Bioinformatics, 31(11), 2015, 1754–1761
doi: 10.1093/bioinformatics/btv037
Advance Access Publication Date: 24 January 2015
Original Paper

OXFORD

Gene expression

Bayesian feature selection for high-dimensional linear regression via the Ising approximation with applications to genomics

Charles K. Fisher* and Pankaj Mehta*

Department of Physics, Boston University, Boston, MA 02215, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on August 11, 2014; revised on November 13, 2014; accepted on January 16, 2015

Abstract

Motivation: Feature selection, identifying a subset of variables that are relevant for predicting a response, is an important and challenging component of many methods in statistics and machine learning. Feature selection is especially difficult and computationally intensive when the number of variables approaches or exceeds the number of samples, as is often the case for many genomic datasets.

Results: Here, we introduce a new approach—the Bayesian Ising Approximation (BIA)—to rapidly calculate posterior probabilities for feature relevance in L2 penalized linear regression. In the regime where the regression problem is strongly regularized by the prior, we show that computing the marginal posterior probabilities for features is equivalent to computing the magnetizations of an Ising model with weak couplings. Using a mean field approximation, we show it is possible to rapidly compute the feature selection path described by the posterior probabilities as a function of the L2 penalty. We present simulations and analytical results illustrating the accuracy of the BIA on some simple regression problems. Finally, we demonstrate the applicability of the BIA to high-dimensional regression by analyzing a gene expression dataset with nearly 30 000 features. These results also highlight the impact of correlations between features on Bayesian feature selection.

Availability and implementation: An implementation of the BIA in C++, along with data for reproducing our gene expression analyses, are freely available at http://physics.bu.edu/~pankajm/BIACode.

Contact: charleskennethfisher@gmail.com or ckfisher@bu.edu or pankajm@bu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Linear regression is one of the most broadly and frequently used statistical tools. Despite hundreds of years of research on the subject (Legendre, 1805), modern applications of linear regression to large datasets present a number of new challenges. Modern applications of linear regression, such as Genome Wide Association Studies (GWAS), often consider datasets that have at least as many potential variables (or features) as there are data points (McCarthy *et al.*, 2008). Applying linear regression to high-dimensional datasets often involves selecting a subset of relevant features, a problem known as feature selection in the literature on statistics and machine learning (Guyon and Elisseeff, 2003). Even for classical least-squares linear regression, it turns out that the associated feature selection problem is quite difficult (Huo and Ni, 2007).

The difficulties associated with feature selection are especially pronounced in genomics and GWAS. In general, the goal of many genomics studies is to identify a relationship between a small number of genes and a phenotype of interest, such as height or body mass index (Burton *et al.,* 2007; McCarthy *et al.,* 2008; Peng *et al.,* 2010; Subramanian *et al.,* 2005; Wu *et al.,* 2009). For example, many GWAS seek to identify specific genetic mutations (called single nucleotide polymorphisms—SNPs) that best explain the variation of a quantitative trait, such as height or body mass index, in a population (Yang *et al.,* 2012). Using various techniques, the trait is regressed against binary variables representing the presence or absence of the SNPs in order to find a subset of SNPs that are highly explanatory for the trait (Peng *et al.,* 2010; Wu *et al.,* 2009). Although the number of individuals genotyped in such a study may be in the thousands or even tens of thousands, this pales in comparison to the number of potential SNPs which can be in the millions (McCarthy *et al.,* 2008). Moreover, the presence or absence of various SNPs tends to be correlated due to chromosome structure and genetic processes that induce the so-called linkage disequilibrium (Yang *et al.,* 2012). As a result, selecting the best subset of SNPs for the regression involves a search for the global minimum of a landscape that is both high dimensional (due to the large number of SNPs) and rugged (due to correlations between SNPs).

The obstacles that make feature selection difficult in GWAS also occur in many other applications of linear regression to big datasets. In fact, the task of finding the optimal subset of features is proven, in general, to be NP-hard (Huo and Ni, 2007). Therefore, it is usually computationally prohibitive to search over all possible subsets of features and one has to resort to other methods of feature selection. For example, forward (or backward) selection adds (or eliminates) one feature at a time to the regression in a greedy manner (Guyon and Elisseeff, 2003). Alternatively, one may use heuristic methods such as Sure Independence Screening (SIS) (Fan and Lv, 2008), which selects features independently based on their correlation with the response, or Minimum Redundancy Maximum Relevance (Ding and Peng, 2005), which penalizes features that are correlated with each other. The most popular approaches to feature selection for linear regression, however, are penalized least-squares methods (Candes and Tao, 2007; Hoerl and Kennard, 1970; Tibshirani, 1996; Zou and Hastie, 2005) that introduce a function that penalizes large regression coefficients. Common choices for the penalty function include an L2 penalty, called ‘Ridge’ regression (Hoerl and Kennard, 1970), and an L1 penalty, commonly referred to as LASSO regression (Tibshirani, 1996).

Penalized methods for linear regression typically have natural interpretations as Bayesian approaches with appropriately chosen prior distributions. For example, L2 penalized regression can be derived by maximizing the posterior distribution obtained with a Gaussian prior on the regression coefficients. Similarly, L1 penalized regression can be derived by maximizing the posterior distribution obtained with a Laplace (i.e. double-exponential) prior on the regression coefficients. While penalized regression methods essentially aim to find the features that maximize a posterior distribution they do not allow one to actually compute posterior probabilities, which provide information about confidence in a Bayesian framework. Calculating these posterior probabilities generally requires Monte Carlo methods, which can be very computationally demanding in high dimensions (George and McCulloch, 1993; Guan *et al.,* 2011; Li and Zhang, 2010). Thus, in order to apply Bayesian approaches to feature selection to high-dimensional problems it is necessary to develop approximate methods for computing posterior probabilities that bypass the need for extensive sampling from the posterior distribution.

