Abstract

In this paper, I present a new solution method for sparse regression using L0 regularization. The model introduces a sparseness mechanism in the likelihood, instead of in the prior, as is done in the spike and slab model. The posterior probability is computed in the variational approximation. The variational parameters appear in the approximate model in a way that is similar to Breiman’s Garrote model. I refer to this method as the variational Garrote (VG). The VG is compared numerically with the Lasso method and with ridge regression. Numerical results on synthetic data show that the VG yields more accurate predictions and more accurately reconstructs the true model than the other methods. The naive implementation of the VG scales cubic with the number of features. By introducing Lagrange multipliers we obtain a dual formulation of the problem that scales cubic in the number of samples, but close to linear in the number of features.
The Variational Garrote

H.J. Kappen
Donders Institute for Neuroscience
Radboud University, Nijmegen, the Netherlands

December 9, 2011

1 Introduction

One of the most common problems in statistics is linear regression. Given $p$ samples of $n$-dimensional input data $x_i^\mu$, $i = 1, \ldots, n$ and 1-dimensional output data $y^\mu$, with $\mu = 1, \ldots, p$, find weights $w_i, w_0$ that best describe the relation

$$y^\mu = \sum_{i=1}^{n} w_i x_i^\mu + w_0 + \xi^\mu$$

for all $\mu$. $\xi^\mu$ is zero-mean noise with inverse variance $\beta$.

The ordinary least square (OLS) solution is given by $\vec{w} = \chi^{-1} \vec{b}$ and $w_0 = \bar{y} - \sum_i w_i \bar{x}_i$, where $\chi$ is the input covariance matrix $\vec{b}$ is the vector of input-output covariances and $\bar{x}_i, \bar{y}$ are the mean values. There are several problems with the OLS approach. When $p$ is small, it typically has a low prediction accuracy due to overfitting. In particular, when $p < n$, $\chi$ is not of maximal rank and so its inverse is not defined. In addition, the OLS solution is not sparse: it will find a solution $w_i \neq 0$ for all $i$. Therefore, the interpretation of the OLS solution is often difficult.

These problems are well-known, and there exist a number of approaches to overcome these problems. The simplest approach is called ridge regression. It adds a regularization term $\frac{1}{2} \lambda \sum_i w_i^2$ with $\lambda > 0$ to the OLS criterion. This has the effect that the input covariance matrix $\chi$ gets replaced by $\chi + \lambda I$ which is of maximal rank for all $p$. One optimizes $\lambda$ by cross validation. Ridge regression improves the prediction accuracy but not the interpretability of the solution.

Another approach is Lasso [1]. It solves the OLS problem under the linear constraint $\sum_i |w_i| \leq t$. This problem is equivalent to adding an $L_1$ regularization term $\lambda \sum_i |w_i|$ to the OLS criterion. The optimization of the quadratic error under linear constraints is called a quadratic program. There exist efficient methods to solve this quadratic programming problem. See [2] for a recent account. Again, $\lambda$ or $t$ are found through cross validation. The advantage of the $L_1$ regularization is that the solution tends to be sparse. This improves both the prediction accuracy and the interpretability of the solution.

The $L_1$ or $L_2$ regularization terms are known as shrinkage priors because their effect is to shrink the size of $w_i$. The idea of shrinkage prior has been generalized by [3] to the form $\lambda \sum_i |w_i|^q$ with $q > 0$ and $q = 1, 2$ corresponding to the Lasso and ridge case, respectively. Better solutions can be obtained for $q < 1$, however the resulting optimization problem is no longer convex and therefore more difficult.

An alternative Bayesian approach to obtain a sparse solution was proposed by [4]. They introduce $n$ variational selector variables $s_i$ such that the prior distribution over $w_i$ is a mixture of a narrow (spike) and wide (slab) Gaussian distribution, both centered on zero. The posterior distribution over $s_i$ indicates whether the solution $w_i$ is non-zero or zero. The computation of the posterior requires the use of Gibbs sampling or any other MCMC method.

Here, we propose a model where we introduce sparseness in the likelihood instead of in the prior: $\Box$

$$y^\mu = \sum_{i=1}^{n} w_i s_i x_i^\mu + \xi^\mu$$

(1)

with $s_i = 0, 1$. The idea of this model is that the bits $s_i$ will identify the predictive inputs $i$. When we optimize the posterior distribution with respect to $s_i$ or using Bayesian integration, one finds the optimal subset of

---

We assume from here on without loss of generality that $\frac{1}{p} \sum_{\mu=1}^{p} x_i^\mu = \frac{1}{p} \sum_{\mu=1}^{p} y^\mu = 0$
relevant features. Since the number of subsets is exponential in \( n \) one could use MCMC sampling. Here, instead we propose to approximate the posterior distribution over \( s_i \) and \( \bar{w} \) using a variational approximation.

Model Eq. 1 has some similarity with Breiman’s non-negative Garrote method [4]. The non-negative Garrote method assumes the same model as in Eq. 1 but with \( 0 \leq s_i \leq 1 \) instead of binary. It computes \( w_i \) using OLS and then finds \( s_i \) by minimizing

\[
\sum_{\mu} \left( y^\mu - \sum_{i=1}^{n} x_i^\mu w_i s_i \right)^2 \quad \text{subject to} \quad s_i \geq 0 \quad \sum_i s_i \leq t
\]

Because of this similarity, we refer to our method as the variational Garrote (VG). As we will see, the VG does not require the OLS solution and can find good solutions also when \( p < n \).

