– **manuscript No.**
(will be inserted by the editor)

The Variational Garrote

Hilbert J. Kappen · Vicenç Gómez

Received: date / Accepted: date

Abstract In this paper, we present a new variational method for sparse regression using L_0 regularization. The variational parameters appear in the approximate model in a way that is similar to Breiman's Garrote model. We refer to this method as the variational Garrote (VG). We show that the combination of the variational approximation and L_0 regularization has the effect of making the problem effectively of maximal rank even when the number of samples is small compared to the number of variables. The VG is compared numerically with the Lasso method, ridge regression and the recently introduced paired mean field method (PMF) [1]. Numerical results show that the VG and PMF yield more accurate predictions and more accurately reconstruct the true model than the other methods. It is shown that the VG finds correct solutions when the Lasso solution is inconsistent due to large input correlations. Globally, VG is significantly faster than PMF and tends to perform better as the problems become denser and in problems with strongly correlated inputs. 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.

Hilbert J. Kappen
Donders Institute for Brain Cognition and Behaviour
Radboud University Nijmegen
6525 EZ Nijmegen, The Netherlands
E-mail: b.kappen@science.ru.nl

Vicenç Gómez
Donders Institute for Brain Cognition and Behaviour
Radboud University Nijmegen
6525 EZ Nijmegen, The Netherlands
E-mail: v.gomez@science.ru.nl

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^μ, with $\mu = 1, \ldots, p$, find weights w_i, w_0 that best describe the relation

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

for all μ. ξ^μ is zero-mean noise with inverse variance β.

The ordinary least square (OLS) solution is given by $\mathbf{w} = \chi^{-1}\mathbf{b}$ and $w_0 = \bar{y} - \sum_i w_i \bar{x}_i$, where χ is the input covariance matrix $\mathbf{b}$ 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 over fitting. In particular, when $p < n$, χ is not of maximal rank and so its inverse is not uniquely 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 χ gets replaced by $\chi + \lambda I$ which is of maximal rank for all p. One optimizes λ by cross validation. Ridge regression improves the prediction accuracy but not the interpretability of the solution.

Another approach is Lasso [2]. 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 can be solved efficiently. See [3] for a recent account. Again, λ or t may be 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 [4] 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 to solve.

An alternative Bayesian approach to obtain a sparse solution using an L_0 penalty was proposed by [5]. 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 input feature i is included in the model or not. Since the number of subsets of features is exponential in n, for large n one cannot compute the solution exactly. In addition, the posterior is a complex high dimensional distribution of the w_i and the other (hyper)

parameters of the model. The computation of the posterior requires thus the use of MCMC sampling [5] or a variational Bayesian approximation [1,6,7,8].

Although Bayesian approaches tend to over fit less than a maximum likelihood or maximum a posteriori method (MAP approach), they also tend to be relatively slow. Here we propose a partial Bayesian approach, where we apply a variational approximation to integrate out the binary (selector) variables in combination with a MAP approach for the remaining parameters. For clarity, we analyse this idea in its most simple form, in the absence of (hierarchical) priors. Instead, we infer the sparsity prior through cross validation. As we will motivate below, we call the method the Variational Garrote (VG).

The paper is organized as follows. In section 2 we introduce the model and we derive the variational approximation. We show that the combination of the variational approximation and L_0 regularization has the effect of making the problem effectively of maximal rank by introducing a 'variational ridge term'. As a result, the solution is well defined even when $p < n$ as long as the number of predictive features is less than p (which is controlled by the sparsity prior).

To gain further insight, in section 3 we study the case when the design matrix is orthogonal. In this case the solution can be computed exactly in closed form with no need to resort to approximations. In the variational approximation, we show for the uni-variate case that the solution is either unique or has two solutions, depending on the input-output correlations, the number of samples p and on the sparsity prior γ. We derive a phase plot and show that the solution is unique, when the sparsity prior is not too strong *or* when the input-output correlation is not too large. The input-output behavior of the VG is shown to be close to optimal as a smoothed version of hard feature selection. We argue that this behavior also holds in the multi-variate case.

In section 4 we compare the VG with a number of other MAP methods, such as Lasso and ridge regression and with the paired mean field method (PMF) [1], a recently proposed variational bayesian method. We show that the VG and PMF significantly outperform the Lasso and ridge regression on a large number of different examples both in terms of the accuracy of the solution, as well as in prediction error. In addition, we show that the VG do not suffer from the inconsistency of the Lasso method when the input correlations are large. We show in detail how all methods compare as a function of the level of noise, the sparsity of the target solution, the number of samples and the number of irrelevant predictors. Globally, VG is significantly faster than PMF and tends to perform better as the problems become denser and in problems with strongly correlated inputs.

