1$ for genotype A (TR306 allele), -1 for genotype B (Harrington allele) and 0 for missing genotype. The mean squared error of prediction, estimated using 10-fold cross-validation, was 1.62 for type-II maximum likelihood and 0.68 for LASSO (the LASSO penalty being estimated by an inner loop of cross-validation). Fig. 4 shows the estimated weights of markers for the two methods. Both the methods perform similar in terms of sign and have same direction for coefficients. The LASSO pattern of weights is more shrunken towards zero as compared to type II ML. Type II ML has thus higher peaks.

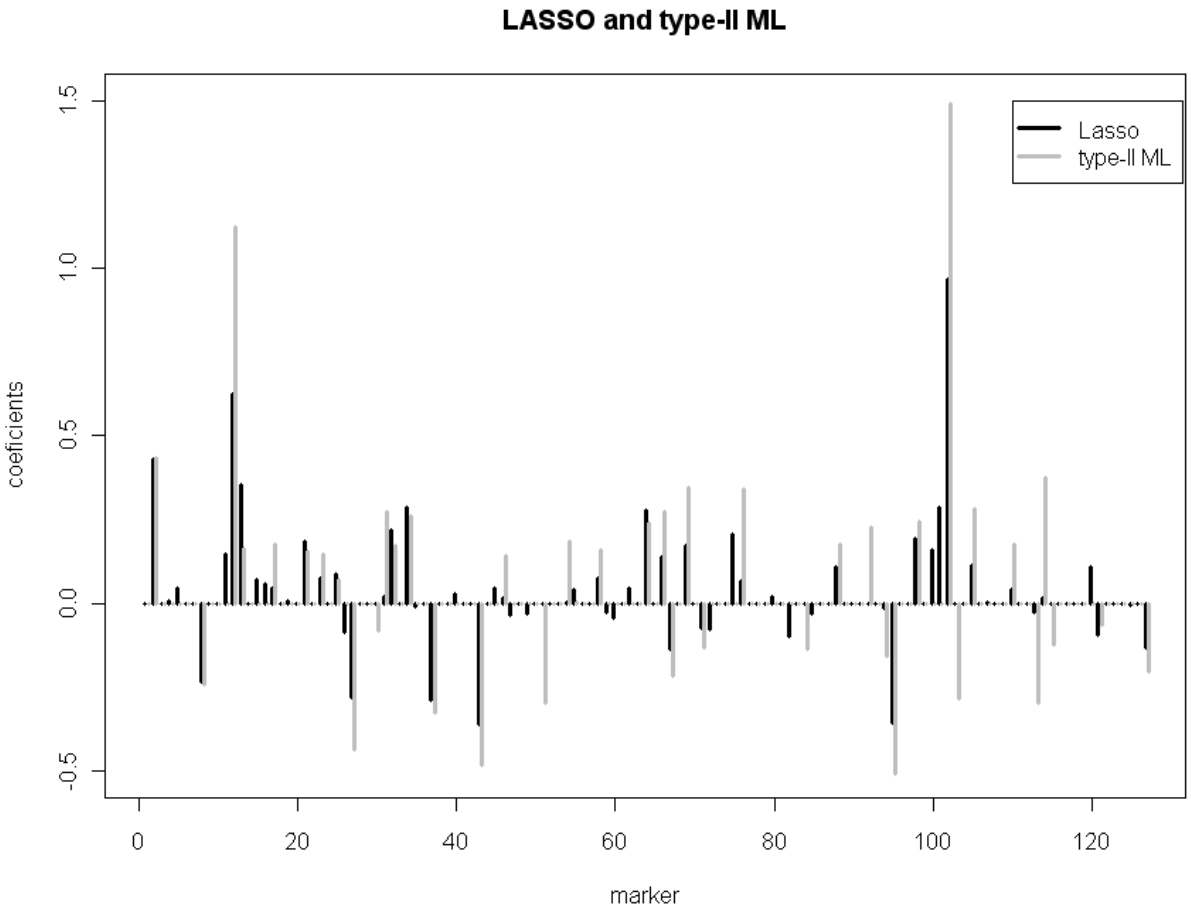

Fig. 4. Marker weights **w** estimated by type II ML and LASSO in the barley data.

6. DISCUSSION

RVM has the attractive property that it automatically selects relevant predictors. Its hyperparameters are estimated by type II ML (empirical bayes). By contrast, methods such as LASSO require crossvalidation to set the penalty hyperparameter. We showed analytically that RVM selects predictors on the basis of the least-squares z-ratio ($|z|>1$) in the case of orthogonal predictors and, for $M = 2$, that this still holds true for correlated predictors when the other z-ratio is large. We also found that RVM prunes the weaker of two highly correlated predictors. In a kernel setting, predictors are likely to be highly correlated, so RVM prunes there. In our simulated and real data, we found that RVM gave higher prediction error than LASSO.