Inspired by the success of statistical physics approaches to hard problems in computer science (Mézard *et al.,* 2002; Monasson *et al.,* 1999) and statistics (Balasubramanian, 1997; Malzahn and Opper, 2005; Nemenman and Bialek, 2002), we study high-dimensional regression with ‘strongly regularizing’ prior distributions. A strongly regularizing prior distribution is one that exerts a significant influence on the posterior distribution even when the sample size goes to infinity. The definition will be made more precise later. In this strongly regularized regime, we show that the marginal posterior probabilities of feature relevance for L2 penalized regression are well-approximated by the magnetizations of an appropriately chosen Ising model—a widely studied model from physics used to describe magnetic materials (Opper and Winther, 2001). For this reason, we call our approach the Bayesian Ising Approximation (BIA) of the posterior distribution. Using the BIA, the posterior probabilities can be computed without resorting to Monte Carlo simulation using an efficient mean field approximation that facilitates the analysis of very high-dimensional datasets. We envision the BIA as part of a two-stage procedure where the BIA is applied to rapidly screen irrelevant variables, i.e. those that have low rank in posterior probability, before applying a more computationally intensive cross-validation procedure to infer the regression coefficients for the reduced feature set. This study is especially well suited to modern feature selection problems where the number of features, p, is often larger than the sample size, n.

Our approach differs significantly from previous methods for feature selection. Traditionally, penalized regression and related Bayesian approaches have focused on the ‘weakly regularized regime’ where the effect of the prior is assumed to be negligible as the sample size tends to infinity. The underlying intuition for considering the weak-regularization regime is that as long as the prior (i.e. the penalty parameter) is strong enough to regularize the inference problem, a less influential prior distribution should be better suited for feature selection and prediction tasks because it ‘allows the data to speak for themselves’ (Gelman *et al.,* 2013). In the machine learning literature, the penalty parameter is usually chosen using cross validation to maximize out-of-sample predictive ability (Tibshirani, 1996; Zou and Hastie, 2005). A similar esthetic is also reflected in the abundant literature on ‘objective’ priors for Bayesian inference (Ghosh *et al.,* 2011). As expected, these weakly regularizing approaches perform well when the sample size exceeds the number of features ($n \gg p$). However, very strong priors may be required for high-dimensional inference where the number of features can greatly exceed the sample size ($p \gg n$). Our BIA approach exploits the large penalty parameter in this strongly regularized regime to efficiently calculate marginal posterior probabilities using methods from statistical physics.

The article is organized as follows: in Section 2.1, we review Bayesian linear regression; in Section 2.2, we derive the BIA using a series expansion of the posterior distribution and describe the associated algorithm for variable selection; and in Section 3.1, we present analytical results and simulations on the performance of the BIA using features with a constant correlation, in Section 3.2 we analyze a real dataset for predicting bodyfat percentage from 12 different body measurements and in Section 3.3 we analyze a real dataset for predicting a quantitative phenotypic trait from data on the expression of 28 395 genes in soybeans.

2 Methods

2.1 Bayesian linear regression

In this section, we briefly review the necessary aspects of Bayesian linear regression. This entire section follows standard arguments, the details of which can be found in many textbooks on Bayesian statistics (see e.g. O’Hagan *et al.,* 2004). The goal of linear regression is to infer the set of coefficients β_j for $j = 1, \ldots, p$ that describe the relationship $y = \mathbf{x}^T\beta + \eta$ from n observations $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$. Here, $\mathbf{x}$ is a $(p \times 1)$ vector of features and $\eta \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian distributed random variable with unknown variance σ^2. Without loss of generality, we will assume throughout this article that the data are standardized with $\sum_i y_i = 0$, $\sum_i y_i^2 = n$, $\sum_i (\mathbf{x}_i)_j = 0$ and $\sum_i (\mathbf{x}_i)_j^2 = n$ so that it is not necessary to include an intercept term in the regression. Penalized least-squares methods estimate the regression coefficients by minimizing a convex objective function in the form of:

$$U(\beta) = \sum_i (y_i - \mathbf{x}_i^T\beta)^2 + \lambda f(\beta), \tag{1}$$

where $f(\beta)$ is a function that penalizes large regression coefficients and λ is the strength of the penalty. Common choices for the penalty function include $f(\beta) = \sum_j \beta_j^2$ for L2 penalized or 'Ridge' regression (Hoerl and Kennard, 1970), and $f(\beta) = \sum_j |\beta_j|$ for L1 penalized or LASSO regression (Tibshirani, 1996). The standard least-squares (and maximum likelihood) estimate $\hat{\beta} = (X^TX)^{-1}X^T\mathbf{y}$ is recovered by setting $\lambda = 0$, where X is the $(n \times p)$ design matrix with columns x_i. Adding a penalty to the least-squares objective function mitigates instability that results from computing the inverse of the X^TX matrix. In the case of the L1 penalty, many of the regression coefficients end up being shrunk exactly to 0 resulting in a type of automatic feature selection (Candes and Tao, 2007; Tibshirani, 1996; Zou and Hastie, 2005).

Bayesian methods combine the information from the data, described by the likelihood function, with a priori knowledge, described by a prior distribution, to construct a posterior distribution that describes one's knowledge about the parameters after observing the data. In the case of linear regression, the likelihood function is a Gaussian

$$P(\mathbf{y}|\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)}{2\sigma^2}\right).$$

In this work, we will use standard conjugate prior distributions for β and σ^2 given by$P(\beta, \sigma^2|\mathbf{s}) = P(\sigma^2)P(\beta|\sigma^2, \mathbf{s})$where

$$P(\sigma^2) \propto (\sigma^2)^{-(a_0+1)} \exp\left(-b_0/\sigma^2\right)$$
$$P(\beta|\sigma^2, \mathbf{s}) \propto \prod_j \left[(1 - s_j)\delta(\beta_j) + (1 + s_j)\sqrt{\frac{\lambda}{2\pi\sigma^2}} \exp\left(-\frac{\lambda\beta_j^2}{2\sigma^2}\right)\right].$$

These distributions were chosen because they ensure that the posterior distribution can be obtained in closed-form (O'Hagan *et al.*, 2004). Here, we have introduced a vector $(\mathbf{s})$ of indicator variables so that $\beta_j = 0$ if $s_j = -1$ and $\beta_j \neq 0$ if $s_j = +1$. We also have to specify a prior for the indicator variables, which we will set to a flat prior $P(\mathbf{s}) \propto 1$ for simplicity. In principle, a_0, b_0 and the penalty parameter on the regression coefficients, λ, are free parameters that must be specified ahead of time to reflect our prior knowledge. We will discuss these parameters in the following section.