2 The variational approximation

Consider the regression model of the form [1]

$$y^\mu = \sum_{i=1}^n w_i s_i x_i^\mu + \xi^\mu \qquad \sum_{i=1}^n s_i \leq t \tag{2}$$

with $s_i = 0, 1$. The bits $s_i = 1$ will identify the predictive inputs i. Using a Bayesian description, and denoting the data by $D : \{\mathbf{x}^\mu, y^\mu\}, \mu = 1, \ldots, p$, the likelihood term is given by

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{s}, \mathbf{w}, \beta) &= \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}\left(y - \sum_{i=1}^n w_i s_i x_i\right)^2\right) \\ p(D|\mathbf{s}, \mathbf{w}, \beta) &= \prod_\mu p(y^\mu|\mathbf{x}^\mu, \mathbf{s}, \mathbf{w}, \beta) \\ &= \left(\frac{\beta}{2\pi}\right)^{p/2} \exp\left(-\frac{\beta p}{2}\left(\sum_{i,j=1}^n s_i s_j w_i w_j \chi_{ij} - 2\sum_{i=1}^n w_i s_i b_i + \sigma_y^2\right)\right) \end{aligned} \tag{3}$$

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 $\mathbf{s}, \mathbf{w}, \beta$. For concreteness, we assume that the prior over $\mathbf{s}$ is factorized over the individual s_i, each with identical prior probability:

$$p(\mathbf{s}|\gamma) = \prod_{i=1}^n p(s_i|\gamma) \qquad p(s_i|\gamma) = \frac{\exp(\gamma s_i)}{1 + \exp(\gamma)} \tag{4}$$

with γ given which specifies the sparsity of the solution. We denote by $p(\mathbf{w}, \beta)$ the prior over the inverse noise variance β and the feature weights $\mathbf{w}$. We will leave this prior unspecified since its choice does not affect the variational approximation. [2]

The posterior becomes

$$p(\mathbf{s}, \mathbf{w}, \beta|D, \gamma) = \frac{p(\mathbf{w}, \beta)p(\mathbf{s}|\gamma)p(D|\mathbf{s}, \mathbf{w}, \beta)}{p(D|\gamma)} \tag{5}$$

Computing the MAP estimate or computing statistics from the posterior is complex in particular due to the discrete nature or $\mathbf{s}$. We propose to compute a variational approximation to the marginal posterior $p(\mathbf{w}, \beta|D, \gamma) =$

[1] We assume from here on without loss of generality that $\frac{1}{p}\sum_{\mu=1}^p x_i^\mu = \frac{1}{p}\sum_{\mu=1}^p y^\mu = 0$

[2] It can be shown that the regression model specified by Eqs. 3 and 4 is identical to the spike and slab model, with the difference that the latter usually contains a (Gaussian) prior over the w_i which could also be added in the above representation[1]. See appendix C for details.

$\sum_{\mathbf{s}} p(\mathbf{s}, \mathbf{w}, \beta|D, \gamma)$ and computing the MAP solution with respect to $\mathbf{w}, \beta$. Since $p(D|\gamma)$ does not depend on $\mathbf{w}, \beta$ we can ignore it.

The posterior distribution Eq. 5 for given $\mathbf{w}, \beta$ is a typical Boltzmann distribution involving terms linear and quadratic in s_i. It is well-known that when the effective couplings $w_i w_j \chi_{ij}$ are small, 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 [9], the TAP approximation [10] or belief propagation (BP) [11]. For introductions into these methods also see [12,13]. 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.

$$\begin{aligned} \log \sum_{\mathbf{s}} p(\mathbf{s}|\gamma) p(D|\mathbf{s}, \mathbf{w}, \beta) &\geq -\sum_{\mathbf{s}} q(\mathbf{s}) \log \frac{q(\mathbf{s})}{p(\mathbf{s}|\gamma) p(D|\mathbf{s}, \mathbf{w}, \beta)} \\ &= -F(q, \mathbf{w}, \beta) \end{aligned} \tag{6}$$