Using a Bayesian description, and denoting the data by \( D : \{ \bar{x}^\mu, y^\mu \}, \mu = 1, \ldots, p \), the likelihood term is given by

\[
p(y|\bar{x}, \bar{s}, \bar{w}, \beta) = \sqrt{\frac{1}{2\pi}} \exp \left( -\frac{1}{2} \left( y - \sum_{i=1}^{n} w_i s_i \bar{x}_i \right)^2 \right)
\]

\[
p(D|\bar{s}, \bar{w}, \beta) = \prod_{\mu} p(y^\mu|\bar{x}^\mu, \bar{s}, \bar{w}, \beta) = \left( \frac{\beta}{2\pi} \right)^{p/2} \exp \left( -\frac{\beta p}{2} \left( \sum_{i,j=1}^{n} s_i s_j w_i w_j \chi_{ij} - 2 \sum_{i=1}^{n} w_i s_i b_i + \sigma_y^2 \right) \right) \quad (2)
\]

with \( b_i = \frac{1}{p} \sum_{\mu} x_i^\mu y^\mu, \sigma_y^2 = \frac{1}{p} \sum_{\mu} (y^\mu)^2, \chi_{ij} = \frac{1}{p} \sum_{\mu} x_i^\mu x_j^\mu \).

We should also specify prior distributions over \( s_i, \bar{w}, \beta \). For concreteness, we assume that the prior over \( s_i \) is factorized over the individual \( s_i \), each with identical prior probability:

\[
p(s_i|\gamma) = \prod_{i=1}^{n} p(s_i|\gamma) \quad p(s_i|\gamma) = \frac{\exp(\gamma s_i)}{1 + \exp(\gamma)} \quad (3)
\]

with \( \gamma \) given which specifies the sparsity of the solution. We denote by \( p(\bar{w}, \beta) \) the prior over the inverse noise variance \( \beta \) and the feature weights \( \bar{w} \). We will leave this prior unspecified since it can be chosen arbitrary.

The posterior becomes

\[
p(\bar{s}, \bar{w}, \beta|D, \gamma) = \frac{p(\bar{w}, \beta)p(\bar{s}|\gamma)p(D|\bar{s}, \bar{w}, \beta)}{p(D|\gamma)} \quad (4)
\]

Computing the MAP estimate or computing statistics from the posterior is complex due to the discrete nature of \( \bar{s} \). We propose to compute a variational approximation to the marginal posterior \( p(\bar{w}, \beta|D, \gamma) = \sum_{\bar{s}} p(\bar{s}, \bar{w}, \beta|D, \gamma) \) and computing the MAP solution with respect to \( \bar{w}, \beta \). Since \( p(D|\gamma) \) does not depend on \( \bar{w}, \beta \) we can ignore it.

Note, that the model as specified by the likelihood Eq. 2 and the prior Eq. 3 is not equivalent to the spike and slab model [4]. The spike and slab model has a quadratic likelihood \( p(D|\bar{w}, \beta) \) that does not depend on selector variables. Instead, it has a prior \( p(\bar{w}|\bar{s}) \) that depends on the \( \bar{s} \) of the form

\[
\exp \left( -\frac{1}{2} \sum_{i=1}^{n} s_i \beta_+ w_i^2 - \frac{1}{2} \sum_{i=1}^{n} (1 - s_i) \beta_- w_i^2 \right)
\]

with \( \beta_+ \) the spike and slab precisions. Thus, the spike and slab model only contains an interaction between \( w_i \) and \( s_i \) of the form \( s_i w_i^2 \), whereas the likelihood Eq. 2 contains in addition interactions of the form \( w_i s_i \).

2 The variational approximation

The posterior distribution Eq. 4 for given \( \bar{w}, \beta \) is a typical Boltzmann distribution involving terms linear and quadratic in \( s_i \). It is well-known that one can obtain good approximations using methods that originated in the statistical physics community and where \( s_i \) denote binary spins. Most prominently, one can use the mean field or variational approximation [6], or belief propagation [7]. For introductions into these methods also see [8, 9]. Here, we will develop a solution based on the simplest possible variational approximation and leave the possible improvements using BP or structured mean field approximations to the future.
We approximate the sum by the variational bound using Jensen’s inequality.

\[
\log \sum_{\tilde{s}} p(\tilde{s}|\gamma)p(D|\tilde{s}, \tilde{w}, \beta) \geq - \sum_{\tilde{s}} q(\tilde{s}) \log \frac{q(\tilde{s})}{p(\tilde{s}|\gamma)p(D|\tilde{s}, \tilde{w}, \beta)} = -F(q, \tilde{w}, \beta)
\]  \hspace{1cm} (5)

$q(\tilde{s})$ is called the variational approximation and can be any positive probability distribution on $\tilde{s}$ and $F(q, \tilde{w}, \beta)$ is called the variational free energy. The optimal $q(\tilde{s})$ is found by minimizing $F(q, \tilde{w}, \beta)$ with respect to $q(\tilde{s})$ so that the tightest bound - best approximation - is obtained.

In order to be able to compute the variational free energy efficiently, $q(\tilde{s})$ must be a tractable probability distribution, such as a chain or a tree with limited tree-width [10]. Here we consider the simplest case where we assume a flat prior $p(\tilde{w})$, since these are the only terms that depend on $\tilde{s}$. Thus, the variational approximation does not depend on the particular choices for the prior $p(\tilde{w}, \beta)$. For concreteness, we assume a flat prior $p(\tilde{w}, \beta) \propto 1$. We set the derivatives of $F$ with respect to $\tilde{m}, \tilde{w}, \beta$ equal to zero. This gives the following set of fixed point equations:

\[
m_i = \sigma \left( \gamma + \frac{\beta p}{2} w_i^2 \chi_{ii} \right) \hspace{1cm} \text{(8)}
\]

\[
\tilde{w} = (\chi')^{-1} \tilde{b} \hspace{1cm} \chi'_{ij} = \chi_{ij} m_j + (1 - m_j) \delta_{ij} \hspace{1cm} \text{(9)}
\]

\[
\frac{1}{\beta} = \sigma^2 y - \sum_{i=1}^n m_i w_i b_i \hspace{1cm} \text{(10)}
\]

with $\sigma(x) = (1 + \exp(-x))^{-1}$ and where in Eq. [10] we have used Eq. [9]. Eqs. [8][10] provide the solution to the variational Garrotes. They can be solved by fixed point iteration as outlined in Appendix [C]. Initialize $\tilde{m}$ at random. Compute $\tilde{w}$ by solving the linear system Eq. [9] and $\beta$ from Eq. [10]. Compute a new solution for $\tilde{m}$ from Eq [8].