We have set up the problem so that identifying which features are relevant is equivalent to identifying those features for which $s_j = +1$. Therefore, we need to compute the posterior distribution for $\mathbf{s}$, which can be determined from Bayes' theorem:

$$\begin{aligned} \log P_\lambda(\mathbf{s}|\mathbf{y}) + C &= \log \int d\beta d\sigma^2 P(\mathbf{y}|\beta, \sigma^2)P(\beta, \sigma^2|\mathbf{s})P(\mathbf{s}) \\ &= \frac{1}{2}\ln|\lambda I| - \frac{1}{2}\ln|\lambda I + X_\mathbf{s}^TX_\mathbf{s}| - \left(a_0 + \frac{n}{2}\right)\ln\left(b_0 + \frac{1}{2}E_\mathbf{s}(\lambda)\right), \end{aligned} \tag{2}$$

where C is a constant and $E_\mathbf{s}(\lambda)$ is the sum of the squared residual errors. In this expression, $q = \sum_j (1 + s_j)/2$ is the number of variables with $s_j = +1$, I is the $(q \times q)$ identity matrix and $X_\mathbf{s}$ is a $(n \times q)$ restricted design matrix which only contains columns corresponding to features where $s_j = +1$. The sum of the squared residual errors is given by $E_\mathbf{s}(\lambda) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^TX_\mathbf{s}\overline{\beta}_\mathbf{s}(\lambda)$, where $\overline{\beta}_\mathbf{s}(\lambda) = (\lambda I + X_\mathbf{s}^TX_\mathbf{s})^{-1}X_\mathbf{s}^T\mathbf{y}$ is the Bayesian estimate for the regression coefficients corresponding to those variables for which $s_j = +1$.

2.2 The Ising approximation

2.2.1 Strongly regularized expansion

In principle, one can directly use Equation (2) to estimate the relevance of each feature using two different approaches. First, we could find the $\mathbf{s}$ that maximizes the posterior probability distribution. Alternatively, we could compute the marginal probabilities of feature relevance, $P_\lambda(s_j = +1|\mathbf{y}) = (1 + \langle s_j\rangle)/2$, where $\langle s_j\rangle$ is the expectation value of s_j with respect to the posterior distribution, and select the features with the largest $P_\lambda(s_j = +1|\mathbf{y})$. In the Bayesian setting, these two point estimates result from the use of different utility functions (Berger, 1985). Here, we will focus on computing the latter, i.e., the expected value of $\mathbf{s}$. The expectation values cannot be evaluated analytically due to the cumbersome restriction of the design matrix to those variables for which $s_j = +1$. Moreover, although the computation of the expectation values can be performed using Monte Carlo methods (George and McCulloch, 1993; Li and Zhang, 2010), the numerical calculations often take a long time to converge for high-dimensional inference problems.

Our main result—which we call the BIA of the posterior distribution for feature selection—is that a second-order series expansion of Equation (2) in λ^{-1} corresponds to an Ising model described by

$$\log P_\lambda(\mathbf{s}|\mathbf{y}) \simeq \frac{n^2}{4\lambda}\left(\sum_i b_i(\lambda)s_i + \frac{1}{2}\sum_{i,j;i\neq j} J_{ij}(\lambda)s_is_j\right) \tag{3}$$

with an error that is $O\left(\lambda^{-3}\mathrm{Tr}[(X_\mathbf{s}^TX_\mathbf{s})^3]\right)$ where $\mathrm{Tr}[\cdot]$ is the matrix trace operator and the external fields and couplings are defined as

$$b_i(\lambda) = r^2(y, x_i) - \frac{1}{n} + \sum_j J_{ij}(\lambda) \tag{4}$$

$$J_{ij}(\lambda) = \lambda^{-1}r^2(x_i, x_j) - \frac{n}{\lambda}\left(r(x_i, x_j)r(y, x_i)r(y, x_j) - \frac{1}{2}r^2(y, x_i)r^2(y, x_j)\right). \tag{5}$$

Here, $r(z_1, z_2)$ is the Pearson correlation coefficient between variables z_1 and z_2. In writing this expression, we have assumed that the hyperparameters a_0 and b_0 are small enough to neglect, though this assumption is not necessary. A detailed derivation of this result is presented in the Supporting Information.

The series expansion converges as long as $\lambda^k > \mathrm{Tr}[(X_\mathbf{s}^TX_\mathbf{s})^k]$ for all $\mathbf{s}$ and integer powers $k \geq 1$, which defines the regime that we call 'strongly regularized'. Since $X_\mathbf{s}$ is the restricted design matrix for standardized data, we can relate $\mathrm{Tr}[(X_\mathbf{s}^TX_\mathbf{s})^k]$ to the covariances between x_j's. In particular, Gershgorin's Circle Theorem (Varga, 2010) implies that the series will converge as long as $\lambda > n(1 + p\tilde{r})$ where $\tilde{r} = \frac{1}{p}\inf_i \sum_{j\neq i}|r(X_i, X_j)|$ (see Supporting Information). For large p, we can replace $\tilde{r}$ by the root-mean-squared correlation between features, $r = \sqrt{p^{-1}(p-1)^{-1}\sum_{i\neq j} r^2(X_i, X_j)}$. This defines a natural scale

$$\lambda^* = n(1 + pr). \tag{6}$$

for the penalty parameter at which the BIA is expected to breakdown. We expect the BIA to be accurate when $\lambda \gg \lambda^*$ and to breakdown when $\lambda \ll \lambda^*$.