The threshold of 1 for the z-ratio is a kind of minimum that is also implicit in the AIC criterion. For $M > N$, it appears too weak. For example, Donoho (1995) advocated pruning based on $|z| < \sqrt{2\log(M)}$ based on the idea that, for large M, the maximum of M independent standard Gaussian deviates is below this threshold with probability close to 1. More recent work proposes thresholds based on the ratio of the actual and potential model sizes (Abramovich et al., 2005). RVM does not have this property.

In line with the original ideas in Tipping (2001), Xu (2007, 2010) extended the RVM approach by adding a (hyper)prior for the variance components. With a uniform prior for the variances his approach reduces to RVM, whereas it relates to the adaptive sparseness method (Figueiredo, 2003) with a Jeffrey's prior. The prior adds a penalty to the marginal likelihood; the penalized marginal likelihood is maximized to obtain the variance components. The prior provides the means for threshold values higher than 1, although we were not yet able to show that analytically.

Penalized methods are often given additional underpinning as giving maximum a posteriori (MAP) estimates in the Bayesian framework (Zou and Hastie, 2005). In the same vein, RVM yields variance estimates that are MAP under a uniform prior for the variances $\boldsymbol{\alpha}$. But what happens in terms of precisions? The posterior density would change with a Jacobian term involving $\prod a_m^2$ that accounts for the transformation to precision and therefore the MAP would change when back-transformed to variance. By contrast, penalized methods are invariant under transformation. The Bayesian underpinning of penalized methods is thus rather thin.

This raises the question whether RVM and Xu's extensions can be thought of as approximations to a fully Bayesian model. Xu (2007, 2010) uses independent scaled inverse chi-square distributions as priors for the variances, which is equivalent to gamma distributions for the precisions. The prior for α_m is thus inverse gamma

$$p(\alpha_m|a,b) \propto \alpha_m^{-(a+1)} \exp\left(-\frac{b}{\alpha_m}\right), \tag{13}$$

which is proper for $a>0$ and $b>0$ and leads to t-priors for the weights. For obtaining more shrinkage, Xu (2007, 2010) used improper priors with $b=0$ and $-1 \leq a \leq 0$. The model is equivalent with the improper δ-prior (ter Braak, 2006; ter Braak et al., 2005).

$$p(\alpha_m|\delta) \propto \alpha_m^{\delta-1}. \tag{14}$$

Their fully Bayesian treatment showed that the posteriors for $\boldsymbol{\alpha}$ and $\mathbf{w}$ are proper if and only if $0 < \delta \leq 1/2$ or equivalently $-1/2 \leq a < 0$ with $b = 0$ in (13) and that the model gives

attractive sigmoidal shrinkage for small δ, similar in form of that of the SCAD penalty (Fan and Li, 2001). Note that the uniform prior ($\delta = 1$) for α_m (RVM) and Jeffey's prior are excluded ($\delta = 0$). The uniform for the standard deviation $\alpha_m^{1/2}$ ($\delta = 1/2$) is not excluded, but does not shrink. We conclude that the empirical Bayes approach in RVM and its extensions by Xu (2010) are not supported as approximations to a fully Bayesian approach; the fully fletched Bayesian model does not even exist for the values used for the parameter of *a* and *b*.

Parameters of priors such as (14) can no longer be estimated in a Bayesian way if they are improper. The reason is that it is impossible to add an additional level to the Bayesian model and to assign them a hyper prior so as to obtain the posterior distribution of $\boldsymbol{\alpha}$ for the assigned hyper prior. It is therefore of interest to define a proper prior for the variances. In terms of the scaled variance ratios $\gamma_m = (\boldsymbol{\phi}_m^T \boldsymbol{\phi}_m)\alpha_m/\sigma^2$ a useful proper prior is

$$p(\gamma_m|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\gamma_m^{a-1}(1+\gamma_m)^{-(a+b)}, \text{ for } a>0 \text{ and } b>0. \quad (15)$$