Within the variational/MAP approximation the predictive model is given by

\[
y = \sum_i m_i w_i x_i + \xi \hspace{1cm} \text{(11)}
\]

with $\langle \xi^2 \rangle = 1/\beta$ and $\tilde{m}, \tilde{w}, \beta$ as estimated by the above procedure.

Let us pause to make some observations about this solution. One might naively expect that the variational approximation would simply consist of replacing $w_is_i$ in Eq. [1] by its variational expectation $w_im_i$. If this were the case, $\tilde{m}$ would disappear entirely from the equations and one would expect in Eq. [9] the OLS solution with the normal input covariance matrix $\chi$ instead of the new matrix $\chi'$ (note, that in the special case that $m_i = 1$
for all $i$, $\chi' = \chi$ and Eq. 9 does reduce to the OLS solution). Instead, $\vec{m}$ and $\vec{w}$ are both to be optimized, giving in general a different solution than the OLS solution\footnote{The technical reason that this does not occur is that in the computation of the expectation with respect to the distribution $q$ that results in Eq. 4 one has $\langle s_i s_j \rangle = m_i m_j$ for $i \neq j$, but $\langle s_i^2 \rangle = \langle s_i \rangle = m_i$.}.

When $m_i < 1$, $\chi'$ differs from $\chi$ by adding a positive diagonal to it. This is similar to the mechanism of ridge regression, but with the important difference that the diagonal term depends on $i$ and is dynamically adjusted depending on the solution for $\vec{m}$. Thus, the sparsity prior together with variational approximation provides a mechanism that solves the rank problem. When all $m_i < 1$, $\chi'$ is of maximal rank. Roughly speaking if $\chi$ has rank $p < n$, $\chi'$ can be still of rank $n$ when no more than $p$ of the $m_i = 1$, the remaining $n - p$ of the $m_i < 1$ making up for the rank deficiency. Note, that the size of $m_i$ (and thus the rank of $\chi'$) is controlled by $\gamma$ through Eq. 8.

\subsection{2.1 Orthogonal and univariate case}

In order to obtain further insight in the solution, consider the case in which the inputs are uncorrelated: $\chi_{ij} = \delta_{ij}$. Then $\chi'_{ij} = \delta_{ij}$ and Eq. 4 gives the solution $w_i = b_i$. Eqs. 8 and 10 become

$$m_i = \sigma \left( \gamma + \beta p b_i^2 \right)$$

$$\frac{1}{\beta} = \sigma_y^2 - \sum_i b_i^2 m_i$$

The term $\sum_i b_i^2 m_i$ is the explained variance and is subtracted from the total output variance to give an estimate of the noise variance $1/\beta$. When $m_i = 1$, the explained variance is maximal. The Garrote solution with $0 \leq m_i \leq 1$ has reduced explained variance with (hopefully) a better prediction accuracy and interpretability.

In the 1-dimensional case these equations become

$$m = \sigma \left( \gamma + \frac{p}{2} \frac{\rho}{1 - \rho m} \right) = f(m)$$

$$\frac{1}{\beta} = \sigma_y^2 (1 - m \rho)$$

with $\rho = b_i^2 / \sigma_y^2$ the squared correlation coefficient.

In Eq. 12 we have eliminated $\beta$ and we must find a solution for $m$ for this non-linear equation. We see that it depends on the input-output correlation $\rho$, the number of samples $p$ and the sparsity $\gamma$. For $p = 100$, the solution for different $\rho, \gamma$ is illustrated in figure 1 (see appendix A). Eq. 12 has one or two solutions for $m$, depending on the values of $\gamma, \rho, p$. The two solutions correspond to two local minima of the free energy $F$, which in the univariate case is given by

$$F = \frac{p}{2} \log(1 - \rho m) - \gamma m + \text{const}.$$

where we have omitted terms independent of $m$. The best variational solution is given by the solution with the lowest free energy, indicated by the solid lines in fig. 1 right.

It is interesting to compare the univariate solution with ridge regression, Lasso or Breiman’s Garrote, which was previously done for the latter three methods in \[1\]. Suppose that data are generated from the model $y = wx + \xi$ with $\langle \xi^2 \rangle = \langle x^2 \rangle = 1$. We compare the solutions as a function of $w$. The OLS solution is approximately given by $w_{\text{ols}} \approx (xy) = w$, where we ignore the statistical deviations of order $1/p$ due to the finite data set size. Similarly, the ridge regression solution is given by $w_{\text{ridge}} \approx \lambda w$, with $0 < \lambda < 1$ depending on the ridge prior. The Lasso solution (for non-negative $w$) is given by $w_{\text{lasso}} = (w - \gamma)^+$, with $\gamma$ depending on the $L_1$ constraint. Breiman’s Garrote solution is given by $w_{\text{garrote}} = (1 - \frac{\gamma}{w})^+ w$ \[1\], with $\gamma$ depending on the $L_1$ constraint. The solutions are given in fig. 2. These methods are quantitatively different. The ridge regression solution is off by a constant multiplicative factor. The Lasso solution is zero for small $w$ and for larger $w$ gives a solution that is shifted downwards by a constant factor. Breiman’s Garrote is identical to the Lasso for small $w$ and shrinks less for larger $w$. The VG gives an almost ideal behavior and can be interpreted as a soft version of variable selection: For small $w$ the solution is close to zero and the variable is ignored, and above a threshold it is identical to the OLS solution.
Figure 1: Left: Phase plot $\rho, \gamma$ for $p = 100$ giving the different solutions for $m$. Solid and dashed lines are from Eq. 24 in the range of $\rho$ values where two solutions for $m$ exist. Dotted line is solution for $\gamma$ when $m = 1/2$, to indicate the transition from $m \approx 0$ to $m \approx 1$. In the lower left corner, the unique solution $m \approx 0$ is found. In the top right corner, the unique solution $m \approx 1$ is found. Between the dashed and the solid line, the two solutions $m \approx 0$ and $m \approx 1$ co-exist. Right: Example solutions for $m$ versus $\rho$ for $\gamma = -10, p = 100$ (top) and for $\gamma = -40, p = 100$ (bottom).