Because higher-order terms in the series can be neglected, the strongly regularized expansion allows us to remove any references to the restricted design matrix, and maps the posterior distribution to the Ising model, which has been studied extensively in the physics literature. Moreover, the magnitude of the couplings (J_{ij}) scales as λ^{-1}, ensuring that the couplings are weak, which will allow us to compute posterior probabilities analytically. To perform feature selection, we are interested in computing marginal probabilities $P_\lambda(s_j = 1|\mathbf{y}) \simeq (1 + m_j(\lambda))/2$, where we have defined the magnetizations $m_j(\lambda) = \langle s_j \rangle$. While there are many techniques for calculating the magnetizations of an Ising model, we focus on the mean field approximation which leads to a self-consistent equation (Opper and Winther, 2001):

$$m_i(\lambda) = \tanh\left[\frac{n^2}{4\lambda}\left(h_i(\lambda) + \frac{1}{2}\sum_{j\neq i} J_{ij}(\lambda) m_j(\lambda)\right)\right]. \tag{7}$$

This mean field approximation provides a computationally efficient tool that approximates Bayesian feature selection for linear regression, requiring only the calculation of the Pearson correlations and solution of Equation (7).

2.2.2 Computing the feature selection path

As with other approaches to penalized regression, our expressions depend on a free parameter (λ) that determines the strength of the prior distribution. As it is usually difficult, in practice, to choose a specific value of λ ahead of time it is often helpful to compute the feature selection path; i.e. to compute $m_j(\lambda)$ over a wide range of λ's. Indeed, computing the variable selection path is a common practice when applying other feature selection techniques such as LASSO regression. To obtain the mean field variable selection path as a function of $\epsilon = 1/\lambda$, we notice that $\lim_{\epsilon\to 0} m_j(\epsilon) = 0$ and so define the recursive formula

$$m_i(\epsilon + \delta\epsilon) \approx \tanh\left[\frac{(\epsilon + \delta\epsilon)n^2}{4}\left(h_i(\epsilon + \delta\epsilon) + \frac{1}{2}\sum_{j\neq i} J_{ij}(\epsilon + \delta\epsilon) m_j(\epsilon)\right)\right]$$

with a small step size $\delta\epsilon \ll 1/\lambda^* = n^{-1}(1 + pr)^{-1}$. We have set $\delta\epsilon = 0.05/\lambda^*$ in all of the examples presented below. We note that our implementation of the BIA is an example of homotopy algorithm and could potentially be improved by applying more advanced methods (Allgower and Georg, 2003).

2.2.3 Remarks

The BIA provides a computationally efficient framework to calculate posterior probabilities of feature relevance as a function of λ without Monte Carlo simulations. We have used a simple, unoptimized C++ implementation of the BIA method. Using this code, computing the entire feature selection path for a genomics dataset with almost 30 000 features took ~15 min on a desktop computer with 24 GB of RAM and two 2.4 GHz 6-core Intel Xeon processors. The bulk of the computational effort—in terms of both processing power and memory usage—is expended computing the ($p \times p$) correlation matrix. For example, computing the feature selection path for the genomics dataset with our naive implementation required ~15 GB of RAM. However, any method designed for efficiently computing large correlation matrices could be applied to improve the computational performance of the BIA. For example, adaptive thresholding estimators could be used to obtain a sparse correlation matrix that requires less memory (Cai and Liu, 2011). In any case, we have left the optimization of the code for future research.

To first order in $\epsilon = \lambda^{-1}$, the posterior distribution corresponds to an Ising model with fields and couplings given by $h_i = r^2(y, x_i)$ $-1/n$ and $J_{ij} = 0$. That is, the indicator variables representing feature relevance are independent, and the probability that a feature is relevant is only a function of its squared correlation with the response. Specifically, $m_j(\lambda) \geq 0$ if $|r(y, x_j)| > 1/\sqrt{n}$ and $m_j(\lambda) \leq 0$ if $|r(y, x_j)| < 1/\sqrt{n}$. Therefore, the BIA demonstrates that methods that rank features by their squared correlation with the response, such as SIS (Fan and Lv, 2008), are actually performing a first-order approximation to Bayesian feature selection in the strongly regularized limit.

The couplings between the spin variables representing feature relevance enter into the BIA with the second-order term in $\epsilon = \lambda^{-1}$. A positive coupling between spins i and j favors models that include both features i and j, whereas a negative coupling favors models that include one feature or the other, but not both. In general, the coupling terms are negative for highly correlated variables which minimizes the redundancy of the feature set.

3 Examples

We have chosen three examples to illustrate different characteristics of the BIA for Bayesian feature selection. (A) First, we consider regression problems with p features that have a constant correlation r. We present some simple analytic expressions in the large p limit that illustrates how different aspects of the problem affect feature selection performance, and study some simulated data. (B) Next, we analyze a dataset on the prediction of bodyfat percentage from various body measurements. The number of features ($p = 12$) is small enough that we can compute the exact posterior probabilities and, therefore, directly assess the accuracy of the BIA for these data. (C) Finally, we demonstrate the applicability of the BIA for feature selection on high-dimensional regression problems by examining a dataset relating the expression of $p = 28\,395$ genes to the susceptibility of soybean plants to a pathogen.