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

In order to be able to compute the variational free energy efficiently, $q(\mathbf{s})$ must be a tractable probability distribution, such as a chain or a tree with limited tree-width [14]. Here we consider the simplest case where $q(\mathbf{s})$ is a fully factorized distribution: $q(\mathbf{s}) = \prod_{i=1}^{n} q_i(s_i)$ with $q_i(s_i) = m_i s_i + (1 - m_i)(1 - s_i)$, so that q is fully specified by the expected values $m_i = q_i(s_i = 1)$, which we collectively denote by $\mathbf{m}$. The expectation values with respect to q can now be easily evaluated and the result is

$$\begin{aligned} F = &\frac{\beta p}{2} \left(\sum_{i,j}^{n} m_i m_j w_i w_j \chi_{ij} + \sum_{i} m_i (1 - m_i) w_i^2 \chi_{ii} - 2 \sum_{i=1}^{n} m_i w_i b_i + \sigma_y^2 \right) \\ &- \gamma \sum_{i=1}^{n} m_i + \sum_{i=1}^{n} \left(m_i \log m_i + (1 - m_i) \log(1 - m_i) \right) - \frac{p}{2} \log \frac{\beta}{2\pi} \end{aligned} \tag{7}$$

where we have omitted terms independent of m, β, w. The first line is due to the likelihood term, the second line is due to the prior on $\mathbf{s}$ and the entropy of $q(\mathbf{s})$. The approximate marginal posterior is then

$$\begin{aligned} p(\mathbf{w}, \beta|D, \gamma) &\propto p(\mathbf{w}, \beta) \sum_{\mathbf{s}} p(\mathbf{s}|\gamma) p(D|\mathbf{s}, \mathbf{w}, \beta) \\ &\approx p(\mathbf{w}, \beta) \exp(-F(\mathbf{m}, \mathbf{w}, \beta, \gamma)) \end{aligned}$$

We can compute the variational approximation $\mathbf{m}$ for given $\mathbf{w}, \beta, \gamma$ by minimizing F with respect to $\mathbf{m}$. In addition, $p(\mathbf{w}, \beta|D, \gamma)$ needs to be maximized

with respect to $\mathbf{w}, \beta$. Note, that the variational approximation only depends on the likelihood term and the prior on γ, since these are the only terms that depend on $\mathbf{s}$. Thus, for given $\mathbf{w}$, the variational approximation does not depend on the particular choices for the prior $p(\mathbf{w}, \beta)$. For concreteness, we assume a flat prior $p(\mathbf{w}, \beta) \propto 1$. We set the derivatives of F with respect $\mathbf{m}, \mathbf{w}, \beta$ equal to zero. This gives the following set of fixed point equations:

$$m_i = \sigma\left(\gamma + \frac{\beta p}{2} w_i^2 \chi_{ii}\right) \tag{8}$$

$$\mathbf{w} = (\chi')^{-1}\mathbf{b} \qquad \chi'_{ij} = \chi_{ij} m_j + (1 - m_j)\chi_{jj}\delta_{ij} \tag{9}$$

$$\frac{1}{\beta} = \sigma_y^2 - \sum_{i=1}^{n} m_i w_i b_i \tag{10}$$

with $\sigma(x) = (1 + \exp(-x))^{-1}$ and where in Eq. 10 we have used Eq. 9. Eqs. 8-10 provide the final solution. They can be solved by fixed point iteration as outlined in Algorithm 1: Initialize $\mathbf{m}$ at random. Compute $\mathbf{w}$ by solving the linear system Eq. 9 and β from Eq. 10. Compute a new solution for $\mathbf{m}$ from Eq.8.

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

$$y = \sum_i m_i w_i x_i + \xi \tag{11}$$

with $\langle \xi^2 \rangle = 1/\beta$ and $\mathbf{m}, \mathbf{w}, \beta$ as estimated by the above procedure. Eq. 11 has some similarity with Breiman's non-negative Garrote method [15]. It computes the solution in a two step approach: it computes first w_i using OLS and then finds m_i by minimizing

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

Because of this similarity, we refer to our method as the variational Garrote (VG). Note, that because of the OLS step the non-negative garrote requires that $p \geq n$. Instead, the variational solution Eqs. 8-10 computes the entire solution in one step (and as we will see does not require $p \geq n$).

Let us pause to make some observations about this solution. One might naively expect that the variational approximation would simply consist of replacing $w_i s_i$ in Eq. 2 by its variational expectation $w_i m_i$. If this were the case, $\mathbf{m}$ would disappear entirely from the equations and one would expect in Eq. 9 the OLS solution with the normal input covariance matrix χ instead of the new matrix χ' (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, $\mathbf{m}$ and $\mathbf{w}$ are both to be optimized, giving in general a different solution than the OLS solution [3].