Figure 2: Univariate solution for different regression methods. All methods yield a shrunked solution (deviation from diagonal line). Variational Garrote (VG) with $\gamma = -10$ and $p = 100$. Ridge regression with $\lambda = 0.5$. Garrote with $\gamma = 1/4$. Lasso with $\gamma = 0.5$. 
2.2 A dual formulation

The solution of the system of Eqs. 8-10 by fixed point iteration requires the repeated solution of the \( n \) dimensional linear system \( \chi' \vec{w} = \vec{b} \). When \( n > p \), we can obtain a more efficient method using a dual formulation.

We define new variables \( z^\mu = \sum_i m_i w_i x_i^\mu \) and add Lagrange multipliers \( \lambda^\mu \):

\[
F = -\frac{p}{2} \log \frac{\beta}{2\pi} + \frac{\beta}{2} \sum_\mu (z^\mu - y^\mu)^2 + \frac{\beta p}{2} \sum_i m_i (1 - m_i) w_i^2 \chi_{ii}
- \gamma \sum_{i=1}^n m_i + n \log(1 + \exp(\gamma)) + \sum_{i=1}^n (m_i \log m_i + (1 - m_i) \log(1 - m_i))
+ \sum_\mu \lambda^\mu (z^\mu - \sum_i m_i w_i x_i^\mu)
\]

(14)

We show in the Appendix B that the solution is found by solving the system of equations:

\[
\sum_{\nu=1}^p A_{\mu \nu} y^\nu = y^\mu \quad (15)
\]

\[
\frac{1}{\beta} = \frac{1}{p} \sum_\nu y^\nu y^\mu \quad (16)
\]

\[
\lambda^\mu = \beta y^\mu \quad (17)
\]

\[
A_{\mu \nu} = \delta_{\mu \nu} + \frac{1}{p} \sum_i m_i \frac{x_i^\mu x_i^\nu}{1 - m_i} \chi_{ii} \quad (18)
\]

\[
w_i = \frac{1}{\beta p \chi_{ii}} \frac{1}{1 - m_i} \sum_\mu \lambda^\mu x_i^\mu \quad (19)
\]

together with Eq. 8.

2.3 Cross validation

One remaining issue is how to choose the level of sparsity \( \gamma \). If we have a prior belief about \( \gamma \) we could optimize its value together with \( \vec{w}, \beta \). Here, instead, we propose compute the VG solution for fixed \( \gamma \) and choose its optimal value through cross validation on independent data. This has the advantage that our result is independent of our (possibly incorrect) prior belief.

But another important advantage is computational. When we increase \( \gamma \) from a negative value to zero in small steps, we obtain a sequence of solutions with decreasing sparseness. These solutions will better fit the data and as a result \( \beta \) increases with \( \gamma \). Thus, increasing \( \gamma \) implements an annealing approach where we sequentially obtain solutions at lower noise levels. We found empirically that this approach is effective to reduce the problem of local minima. To further deal with the effect of hysteresis we recompute the solution from \( \gamma_{\text{max}} \) down to \( \gamma_{\text{min}} \).

The minimal value of \( \gamma \) is chosen as the largest value such that \( m_i = \epsilon \), with \( \epsilon \) small. We find from Eqs. 8-10 that

\[
\gamma_{\text{min}} = -\frac{pb^2 \chi_{ii}}{2\sigma_y^2} + \sigma^{-1}(\epsilon) + \mathcal{O}(\epsilon)
\]

(20)

We heuristically set the maximal value of \( \gamma \) as well as the step size. The VG algorithm is summarized in Appendix C.

3 Numerical examples

In the following examples, we compare the Lasso, ridge regression and the VG. For each example, we generate a training set, a validation set and a test set. Inputs are generated from a mean zero multi-variate Gaussian distribution with specified covariance structure. We generate outputs \( y^\mu = \sum_i \hat{w}_i x_i^\mu + d\xi^\mu \) with \( d\xi^\mu \in \mathcal{N}(0, \hat{\sigma}) \). For each value of the hyper parameters (\( \gamma \) in the case of VG, \( \lambda \) in the case of ridge regression and Lasso), we
optimize the model parameters on the training set. For the Lasso, we used the method described in [2]. We optimize the hyper parameters that minimize the quadratic error on the validation set. Finally, we report the mean squared error on the training set, validation set and test set as well as the mean absolute error in the estimated parameters. For the VG, we define the solution as \( \hat{w} = m_i w_i \).