3.1 Features with a constant correlation

Correlations between features are detrimental to feature selection. For example, suppose that we observe a response variable y given by $y = \beta x_1 + \eta$. A second feature x_2 that is strongly correlated with x_1 will also be correlated with y. Thus, identifying which feature, x_1 or x_2, is the relevant one is not an easy task. Of course, the reasoning becomes more complicated in high dimensions, but similar effects are observed in high-dimensional regression with the LASSO (Tibshirani, 1996; Zou and Hastie, 2005). Given these observations, we use this section to analyze a simple model of BIA feature selection that allows us to examine many of the characteristics that influence feature selection performance. Specifically, we consider a simple, analytically tractable, model in which we are given p features that are correlated with each other with a constant Pearson correlation coefficient, r. The response, $\tilde{y}$, is a linear function of the first $\tilde{p} \leq p$ variables, which have equal true regression coefficients $\beta_j = \beta$ for $j \leq \tilde{p}$. That is, $\tilde{y} = \beta \sum_{j=1}^{j=\tilde{p}} x_j + \tilde{\eta}$ where $\tilde{\eta} \sim \mathcal{N}(0, \tilde{\sigma}^2)$ is a Gaussian noise. We are interested in studying the behavior of this model when the number of features is large ($p \gg 1$). To simplify analytic expressions, it is helpful to define the number of samples as $n = \theta p$, and the number of relevant features as $\tilde{p} = \phi p$. Furthermore, we assume that the correlation between features scales as $r = \alpha p^{-1}$ so that the correlation between y and x_j stays constant in the large p limit.

Figure 1a presents an example feature selection path computed using the BIA for a simulation of this model. This variable selection path was generated for data simulated from a linear model using

$p = 200$ features with a constant correlation $r = 2/p$, $n = 100$, $\tilde{p} = 10$ and $\omega^2 = \tilde{\sigma}^2/\beta^2 = 1$. Figure 1a demonstrates that all but one of the relevant features (red) have higher posterior probabilities than the irrelevant features (black) as long as $\lambda > \lambda^*$. In fact, there is a clear gap in posterior probability separating the relevant and irrelevant features, and the correct features can be easily selected by visible inspection of the feature selection path in Fig. 1a. The BIA breaks down beyond the threshold of the penalty parameter and the feature selection performance of the BIA deteriorates, as demonstrated by the mixing of the probabilities for the relevant (red lines) and irrelevant (black lines) features in Fig. 1a.

The indicator variables characterizing the feature selection problem can be divided into two groups: relevant features with $j \leq \tilde{p}$ and magnetization $m_{(+)}$, and irrelevant features with $j > \tilde{p}$ and magnetization $m_{(-)}$. Note that an algorithm that performs perfect variable selection will have $m_{(+)} = +1$ and $m_{(-)} = -1$. The Pearson correlation coefficient of a relevant feature ($j \leq \tilde{p}$) with the standardized response $y = \tilde{y}/\sqrt{\mathrm{VAR}(\tilde{y})}$ is given by

$$r(y, x_{j=1\ldots\tilde{p}}) \equiv r_{(+)} = \frac{1 + r(\tilde{p} - 1)}{\sqrt{\omega^2 + \tilde{p}(r\tilde{p} + 1 - r)}},$$

where $\omega^2 = \tilde{\sigma}^2/\beta^2 \sim \mathrm{O}(1)$ is an inverse signal-to-noise ratio. Similarly, the Pearson correlation coefficient of an irrelevant variable ($j > \tilde{p}$) with the standardized response is

$$r(y, x_{j=\tilde{p}+1,\ldots,p}) \equiv r_{(-)} = \frac{r\tilde{p}}{\sqrt{\omega^2 + \tilde{p}(r\tilde{p} + 1 - r)}}.$$

Note that correlations make this problem incredibly difficult when the number of true features is large, i.e. $r_{(-)}/r_{(+)} \to 1$ as $\tilde{p} \to \infty$ for $r > 0$. If we choose $\lambda = \theta p^2$ to ensure that the problem is always in the strongly regularized regime, the magnetizations can be computed explicitly to order $1/p$ giving

$$m_{(+)} \approx \frac{\theta - \phi(1 - \alpha\theta)}{4\phi}\frac{1}{p} + \mathrm{O}\left(\frac{1}{p^2}\right),$$

$$m_{(-)} \approx -\frac{1 + \alpha\phi - \alpha^2\phi\theta}{4(1 + \alpha\phi)}\frac{1}{p} + \mathrm{O}\left(\frac{1}{p^2}\right).$$

In general, we say that feature selection performance is good, on average, as long as $m_{(-)} < 0 < m_{(+)}$, because relevant features have $P(s_j = +1|\mathbf{y}) > 1/2$ and irrelevant features have $P(s_j = +1|\mathbf{y}) < 1/2$. Figure 1b shows that the average feature selection performance is good in this sense within a large volume of the phase space. Specifically, $m_{(-)} < 0 < m_{(+)}$ when

$$\frac{1}{1 + \alpha\phi} < \frac{\theta}{\phi} < \frac{1 + \alpha\phi}{(\phi\alpha)^2}.$$

However, $m_{(-)} < m_{(+)}$ even if the stronger statement $m_{(-)} < 0 < m_{(+)}$ is not satisfied. As a result, there is always a gap between the posterior probabilities of the relevant and irrelevant features. Nevertheless, the gap between the relevant and irrelevant features shrinks with increasing correlations, suggesting that feature selection performance will be strongly affected by sample-to-sample fluctuations, which we have neglected here.