For $a = b = \delta$ small, (15) gives very similar shrinkage properties as shown in ter Braak (2006) for (14). This prior is closely related to the beta distribution; if $s_m \sim$ Beta(a,b), then $\gamma_m = s_m/(1-s_m)$ follows distribution (15). Conversely, $s_m = \gamma_m/(1+\gamma_m)$ which can be interpreted as shrinkage coefficient; it relates the shrunken estimate $\widetilde{w}_m$ to the least-squares estimate $\widehat{w}_m$ via $\widetilde{w}_m = s_m \widehat{w}_m$ in the orthogonal predictors case. Whereas (15) implies a proper Beta(a,b) prior for s_m, (14) implies the improper Beta($\delta, -\delta$) prior. The model with the proper prior is a rival for methods in which discrete mixtures of weights (George and McCulloch, 1993; Johnstone and Silverman, 2004) or variances (Meuwissen et al., 2001) give sparsity and is of interest for further study; see *e.g.* Polson and Scott (2009). Such models are needed as this paper suggests that Type II ML in the linear model with individual variance parameters is not the general answer in high dimensional prediction problems.

Appendix A: Derivation of equation (7)

Here we convert the marginal likelihood $L(\boldsymbol{\alpha})$ from a form that uses $N \times N$ matrices to one that uses $M \times M$ matrices. We start with

$$L(\boldsymbol{\alpha}) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{t}^T\mathbf{C}^{-1}\mathbf{t}\right), \quad (A.1)$$

where $\mathbf{C} = \sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T$. The Matrix Determinant Lemma gives (Golub and van Loan, 1989; Roweis, 1999)

$$|\mathbf{C}| = |\sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T| = |\sigma^2\mathrm{I}||\mathbf{A}|\left|\mathbf{A}^{-1} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\sigma^{-2}\right| = |\sigma^2\mathrm{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{A}|. \quad (A.2)$$

The Matrix Identity Lemma or Woodbury formula gives (Bishop, 2006; Golub and van Loan, 1989; Roweis, 1999)

$$\mathbf{C}^{-1} = (\sigma^2\mathrm{I} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^T)^{-1} = \sigma^{-2}[I - \boldsymbol{\Phi}(\sigma^2\mathbf{A}^{-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}], \quad (A.3)$$

so that

$$\mathbf{t}^T\mathbf{C}^{-1}\mathbf{t} = \sigma^{-2}(\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}(\sigma^2\mathbf{A}^{-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}). \quad (A.4)$$

On inserting (A.2) and (A.4) in (A.1) and deleting the terms that do not depend on $\boldsymbol{\alpha}$, we obtain

$$L(\boldsymbol{\alpha}) \propto |\mathbf{I} + \sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{A}|^{-1/2}\exp\left(\frac{1}{2\sigma^2}\mathbf{t}^T\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sigma^2\mathbf{A}^{-1})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right), \quad (A.5)$$

which is (7).

Appendix B: Derivation of equation (8)

If $\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} = \mathbf{I}$, (A.5) decomposes as a product of individual likelihoods $L(\alpha_m)$ with

$$L(\alpha_m) \propto (1 + \sigma^{-2}\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m\alpha_m)^{-1/2}\exp\left(\frac{1}{2\sigma^2}\mathbf{t}^{\boldsymbol{T}}\boldsymbol{\phi}_m(\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m + \sigma^2/\alpha_m)^{-\mathbf{1}}\boldsymbol{\phi}_m^T\mathbf{t}\right). \quad \text{(B.1)}$$

With $\widehat{w}_m = \boldsymbol{\phi}_m^T\mathbf{t}/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the least-squares estimate, and $v_m = \sigma^2/\boldsymbol{\phi}_m^T\boldsymbol{\phi}_m$, the variance of $\widehat{w}_m$, (B.1) can be written as

$$L(\alpha_m) \propto (1 + v_m^{-1}\alpha_m)^{-1/2}\exp\left(\frac{\widehat{w}_m^2 v_m^{-2}\alpha_m}{2(1+v_m^{-1}\alpha_m)}\right), \quad \text{(B.2)}$$

which is (8).

Appendix C: Derivation of equation (10)

Next, we consider the case of two correlated predictor variables with weights with variance parameters α_1 and α_2. On defining $\boldsymbol{\gamma} = \boldsymbol{\alpha}/\sigma^2$, $c = \boldsymbol{\phi}_1^T\mathbf{t}/\sigma$, $d = \boldsymbol{\phi}_2^T\mathbf{t}/\sigma$, (A.5) becomes

$$L(\gamma_1,\gamma_2) \propto \left|I + \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\begin{bmatrix}\gamma_1 & 0\\ 0 & \gamma_2\end{bmatrix}\right|^{-1/2}\exp\left[\frac{1}{2}\begin{bmatrix}c\\ d\end{bmatrix}\left[\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix} + \begin{bmatrix}\gamma_1 & 0\\ 0 & \gamma_2\end{bmatrix}^{-1}\right]^{-1}\begin{bmatrix}c & d\end{bmatrix}\right]. \quad \text{(C.1)}$$