In the first example (Example 1), we take independent inputs \( x_i^p \in \mathcal{N}(0, 1) \) and a teacher weight vector with only one non-zero entry: \( \hat{w} = (1, 0, \ldots, 0) \), \( n = 100 \) and \( \delta = 1.0 \). The training set size \( p = 50 \), validation set size \( p_v = 50 \) and test set size \( p_t = 400 \). We choose \( \epsilon = 0.001 \) in Eq. 20, \( \gamma_{\text{max}} = 0.02 \gamma_{\text{min}}, \Delta \gamma = -0.02 \gamma_{\text{min}} \). For details of the algorithm see Appendix C. Results for a single run of the VG are shown in fig. 2. In fig. 2, we plot the minimal variational free energy \( F \) versus \( \gamma \) for both the forward and backward run. Note, the hysteresis effect due to the local minima. For each \( \gamma \), we use the solution with the lowest \( F \). In fig. 2, we plot the training error and validation error versus \( \gamma \). The optimal \( \gamma \approx -14.8 \) is denoted by a star and the corresponding \( \gamma_{\text{min}} \) of the remaining components versus \( \gamma \). Note the robustness of the VG solution in the sense of the large range of \( \gamma \) values for which the correct solution is found. In fig. 3, we plot the optimal solution \( \hat{w} = m_i w_i \) versus \( i \).

In fig. 4 we show the Lasso (top row) and ridge regression (bottom row) results for the same data set. The optimal value for \( \lambda \) minimizes the validation error (star). In fig. 4, we see that the Lasso selects a number of incorrect features as well. Fig. 4 also shows that the Lasso solution with a larger \( \lambda \) in the range \( 0.45 < \lambda < 0.95 \) could select the single correct feature, but would then estimate \( w \) too small due to the large shrinkage effect. Ridge regression gives very bad results. The non-zero feature is too small and the remaining features have large values. The noise \( \hat{\sigma} \) is denoted by a star and the corresponding \( \sigma_{\text{est}} \) of the remaining components versus \( \gamma \). Note from fig. 4, that ridge regression yields a non-sparse solution for all values of \( \lambda \).

In table 1 we compare the performance of the VG, Lasso and ridge regression on 10 random instances. We see that the VG significantly outperforms the Lasso method and ridge regression both in terms of prediction error, the accuracy of the estimation of the parameters and the number of non-zero parameters.

In the second example (Example 2), we consider the effect of correlations in the input distribution. Following [2] we generate input data from a multi-variate Gaussian distribution with covariance matrix \( \chi_{ij} = \delta^{[i-j]} \), with \( \delta = 0.5 \). In addition, we choose multiple features non-zero: \( \hat{w}_i = 1, i = 1, 2, 5, 10, 50 \) and all other \( \hat{w}_i = 0 \). We use \( n = 100, \sigma = 1 \) and \( p/p_t / p_v = 50/50/400 \). In table 2 we compare the performance of the VG, Lasso and ridge regression on 10 random instances. We see that the VG significantly outperforms the Lasso method and ridge regression both in terms of prediction error and accuracy of the estimation of the parameters.

In fig. 5 we show the accuracy of the VG method and the Lasso method as a function of the noise \( \hat{\sigma} \). We generate data according to Example 2 and vary \( \hat{\sigma}^2 \) in the range \( 1e-4 \) to 10. Fig. 5 shows the error \( \| \delta \hat{w} \|_2 \) as

\[ \text{(http://www-stat.stanford.edu/~tibs/glmnet-matlab)} \]
Table 1: Results for Example 1. Train is MSE on the training set. Val is MSE on the validation set. Test is MSE on the test set. # non-zero is number of non-zero elements in the solution for Lasso and is $\sum_{i=1}^{n} |w_i|$ for VG. $\|\delta w\|_1 = \sum_{i=1}^{n} |\hat{w}_i - w_i|$ and $\|\delta w\|_2 = \sum_{i=1}^{n} (\hat{w}_i - w_i)^2$ with $\hat{w}$ the estimated solution for Ridge and Lasso and $\hat{w} = \hat{v}$ for VG.

|       | Train     | Val       | Test      | # non-zero | $\|\delta w\|_1$ | $\|\delta w\|_2$
|-------|-----------|-----------|-----------|-------------|------------------|------------------|
| Ridge | 0.62 ± 0.43 | 1.79 ± 0.31 | 1.77 ± 0.15 | -           | 3.91 ± 0.92  | 0.79 ± 0.07    |
| Lasso | 0.72 ± 0.25 | 1.09 ± 0.25 | 1.20 ± 0.21 | 10.5 ± 6.5 | 1.01 ± 0.60  | 0.19 ± 0.11    |
| VG    | 0.85 ± 0.25 | 0.97 ± 0.20 | 1.03 ± 0.12 | 1.4 ± 0.8  | 0.43 ± 0.40  | 0.04 ± 0.05    |
| True  | 0.93 ± 0.14 | 0.87 ± 0.20 | 0.98 ± 0.04 | 1           | 0              | 0              |

Table 2: Results for Example 2. Train is MSE on the training set. Val is MSE on the validation set. Test is MSE on the test set. # non-zero is number of non-zero elements in the solution for Lasso and is $\sum_{i=1}^{n} (m_i > 0.5)$ for VG. $\|\delta w\|_1 = \sum_{i=1}^{n} |\hat{w}_i - w_i|$ and $\|\delta w\|_2 = \sum_{i=1}^{n} (\hat{w}_i - w_i)^2$ with $\hat{w}$ the estimated solution for Ridge and Lasso and $\hat{w} = \hat{v}$ for VG.

|       | Train     | Val       | Test      | # non-zero | $\|\delta w\|_1$ | $\|\delta w\|_2$
|-------|-----------|-----------|-----------|-------------|------------------|------------------|
| Ridge | 0.29 ± 0.31 | 3.52 ± 0.69 | 3.48 ± 0.35 | -           | 11.3 ± 0.69    | 2.79 ± 0.13    |
| Lasso | 0.78 ± 0.39 | 1.40 ± 0.33 | 1.47 ± 0.21 | 16.4 ± 6.0 | 2.00 ± 0.64    | 0.51 ± 0.19    |
| VG    | 0.83 ± 0.26 | 1.17 ± 0.20 | 1.21 ± 0.17 | 4.9 ± 0.3  | 0.96 ± 0.48    | 0.30 ± 0.36    |
| True  | 0.93 ± 0.14 | 0.87 ± 0.20 | 0.98 ± 0.04 | 5           | 0              | 0              |
defined above. We distinguish three noise domains: for very large noise both VG and LASSO produce errors of $\mathcal{O}(1)$ and fail to find the predictive features. For intermediate and low noise levels, VG is approximately 1 order of magnitude more accurate than Lasso. In the limit of zero noise, the error of VG keeps on decreasing whereas the Lasso error saturates to a constant value.