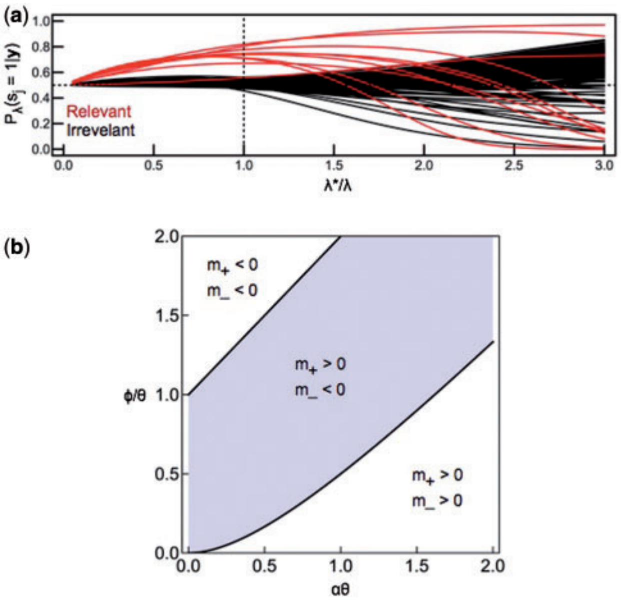

Fig. 1. Performance of BIA feature selection. (**a**) An example variable selection path as a function of decreasing regularization. The relevant variables are red, and the irrelevant variables are black. The dashed vertical line is at $\lambda = \lambda^* = n(1 + rp)$, which is the estimated breakdown point of the approximation. Simulations were performed with $p = 200$, $n = 100$, $\sim p^{\sim} = 10$, $r = 2/p$ and $\omega^2 = 1$. (**b**) A phase diagram illustrating the regions of parameter space where $m_{(-)} < 0 < m_{(+)}$ computed with $\lambda = \theta p^2$

3.2 Bodyfat percentage

Bodyfat percentage is an important indicator of health, but obtaining accurate estimates of bodyfat percentage is challenging. For example, underwater weighing is one of the most accurate methods for measuring bodyfat percentage but it requires special equipment, e.g. a pool. Here, we analyze a well-known dataset obtained from StatLib (http://lib.stat.cmu.edu/datasets/) on the relationship between bodyfat percentage and various body measurements from $n = 252$ men (Penrose *et al.*, 1985). The $p = 12$ features included in our regression are age and body mass index (height/mass^2), as well as circumference measurements of the neck, chest, waist, hip, thigh, knee, ankle, upper arm, forearm and wrist. All of the data were standardized to have mean 0 and variance 1. Therefore, there are $2^{12} = 4096$ potential combinations of features.

For our purposes, the most interesting part about the bodyfat dataset is that the number of features is small enough to compute the posterior probabilities exactly using Equation (2) by enumerating all of the 4096 feature combinations. The exact posterior probabilities as a function of λ^{-1} are shown in Fig. 2a.The posterior probabilities computed from recursive solution of the BIA are shown in Fig. 2b. Comparing Fig. 2a with Fig. 2b demonstrates that the posterior probabilities computed from the BIA are very accurate for $\lambda \gg \lambda^*$, with $\lambda^* = n(1 + pr)$ and r the root-mean-squared correlation between features. However, the approximation breaks down for $\lambda \ll \lambda^*$ as expected. Figure 2c provides another representation of the breakdown of the BIA upon approaching the breakdown point of the penalty (λ^*). The root-mean-squared error given by $\mathrm{RMSE}\ (\lambda) = \sqrt{p^{-1}\sum_j (P_\lambda^{\mathrm{exact}}(s_j = 1|\mathbf{y}) - P_\lambda^{\mathrm{BIA}}(s_j = 1|\mathbf{y}))^2}$ is sigmoidal, with an inflection point close to λ^*.

In the strongly regularized regime with $\lambda \gg \lambda^*$, the exact Bayesian probabilities and those computed using the BIA both rank waist and chest circumference as the most relevant features. Below the breakdown point of the penalty parameter, however, the BIA suggests solutions that are too sparse. That is, it underestimates many of the posterior probabilities describing whether or not the features are relevant. Far below the breakdown point of the penalty parameter (beyond the range of the graph in Fig. 2), the BIA ranks age and body mass index as the most relevant variables even though these have some of the smallest correlations with the response. Age and body mass index also become increasingly important for small λ's in the exact calculation; though, they are never ranked as the most relevant variables. The change in the rankings of the features as a function of λ highlights the importance of the coupling terms ($J_{ij}(\lambda)$) that punish correlated features.

3.3 Gene expression

In 2010, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) (Prill *et al.*, 2010) initiative issued a challenge to predict the response of soybean plants to a pathogen from data on gene expression (Zhou *et al.*, 2009). These DREAM5 training data consist of a response of $n = 200$ different soybean plants to a pathogen (specifically, the response is a measure of the amount of pathogen in an infected tissue sample) along with the expressions of $p = 28\,395$ genes. The team (Loh *et al.*, 2011) that achieved the highest rank correlation on a blind test set of 30 other soybean plants trained their model using elastic net regression to predict the ranks of the responses in the training set. The ranks were used rather than the actual values of the responses to mitigate the effects of outliers, and the value of the penalty parameter was chosen using cross validation. Loh *et al.* (2011) found that their cross-validation procedure for elastic net regression favored sparse models with only a few features, and they highlighted 12 of these features that were frequently chosen by their procedure. However, even the best teams achieved only modest performance on the test data (Loh *et al.*, 2011). Nevertheless, the soybean gene expression dataset presents a good benchmark to compare Bayesian feature selection with the BIA to feature selection using cross-validated penalized regression for a very high-dimensional inference problem.

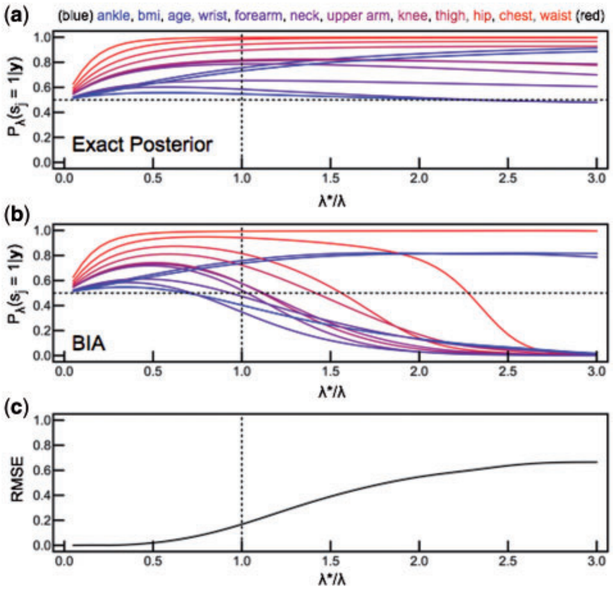

Fig. 2. Comparison of exact Bayesian marginal probabilities to the BIA for the bodyfat data. (**a**) Exact Bayesian marginal probabilities for decreasing regularization. (**b**) BIA approximations of the marginal probabilities for decreasing regularization. (**c**) RMSE between the exact and BIA probabilities as a function of decreasing regularization. The dashed vertical line is at $\lambda = \lambda^* = n(1 + rp)$, which is the estimated breakdown point of the approximation. The variables have been color coded (blue to red) by increasing squared Pearson correlation coefficient with the response (bodyfat percentage)