Differentiating $L(\gamma_1,\gamma_2)$ with respect to γ_1 with Mathematica and then setting derivatives to zero gave a ratio for $\hat{\gamma}_1$ with numerator

$$-1 + c^2 - 2\hat{\gamma}_2 + 2c^2\hat{\gamma}_2 - \hat{\gamma}_2^2 + c^2\hat{\gamma}_2^2 - 2cd\rho\hat{\gamma}_2 - 2cd\rho\hat{\gamma}_2^2 + \rho^2\hat{\gamma}_2 + \rho^2\hat{\gamma}_2^2 + d^2\rho^2\hat{\gamma}_2^2 \quad \text{(C.2)}$$

and denominator $(1 + \hat{\gamma}_2 - \hat{\gamma}_2\rho^2)^2$. Simplifying with Mathematica did not help and was done by hand by collecting terms that involved *c* or *d* and those that did not. The terms involving *c* or *d* are

$$c^2\hat{\gamma}_2^2 - 2cd\rho\hat{\gamma}_2^2 + d^2\rho^2\hat{\gamma}_2^2 = \hat{\gamma}_2^2(c - d\rho)^2, \quad \text{(C.3)}$$

$$2c^2\hat{\gamma}_2 - 2cd\rho\hat{\gamma}_2 = 2c\hat{\gamma}_2(c - d\rho) \quad \text{(C.4)}$$

and c^2, resulting in

$$\hat{\gamma}_2^2(c - d\rho)^2 + 2c\hat{\gamma}_2(c - d\rho) + c^2 = (\hat{\gamma}_2(c - d\rho) + c)^2 \quad \text{(C.5)}$$

and the terms involving neither *c* nor *d* are

$$-1 - 2\hat{\gamma}_2 - \hat{\gamma}_2^2 + \rho^2\hat{\gamma}_2 + \rho^2\hat{\gamma}_2^2 = -(1 + \hat{\gamma}_2)\left(1 + \hat{\gamma}_2(1 - \rho^2)\right), \quad \text{(C.6)}$$

so that by insertion

$$\hat{\gamma}_1 = \frac{\left(c+\hat{\gamma}_2(c-\rho d)\right)^2 - (1+\hat{\gamma}_2)\left(1+\hat{\gamma}_2(1-\rho^2)\right)}{(1+\hat{\gamma}_2(1-\rho^2))^2}. \quad \text{(C.7)}$$

The expression for $\hat{\gamma}_2$ was obtained by symmetry.

Acknowledgements

We thank Laura Astola for help with Mathematica in the extended RVM model, Luke Tierney for suggesting the Beta prior for the shrinkage coefficient and the reviewers for constructive comments. Jamil's research was supported by a grant from Higher Education Commission of Pakistan through NUFFIC (The Netherlands).

References

Abramovich, F., Benjamini, Y., Donoho, D.L., Johnstone, I.M., 2005. Adapting to unknown sparsity by controlling the false discovery rate. Ann**.** Statist. 34, 584-653.

Bates, D., Maechler, M., Bolker, B., 2011. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39, http://CRAN.R-project.org/package=lme4.

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Bishop, C., 2006. Pattern Recognition and Machine Learning Springer, New York.

Donoho, D., Johnstone, J., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425-455.

Donoho, D.L., 1995. De-noising via soft-thresholding. IEEE T. Inform**.** Theory 41, 613-627

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least Angle Regression. Ann**.** Statist. 32, 407-451.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348-1360.

Faul, A.C., Tipping, M.E., 2002. Analysis of sparse Bayesian learning, in: Dietterich, T.G., Becker, S., Ghahramani, Z. (Eds.), Adv**.** Neur **.** Inform. Process. Syst. 14, Vols 1 and 2, pp. 383-389.

Figueiredo, M.A.T., 2003. Adaptive Sparseness for Supervised Learning. IEEE Trans. Pattern Anal**.** Machine Intell. 25, 1150-1159.

Frank, I.E., Friedman, J.H., 1993. A Statistical View of Some Chemometrics Regression Tools. Technometrics 35, 109-135.

Friedman, J.H., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1-21.

George, E.I., McCulloch, R.E., 1993. Variable Selection Via Gibbs Sampling. J. Am. Stat. Assoc. 88, 881-889.

Golub, G.H., van Loan, C.F., 1989. Matrix computations, The John Hopkins University Press, Baltimore.