It is well-known that the Lasso method may yield inconsistent results when input variables are correlated. \cite{efron2004} derive a necessary and sufficient condition for consistency. In addition, they give a number of examples where Lasso gives inconsistent results. Their simplest example has three input variables, $x_1, x_2, x_3, x_1, x_2, \xi, e$ are independent and Normal distributed random variables, $x_3 = 2/3x_1 + 2/3x_2 + \xi$ and $y = \sum_{i=1}^{3} \tilde{w}_ix_i + e$, $p = 1000$. When $\hat{w} = (2, 3, 0)$ this example violates the consistency condition. The Lasso and VG solutions for different values of $\lambda$ and $\gamma$, respectively are shown in fig. 6. The average results over 100 instances of these problems is shown in table 3.

In fig. 7 we show how the Lasso, the ridge regression and the VG behave as a function of the number of samples. The VG is significantly more accurate than the Lasso, for almost all values of $p$. This is evident from the $\|\delta w\|_2$ and validation error. This is also evident by comparing the $L_0$ and $L_1$ norms of the solutions found. The Lasso solution find a solution with the same (correct) $L_1$ norm as the VG, but spreads this solution over many more features, as is witnessed by the $L_0$ norm.

In fig. 8 we show how the Lasso, the ridge regression and the VG behave as a function of the number of samples. For the VG we use the dual method described in section 2.2. The VG has constant quality in terms of the error $\|\delta w\|_2$ and close to optimal norms $L_0 = L_1 = 5$. The computation time of the VG scales approximately linear with $n$. The Lasso is significantly faster, but its quality deteriorates with $n$.  

---

Table 3: Accuracy of Ridge, Lasso and Garrote for Example 1a,b from \cite{efron2004}. $p = p_0 = 1000$. Parameters $\lambda$ (Ridge and Lasso) and $\gamma$ (Garrote) optimized through cross validation. $\|\delta w\|_{1,2}$ as before, $\max(|w_3|)$ is maximum over 100 trials of the absolute value of $w_3$. Example 1a is inconsistent for Lasso and yields much larger errors than the Garrote. Example 1b is consistent and the quality of the Lasso and Garrote are similar. Ridge regression is bad for both examples.

| Example 1a | Example 1b |
|------------|------------|
| $\|dw\|_1$ | $\|dw\|_2$ | $\max(|w_3|)$ | $\|dw\|_1$ | $\|dw\|_2$ | $\max(|w_3|)$ |
| Ridge | 0.6366 ± 0.1811 | 0.1504 ± 0.0787 | 0.48 | 0.1905 ± 0.1113 | 0.0185 ± 0.0206 | 0.27 |
| Lasso | 0.1916 ± 0.1363 | 0.0199 ± 0.0260 | 0.30 | 0.0628 ± 0.0364 | 0.0030 ± 0.0032 | 0 |
| Garrote | 0.0491 ± 0.0257 | 0.0020 ± 0.0018 | $2e - 14$ | 0.0491 ± 0.0257 | 0.0020 ± 0.0018 | $2e - 14$ |

Figure 5: Accuracy of VG and Lasso in terms of the error $\|\delta w\|_2$ as a function of the noise. Data according to Example 2 with $\delta = 0.5$. All results are averages over 10 runs. The variance in these results is of order 50.
Figure 6: Top left) Lasso solution versus $\lambda$ is called inconsistent because it does not contain a $\lambda$ for which the correct sparsity $w(3) = 0$ is obtained. Top right) Garrote solution for $v$ versus $\gamma$ contains large range of $\gamma$ for which the correct solution is obtained. Bottom left) Garrote solution for $m$. Curves for $m_1,2$ are identical. Bottom right) Garrote solution for $w$. 
Figure 7: Performance of different methods as a function of the number of samples $p$. Data are generated as in Example 2. $n = 500$ features and $\beta = 1, \delta = 0.5$. The number of samples is varied from $p = [10, 20, 50, 100, 250, 500]$. Optimal hyper parameters found through cross validation. Cross validation set of size $p_v = 50$. Results are average over 10 runs. Top left: Quadratic error $\|\delta w\|_2 = \sum_{i=1}^{n}(\hat{w}_i - w_i)^2$. Top right: $L_0$ norm of the solution found. For the VG the number of non-zero components is defined as those $i$ for which $m_i > 0.5$. Bottom left: $L_1$ norm of the solution found. For the VG the $L_1$ norm of the vector with components $w_i m_i$ is used. Bottom right: validation error.
Figure 8: Performance of different methods as a function of the number of features $n$. Data are generated as in Example 2. $p = 100, \beta = 2, \delta = 0$. Optimal hyper parameters are found through cross validation. Cross validation set of size $p_v = 100$. The number of features is varied in steps of 100 from $n = 100$ to $n = 1000$, and $n = 2000$ and $n = 5000$. For each value averages over 10 runs are plotted. Top left: Quadratic error $\|\delta w\|_2^2 = \sum_{i=1}^{n}(\hat{w}_i - w_i)^2$. Top right: $L_0$ norm of the solution found. For the VG the number of non-zero components is defined as those $i$ for which $m_i > 0.5$. Bottom left: $L_1$ norm of the solution found. For the VG the $L_1$ norm of the vector with components $w_i m_i$ is used. Bottom right: Matlab CPU time in seconds for VG and C implementation CPU time in seconds for Lasso.
In this paper, I have introduced a new model for sparse regression by introducing binary switch variables $\bar{s}$ that multiply the feature weights. When using a Bayesian framework one obtains a posterior distribution that is intractable due to the binary nature of $\bar{s}$. In this paper we have proposed to approximate the variables $\bar{s}$ using a variational approximation. We have demonstrated that this approach is efficient and accurate and yields results that significantly outperform the Lasso approach.