We used the BIA to compute the posterior probabilities for all $p = 28\,395$ features as a function of λ^{-1} using the ranks of the responses of the soybean plants to the pathogen as our y variable. As before, all of the data were standardized to have mean 0 and variance 1. Figure 3a compares the posterior probabilities of the 12 features highlighted by Loh *et al.* (2011) (red lines) to the distribution of posterior probabilities for all of the features (gray area). Visual inspection of Fig. 3a suggests that the 12 features identified by Loh *et al.* (2011) have some of the highest posterior probabilities among all 28 395 features. Similarly, Fig. 3b shows that only a small percentage of features have higher posterior probabilities than those identified by Loh *et al.* (2011), demonstrating that there is generally a pretty good agreement between features that are predictive (i.e. those that perform well in cross validation) and those with high posterior probabilities computed with the BIA.

Although our analyses of the soybean gene expression data identify similar features as cross-validated elastic net regression, the posterior probabilities all fall in the range $P_\lambda(s_j|\mathbf{y}) = 1/2 \pm 0.001$. The small range of posterior probabilities around the value representing random chance ($P_\lambda(s_j|\mathbf{y}) = 1/2$) is consistent with the highly variable out-of-sample performance discussed by Loh *et al.* (2011). One reason for the generally poor performance of feature selection on these data, aside from the underdetermined nature of the problem, is that the expressions of the genes are significantly correlated ($r \approx 0.29$). To demonstrate this, we constructed synthetic datasets with varying numbers of relevant and irrelevant genes and computed the rate at which true features were identified by the BIA (Fig. 4 and Supporting Information). Like the original data, these synthetic datasets each contained $n = 200$ distinct samples. The true positive rate (or sensitivity) was defined as the fraction of true features among the q features with the highest BIA posterior probabilities at $\lambda = 0.5\lambda^*$. Comparing the true positive rates of BIA feature selection on synthetic data using genes with a strong correlation ($r \approx 0.28$, Fig. 4a) and synthetic data with a weak correlation obtained by randomly shuffling the genes ($r \approx 0.07$, Fig. 4b) clearly demonstrates the dramatic effect that interfeature correlations have on feature selection performance. This highlights the importance of strong

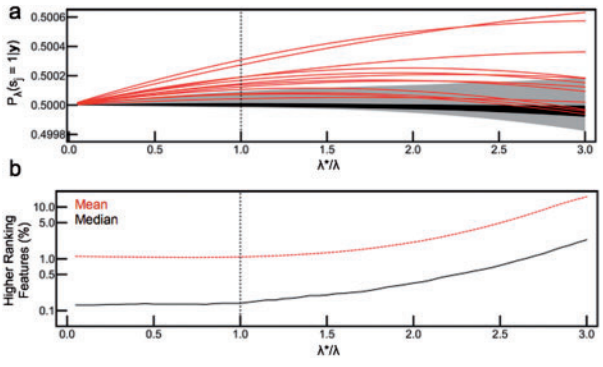

Fig. 3. Feature selection path for the gene expression data. The problem is severely under-determined, involving the prediction of a quantitative phenotype from the expressions of $p = 28\,395$ genes given a sample size of $n = 200$ and, therefore, the posterior probabilities remain close to $P_\lambda(s_j = 1|\mathrm{y}) = 1/2$. (**a**) Features selected in a previous study (red lines) by cross validation with the elastic net have high ranking posterior probabilities. Gray area represents the 1–99% quantiles, and the black area represents the 25–75% quantiles. (**b**) The median (solid black line) and mean (dashed red line) percentage of features with higher posterior probabilities than those identified by Loh *et al.* (2011). The vertical axis is a logarithmic scale. The dashed vertical line is at the breakdown point $\lambda = \lambda^* = n(1 + rp)$

regularization procedures that specifically account for correlation between genes in high-dimensional genomic studies.

4 Discussion

To summarize, we have shown that Bayesian feature selection for L2 penalized regression, in the strongly regularized regime, corresponds to an Ising model, which we call the BIA. Mapping the posterior distribution to an Ising model that has simple expressions for the local fields and couplings using a controlled approximation opens the door to analytical studies of Bayesian feature selection using the vast number of techniques developed in physics for studying the Ising model. It will be interesting to see if our analyses can be generalized to study Bayesian feature selection for many statistical techniques other than linear regression, as well as other prior distributions. From a practical standpoint, the BIA provides an algorithm to efficiently compute Bayesian feature selection paths for L2 penalized regression. Using our approach, it is possible to compute posterior probabilities of feature relevance for very high-dimensional datasets such as those typically found in genomic studies.

Unlike most previous work of feature selection, the BIA is ideally suited for large genomic datasets where the number of features can be much greater than the sample size, $p \gg n$. The underlying reason for this is that we work in strongly regularized regime where the prior always has a large influence on the posterior probabilities. This is in contrast to previous works on penalized regression and related Bayesian approaches that have focused on the ‘weakly regularized regime’ where the effect of the prior is assumed to be small. Moreover, we have identified a sharp threshold for the regularization parameter $\lambda^* = n(1 + pr)$ where the BIA is expected to break down. This threshold depends on the sample size, n, number of features, p, and root-mean-squared correlation between features, r. The threshold at which the BIA breaks down occurs precisely at the transition from the strongly regularized to the weakly regularized regimes where the prior and the likelihood have a comparable influence on the posterior distribution.