[3] 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. 7 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 $m_i < 1$, χ' differs from χ by rescaling with m_i and adding a positive diagonal to it, a 'variational ridge'. 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 $\mathbf{m}$. Thus, the sparsity prior together with variational approximation provides a mechanism that solves the rank problem. When all $m_i < 1$, χ' is of maximal rank. Each m_i that approaches 1, reduces the rank by one. Thus, if χ has rank $p < n$, χ' 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 χ') is controlled by γ through Eq. 8.

In the above procedure, we compute the VG solution for fixed γ and choose its optimal value through cross validation on independent data [16]. This has the advantage that our result is independent of our (possibly incorrect) prior belief.

But another important advantage of varying γ manually is that it helps to avoid local minima. When we increase γ from a negative value $\gamma_{\min}$ to a maximal value $\gamma_{\max}$ in small steps, we obtain a sequence of solutions with decreasing sparseness. These solutions will better fit the data and as a result β increases with γ. Thus, increasing γ implements an annealing mechanism 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 (see section 3) we recompute the solution from $\gamma_{\max}$ down to $\gamma_{\min}$ and choose the solution with lowest free energy.

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

$$\gamma_{\min} = -\frac{p b_i^2 \chi_{ii}}{2\sigma_y^2} + \sigma^{-1}(\epsilon) + \mathcal{O}(\epsilon) \tag{12}$$

with $\sigma^{-1}(x) = \log(x/(1-x))$. We heuristically set the maximal value of γ as well as the step size.

In appendix B we provide an alternative fixed point iteration scheme that is more efficient in the large n small p limit. Whereas Eqs. 8-10 require the repeated solution of a n-dimensional linear system, the dual formulation, Eqs. (8),(21),(24)-(27), requires the repeated solution of a p dimensional linear system. Algorithm 1 summarizes the VG method.

3 Orthogonal and uni-variate case

In order to obtain further insight in the solution, consider the case in which the inputs are uncorrelated: $\chi_{ij} = \delta_{ij}$. In this case, we can derive the MAP solution of Eq. 5 exactly, without the need to resort to the variational approximation. Eq. 5 reduces to a distribution that factorizes over i with log probability

input : Data $D : \{\mathbf{x}^\mu, y^\mu\}, \mu = 1, \ldots, p$; ϵ and step-size $\Delta\gamma$
output : $\mathbf{w}, \mathbf{m}, \beta, \gamma$ solution with minimal cross validation error

Preprocess data such that $\sum_\mu x_i^\mu = \sum_\mu y^\mu = 0$ and partition D in D^{train}, D^{val}
Compute $b_i = \frac{1}{p}\sum_\mu x_i^\mu y^\mu$ and if $n < p$ compute $\chi_{ij} = \frac{1}{p}\sum_\mu x_i^\mu x_j^\mu$
Compute $\gamma_{\min}$ from ϵ and $\gamma_{\max}$ from $\gamma_{\min}$ and $\Delta\gamma$
for $\gamma = \gamma_{\min} : \Delta\gamma : \gamma_{\max}$ **do** // FORWARD PASS
  $\eta \leftarrow 1$
  while *not converged* **do**
    Compute $\mathbf{w}, \beta$ from Eqs. (9)-(10) ($n < p$) or Eqs. (21), (24)-(27) ($n > p$);
    Compute $\mathbf{m}'$ using a smoothed version of Eq. (8): $m_i' \leftarrow (1-\eta)m_i + \eta\sigma(\ldots)$
    if $\max_i |m_i' - m_i| > 0.1$ **then**
      $\eta \leftarrow \eta/2$
    $\mathbf{m} \leftarrow \mathbf{m}'$
  Store solution $(\mathbf{w}_1, \mathbf{m}_1, \beta_1)_\gamma$ and $F_1(\gamma) \leftarrow F((\mathbf{w}_1, \mathbf{m}_1, \beta_1)_\gamma)$ from Eq. (7)
for $\gamma = \gamma_{\max} : -\Delta\gamma : \gamma_{\min}$ **do** // BACKWARD PASS
  As **5 − 11**
  Store solution $(\mathbf{w}_2, \mathbf{m}_2, \beta_2)_\gamma$ and $F_2(\gamma) \leftarrow F((\mathbf{w}_2, \mathbf{m}_2, \beta_2)_\gamma)$ from Eq. (7)
for $\gamma = \gamma_{\min} : \Delta\gamma : \gamma_{\max}$ **do**
  Choose solution $(\mathbf{w}, \mathbf{m}, \beta)_\gamma$ that has minimal $F_{1,2}(\gamma)$
  Compute cross validation error on D^{val} using Eq. (11)
Select $\mathbf{w}, \mathbf{m}, \beta, \gamma$ with minimal cross validation error