There are many points that need further investigation. The variational approximation can by replaced by any other approximate inference method. It would be interesting to see whether the results could be further improved by using Belief Propagation, the Cluster Variation Method or Monte Carlo Sampling.

I have not explored the use of different priors on $\bar{w}$ or on $\beta$. In addition, a prior could be imposed on $\gamma$. It is likely that for particular problems the use of a suitable prior could further improve the results.

One of the important advantages of the Lasso over the VG is that the Lasso problem is convex whereas the VG may have local minima, as is illustrated in fig. 1. This is clearly a disadvantage of the VG, but is shared by all other methods that use $L_0$ regularization. We have not much explored the issue of local minima. For the examples that we presented the annealing approach that results from increasing $\gamma$, followed by a 'heating' phase to detect hysteresis, seems to work well. We have noted that for problems with a large number of features ($O(n)$), small $p$ and high noise this approach may fail. The local minima problem depends on the approximation, and may be more or less severe when using other approximation schemes.

For efficient implementation one may use the fact that the data matrix $x_i^m$ is sparse. For instance in genetic regression problems this may be the case, when $x_i$ denotes single nucleotide polymorphisms (SNPs) that have values $x_i = 0, 1, 2$. For each $i$ one may code them such that the value $x_i^m = 0$ is the most frequent one. In that case, substracting the mean from the data such that $\sum_{\mu} x_i^m = 0$ will yield $x_i^m$ non-sparse. The model as presented here should then be slightly generalized to estimate a constant bias as well.

In principle, the model Eq. 11 when used in a full Bayesian setting is non-linear: The predictive distribution

$$p(y|x, D) = \int d\bar{w}d\beta \sum_{\bar{s}} p(y|\bar{s}, \bar{w}, \beta, \bar{x}) p(\bar{s}, \bar{w}, \beta|D)$$  \hspace{1cm} (21)$$

is a mixture of $2^n$ Gaussians. Only in the variational/MAP approximation does this model reduce to the linear model Eq. 11. It would be of interest to explore the Bayesian extension of the VG. However note, that in our experiments we find that $n_i \approx 0$ or $n_i \approx 1$ most of the time and encounter very few intermediate or multi-modal solutions. This suggest that the Bayesian integration, at least as far as the binary variables is concerned may be of limited additional value for the examples considered in the paper.

We have seen, that the performance of the VG is excellent also in the zero noise limit. In this limit, the regression problem reduces to a compressed sensing problem \cite{12,13}. The performance of compressed sensing with $L_0$ sparseness penalty was analyzed theoretically in \cite{14}, showing the superiority of the $L_1$ penalty in comparison to the $L_2$ penalty and suggesting the optimality of the $L_0$ penalty. Our numerical result are in agreement with this finding.

References

[1] R. Tibshirani. Regression shrinkage and regression via the Lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

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

[3] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.

[4] E. George and R. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:884–889, 1993.

[5] L. Breiman. Better subset selection using the non-negative garotte. *Technometrics*, 37:373–384, 1995.

[6] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in graphical models*. MIT press, 1998.

[7] Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475, 1999.

[8] M. Opper and D. Saad. *Advanced mean field theory*. MIT Press, 2001.
Figure 9: $f(m)$ vs $m$. Left: $p = 100, \gamma = -10$, different lines correspond to different values of $0 < \rho < 1$. The solution for $m$ is given by the intersection $f$ with the diagonal line. The solution for $m$ is unique and increases with increasing $\rho$. Right: Same as left, but with $p = 100, \gamma = -30$. For given $\rho$ there are three solutions for $m$. The solutions close to $m \approx 0, 1$ correspond to local minima of $G$. The intermediate solution corresponds to a local maximum of $G$.

[9] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families and variational inference. *Foundations and trends in machine learning*, 1:1–305, 2008.

[10] D. Barber and W. Wiegerinck. Tractable variational structures for approximating graphical models. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11 of *Advances in Neural Information Processing Systems*, pages 183–189. MIT Press, 1999.

[11] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.

[12] E.J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[13] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

[14] Y. Kabashima, T. Wadayama, and T. Tanaka. A typical reconstruction limit for compressed sensing based on $l_p$-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, page L09003, 2009.

A Phase plot computation for the univariate case

$f(m)$ in Eq. 12 is an increasing function of $m$ and crosses the line $m$ either 1 or three times, depending on the values of $p$ and $\gamma$ (see fig. 9).