This study also highlights the importance of accounting for correlations between features when assessing statistical significance in large datasets. When the number of features is large, even small correlations can cause a huge reduction in the posterior probabilities of features. For example, our analysis of a dataset including the expression of 28 395 genes demonstrates that the resulting posterior probabilities of gene relevance may be very close to value representing random chance $P_\lambda(s_j|\mathbf{y}) = 1/2$ when $p \gg n$ and the genes are moderately correlated, e.g. $r \approx 0.29$. This is likely to have important implications for assessing the results of GWAS studies where such correlations are often ignored.

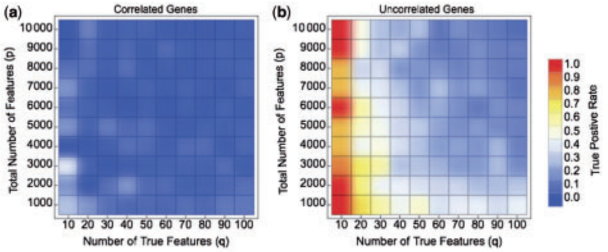

Fig. 4. True positive rates for feature selection with (**a**) correlated and (**b**) uncorrelated genes. Features were selected by taking the q genes with highest posterior probability at $\lambda = 0.5\lambda^*$. The uncorrelated genes were created by randomly shuffling the correlated genes. The root mean squared correlation among the correlated genes was $r = 0.28$ compared with $r = 0.7$ for the uncorrelated genes

Moreover, we suggest that it is generally not reasonable to choose a posterior probability threshold for judging significance on very high-dimensional problems. Instead, the BIA can be used as part of a two-stage procedure, where the BIA is applied to rapidly screen irrelevant variables, i.e. those that have low rank in posterior probability, before applying a more computationally intensive cross-validation procedure to infer the regression coefficients. The computational efficiency of the BIA and the existence of a natural threshold for the penalty parameter where the BIA works make this procedure ideally suited for such two-stage procedures.

Acknowledgements

We thank Alex Lang, Javad Noorbakhsh and David Schwab for comments on this manuscript.

Funding

P.M. and C.K.F. were supported by a Sloan Research Fellowship and a Simons Investigator Award [to P.M.], and internal funding from Boston University.

Conflict of Interest: none declared.

References

Allgower,E.L. and Georg,K. (2003) *Introduction to Numerical Continuation Methods,* **Vol. 45**. SIAM, Philadelphia.

Balasubramanian,V. (1997) Statistical inference, Occam’s razor, and statistical mechanics on the space of probability distributions. *Neural Comput.,* **9**, 349–368.

Berger,J.O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer, New York.

Burton,P.R. *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* **447**, 661–678.

Cai,T. and Liu,W. (2011) Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Stat. Assoc.,* **106**, 672–684.

Candes,E. and Tao,T. (2007) The dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.,* **35**, 2313–2351.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.,* **3**, 185–205.

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. B,* **70**, 849–911.

Gelman,A. *et al.* (2013) *Bayesian Data Analysis.* CRC Press, London.

George,E.I. and McCulloch,R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.,* **88**, 881–889.

Ghosh,M. *et al.* (2011) Objective priors: an introduction for frequentists. *Stat. Sci.,* **26**, 187–202.

Guan,Y. *et al.* (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.,* **5**, 1780–1815.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learning Res.,* **3**, 1157–1182.

Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics,* **12**, 55–67.

Huo,X. and Ni,X. (2007) When do stepwise algorithms meet subset selection criteria? *Ann. Stat.,* **35**, 870–887.

Legendre,A.M. (1805) *Nouvelles Méthodes Pour la Détermination des Orbites des Cometes.* F. Didot, Paris.

Li,F. and Zhang,N.R. (2010) Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Am. Stat. Assoc.,* **105**, 1202–1214.

Loh,P.-R. *et al.* (2011) Phenotype prediction using regularized regression on genetic data in the dream5 systems genetics b challenge. *PLoS ONE,* **6**, e29095.

Malzahn,D. and Opper,M. (2005) A statistical physics approach for the analysis of machine learning algorithms on real data. *J. Stat. Mech.: Theory Exp.,* **2005**, P11001.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.,* **9**, 356–369.

Mézard,M. *et al.* (2002) Analytic and algorithmic solution of random satisfiability problems. *Science,* **297**, 812–815.

Monasson,R. *et al.* (1999) Determining computational complexity from characteristic phase transitions. *Nature,* **400**, 133–137.

Nemenman,I. and Bialek,W. (2002) Occam factors and model independent Bayesian learning of continuous distributions. *Phys. Rev. E,* **65**, 026137.

O'Hagan,A. *et al.* (2004) *Bayesian Inference*. Arnold, London.

Opper,M. and Winther,O. (2001) 2 from naive mean field theory to the tap equations. In: *Advanced Mean Field Methods: Theory and Practice*, Opper,M. and Saad,D. (eds), MIT Press, Cambridge, MA, pp. 7–20.

Peng,J. *et al.* (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.,* **4**, 53.

Penrose,K.W. *et al.* (1985) Generalized body composition prediction equation for men using simple measurement techniques. *Med. Sci. Sports Exerc.,* **17**, 189.

Prill, R.J. *et al.* (2010) Towards a rigorous assessment of systems biology models: the dream3 challenges. *PLoS ONE,* **5**, e9202.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA,* **102**, 15545–15550.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B,* **58**, 267–288.

Varga,R.S. (2010) *Geršgorin and His Circles*. Springer Science & Business, Heidelberg.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics,* **25**, 714–721.

Yang,J. *et al.* (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.,* **44**, 369–375.

Zhou,L. *et al.* (2009) Infection and genotype remodel the entire soybean transcriptome. *BMC Genomics,* **10**, 49.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B,* **67**, 301–320.