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 55-67.

Li, Y., Campbell, C., Tipping, M., 2002. Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 18, 1332-1339.

Johnstone, I.M., Silverman, B.W., 2004. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. Ann**.** Statist. 32, 1594-1649.

Karatzoglou, K., Smola, A., Hornik, K., Zeileis, A, 2004. kernlab - An S4 Package for Kernel Methods in R. J. Statist. Soft. 11, 1-9.

MacKay, D.J.C., 1992. The Evidence Framework Applied to Classification Networks. Neural Comput. 4, 720-736.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819-1829.

Neal, R., 1996. Bayesian Learning for Neural Networks. Springer, Berlin.

O'Hagan, A., Forster, J., 2004. Kendall's advanced theory of statistics: Bayesian inference. Arnold, London.

Polson, N.G., Scott, J.G., 2009. Alternative global-local shrinkage priors using hypergeometric-beta mixtures, Technical report. University of Chicago, doi: 10.1.1.161.3592.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing. www.R-project.org, Vienna.

Rogers, S., Girolami, M., 2005. A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics 21, 3131-3137.

Roweis, S., 1999. Matrix identities. http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf.

Searle, S.R., Casella, G., McCulloch, C.E., 2008. Variance components. John Wiley, New York.

ter Braak, C.J.F., 2006. Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising. Comput. Stat. Data Anal. 51, 1232-1242.

ter Braak, C.J.F., 2009. Regression by L1 regularization of smart contrasts and sums (ROSCAS) beats PLS and elastic net in latent variable model. J. Chemometr. 23, 217-228.

ter Braak, C.J.F., Boer, M.P., Bink, M.C.A.M., 2005. Extending Xu's Bayesian Model for Estimating Polygenic Effects Using Markers of the Entire Genome. Genetics 170, 1435-1438.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. Royal Stat. Soc. B 58, 267-288.

Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. J. Machine Learn. Res. 1, 211-244.

Tipping, M.E., Faul, A., 2003. Fast marginal likelihood maximisation for sparse Bayesian models, in: Bishop, C.M., Frey, B.J. (Eds.), Proc. 9th Internat. Workshop on Artificial Intelligence and Statist. Key West, FL.

Wehrens, R., Mevik, B.-H., 2006. The PLS package 2.0-0. Multivariate regression by partial least squares regression (PLSR) and principal component regression (PCR). http://cran.r-project.org/doc/packages/.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics Chemometr**.** Intell**.** Lab**.** Sys. 58, 109-130.

Xu, S., 2007. An Empirical Bayes Method for Estimating Epistatic Effects of Quantitative Trait Loci. Biometrics 63, 513-521.

Xu, S., 2010. An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity 105:483-494.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Royal Stat. Soc. B 67, 301-320.

Table 2. Median mean-squared prediction errors (MSEP) for the simulated examples 1-7 for six methods based on 100 replications. In parentheses are the corresponding standard errors (of the medians) estimated by using 1000 bootstrap resamplings of the 100 MSEPs. For each example the smallest mean-square is in bold (NA = not available as rvm ended with an error).

Method	Example 1		Example 2		Example 3		Example 4		Example 5		Example 6		Example 7	
	MSEP	se	MSEP	se	MSEP	se	MSEP	se	MSEP	se	MSEP	se	MSEP	se
LASSO	**12.4**	(0.34)	13.5	(0.31)	311.0	(4.3)	483.4	(8.8)	436.6	(8.2)	**663.8**	(14.3)	1134.9	(19.3)
PLS	13.4	(0.38)	**11.0**	(0.30)	**273.4**	(4.4)	**351.5**	(4.9)	**361.7**	(6.7)	750.1	(13.8)	**989.7**	(19.9)
Type-II ML	12.9	(0.40)	13.9	(0.35)	380.1	(7.5)	1132.7	(27.0)	601.4	(11.2)	1129.5	(27.1)	1278.4	(37.1)
REML	12.8	(0.37)	13.8	(0.29)	379.7	(8.3)	1155.9	(37.6)	599.5	(10.1)	1110.7	(32.2)	1314.1	(36.1)
RVM_{rbf}	24.0	(0.81)	20.6	(1.12)	347.0	(3.2)	605.7	(7.3)	437.3	(6.17)	1061.0	(11.5)	1044.2	(10.7)
RVM_{lin}	NA		NA		431.5	(7.4)	512.3	(6.4)	603.1	(8.5)	884.2	(12.9)	1279.6	(9.5)
σ^2	9		9		225		225		225		225		225	
N	20		20		50		50		50		50		50	