Thus, there are regions of parameter space $\gamma, p, \rho$ where the univariate solution is unique and others for which there are two solutions. The transition between these two regions is when $f'(m) = 1$ and $f(m) = m$. These two equations imply

$$(1 + \frac{1}{2}p\rho^2)m^2 + (-2\rho - \frac{1}{2}p\rho^2)m + 1 = 0 \quad (22)$$

This quadratic equation in $m$ has either zero, one or two solutions, corresponding to no touching, touching once and touching twice, respectively. Denote $a = 1 + \frac{1}{2}p\rho^2, b = 2\rho + \frac{1}{2}p\rho^2$. The critical value for $\rho, p$ is when Eq. 22 has one solution for $m$, which occurs when

$$D = b^2 - 4a = \frac{1}{4}p^2(1 - \rho) + 4(\rho^2 - 1) = 0$$

This is a quadratic equation in $p$ and we solve for $p$ in terms of $\rho$:

$$p = \frac{4}{\rho^2} \sqrt{1 - \rho} \left( \sqrt{1 - \rho} \pm \sqrt{2} \right)$$
Since $\rho \leq 1$ there is a positive and a negative solution for $p$. The latter we disregard.

$$p^* = \frac{4}{\rho^2} \sqrt{1 - \rho} \left( \sqrt{1 - \rho + \sqrt{2}} \right)$$  (23)

$p^*$ is a decreasing function of $\rho$.

For $p > p^*$, $D > 0$ and we find two solutions for $m$, which we denote by $m_{1,2} = \frac{b \pm \sqrt{D}}{2a}$. Note, that the solutions in these critical points only depend on $\rho, p$. For each of these solutions we must find $\gamma$ such that $f(m) = m$, which is given by

$$\gamma_i = \log \frac{m_i}{1 - m_i} - \frac{p}{2} \frac{1 - \rho m_i}{1 - \rho m_i} \quad i = 1, 2$$  (24)

For $p < p^*$, $D < 0$ and we find no solutions for $m$. In this case the conditions $f'(m) = 1$ and $f(m) = m$ cannot be jointly satisfied.

We illustrate the phase plot $\rho, \gamma$ for $p = 100$ in fig. 1.

**B Derivation of dual fixed point equations**

We compute the derivatives of Eq. 14:

$$\frac{\partial F}{\partial w_i} = m_i \left( \beta p (1 - m_i) \chi_{ii} w_i - \sum_{\mu} \lambda^\mu x_i^\mu \right)$$

$$\frac{\partial F}{\partial z^\mu} = \beta (z^\mu - y^\mu) + \lambda^\mu$$

$$\frac{\partial F}{\partial \beta} = -\frac{p}{2\beta} + \frac{1}{2} \sum_{\mu} (z^\mu - y^\mu)^2 + \frac{p}{2} \sum_i m_i (1 - m_i) w_i^2 \chi_{ii}$$

$$\frac{\partial F}{\partial m_i} = \frac{\beta p}{2} (1 - 2m_i) w_i^2 \chi_{ii} - \gamma + \sigma^{-1}(m_i) - \sum_{\mu} \lambda_{\mu} w_i x_i^\mu$$

$$\frac{\partial F}{\partial \lambda^\mu} = z^\mu - \sum_i m_i w_i x_i^\mu$$

By setting $\frac{\partial F}{\partial w_i} = \frac{\partial F}{\partial z^\mu} = 0$ we obtain Eq. 19 and $z^\mu = y^\mu - \frac{1}{\beta} \lambda^\mu$. Setting the remaining derivatives to zero, and eliminating $w_i$ and $z^\mu$ we obtain Eq. 8 and

$$\beta = \frac{1}{p} \sum_{\mu \nu} \lambda_{\mu} \lambda_{\nu} A_{\mu \nu}$$  (25)

$$\beta y^\mu = \sum_{\nu} A_{\mu \nu} \lambda^\nu$$  (26)

with $A_{\mu \nu}$ given by Eq. 18. For given $A_{\mu \nu}$, let $\hat{y}$ denote the solution of $A \hat{y} = y$. Then it is easy to verify that Eqs. 16, 17 solve the system of Eqs. 25, 26.
C  The Variational Garrote algorithm

Here we give a summary of the VG algorithm.

1. Preprocess data $x_i^\mu, y^\mu, \mu = 1, \ldots, p$ such that $\sum_\mu x_i^\mu = \sum_\mu y^\mu = 0$.
2. Compute $b_i = \frac{1}{p} \sum_\mu x_i^\mu y^\mu$
3. For $n < p$, compute $\chi_{ij} = \frac{1}{p} \sum_\mu x_i^\mu x_j^\mu$.
4. Choose $\epsilon$ and compute $\gamma_{\min}$ from Eq. 20.
5. Initialize all $m_i = 0$.
6. For $\gamma = \gamma_{\min}: \Delta \gamma: \gamma_{\max}$ do
   a) Set $\eta = 1$.
   b) While not converged:
      i. Compute $\vec{w}, \beta$ from Eqs. 9-10 ($n < p$) or Eqs. 15-19 ($n > p$).
      ii. Compute $m$ using a smoothed version of Eq. 8: $m'_i = (1 - \eta)m_i + \eta \sigma(\ldots)$.
      iii. When max$_i |m'_i - m_i| > 0.1$ divide $\eta$ by two.
      iv. $\vec{m} = \vec{m}'$.
   c) Store solution as $\vec{w}_1, \vec{m}_1, \beta_1 |\gamma$.
   d) Compute $F_1(\gamma) = F(\vec{w}_1, \vec{m}_1, \beta_1 |\gamma)$ from Eq. 7.
7. For $\gamma = \gamma_{\max}: -\Delta \gamma: \gamma_{\min}$ do
   a) As 6 (a)
   b) As 6 (b)
   c) Store solution as $\vec{w}_2, \vec{m}_2, \beta_2 |\gamma$.
   d) Compute $F_2(\gamma) = F(\vec{w}_2, \vec{m}_2, \beta_2 |\gamma)$ from Eq. 7.
8. For $\gamma = \gamma_{\min}: \Delta \gamma: \gamma_{\max}$
   a) choose solution $\vec{w}, \vec{m}, \beta |\gamma$ that has minimal $F_{1,2}(\gamma)$.
   b) compute cross validation error, defined as the quadratic error on validation data using Eq. 11.
9. Output solution $\vec{w}, \vec{m}, \beta, \gamma$ with minimal cross validation error.