Algorithm 1: The Variational Garrote algorithm.

proportional to

$$L = \frac{p}{2}\log\beta - \frac{\beta p}{2}\left(\sum_{i=1}^n s_i(w_i^2 - 2w_i b_i) + \sigma_y^2\right) + \gamma\sum_{i=1}^n s_i$$

Maximizing wrt w_i, β yields $w_i = b_i$, $\beta^{-1} = \sigma_y^2 - \sum_{i=1}^n s_i b_i^2$ and

$$L = \frac{p}{2}\log\beta + \sum_{i=1}^n s_i\left(\frac{\beta p}{2}b_i^2 + \gamma\right) - \frac{\beta p}{2}\sigma_y^2$$

Assume without loss of generality that b_i^2 are sorted in decreasing order. L is maximized by setting $s_i = 1$ when $\frac{\beta p}{2}b_i^2 + \gamma > 0$ and $s_i = 0$ otherwise. Thus, the optimal solution is $s_{1:k} = 1, s_{k+1:n} = 0$, $\beta^{-1} = \sigma_y^2 - \sum_{i=1}^k b_i^2$ with k the smallest integer such that

$$\frac{\beta p}{2}b_{k+1}^2 + \gamma < 0 \tag{13}$$

By varying γ from small to large, we find a sequence of solutions with decreasing sparsity.

In the variational approximation the solution is very similar but not identical. Eq. 9 gives the same solution $w_i = b_i$. Eqs. 8 and 10 become

$$m_i = \sigma\left(\gamma + \frac{\beta p}{2}b_i^2\right)$$
$$1/\beta = \sigma_y^2 - \sum_i b_i^2 m_i$$

which we can interpret as the variational approximations of Eq. 13, with $m_{1:k} \approx 1$ and $m_{k+1:n} \approx 0$. 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 $1/\beta$.

Note that the posterior is factorized in s_i, the variational approximation is not identical to the exact map solution Eq. 13, although the results are very similar. The relation is $s_i = 0 \Leftrightarrow m_i < 0.5$ and $s_i = 1 \Leftrightarrow m_i > 0.5$.

In order to further analyze the variational solution, we consider the 1-dimensional case. The variational equations become

$$m = \sigma\left(\gamma + \frac{p}{2}\frac{\rho}{1-\rho m}\right) = f(m) \tag{14}$$

$$\frac{1}{\beta} = \sigma_y^2(1 - m\rho) \tag{15}$$

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

In Eq. 14, we have eliminated β and we must find a solution for m for this non-linear equation. We see that it depends on the input-output correlation ρ, the number of samples p and the sparsity γ. For $p = 100$, the solution for different ρ, γ is illustrated in figure 10 (see appendix A). Eq. 14 has one or three solutions for m, depending on the values of γ, ρ, p. The three solutions correspond to two local minima and one local maximum of the free energy F. For $\gamma = -40$ and $\gamma = -10$, we plot the stable solution(s) for different values of ρ in the inserts in fig. 1. The best variational solution for m is given by the solution with the lowest free energy, indicated by the solid lines in the inserts in fig. 1.

Fig. 1 further shows the phase plot of γ, ρ that indicates that the variational solution is unique for $\gamma > \gamma^*$ or for $\rho < \rho^*$. The solid line for $0 < \rho < \rho^*$ in fig. 1 indicates a smooth (second order) phase transition from $m = 0$ to $m = 1$. For $\rho > \rho^*$, the transition from $m = 0$ to $m = 1$ is discontinuous: for each ρ there is a range of values of γ where two variational solutions $m \approx 0$ and $m \approx 1$ co-exist. For comparison, we also show the line $\gamma = -p\rho/2$ that separates the solution $s = 0$ and $s = 1$ according the the exact (non-variational) solution Eq. 13.

The multi-valued variational solution results in a hysteresis effect. When the solution is computed for increasing γ, the $m \approx 0$ solution is obtained until it no longer exists. If the sequence of solutions is computed for decreasing γ the $m \approx 1$ solution is obtained for values of γ where previously the $m \approx 0$ solution was obtained.

From this simple one-dimensional case we may infer that the variational approximation is relatively easy to compute in the uni-modal region (small ρ or γ not too negative) and becomes more inaccurate in the region where multiple minima exist (region between the dot-dashed and dashed lines in fig. 1) .

It is interesting to compare the uni-variate solution of the variational garrote with ridge regression, Lasso or Breiman's Garrote, which was previously done for the latter three methods in [2]. Suppose that data are generated from

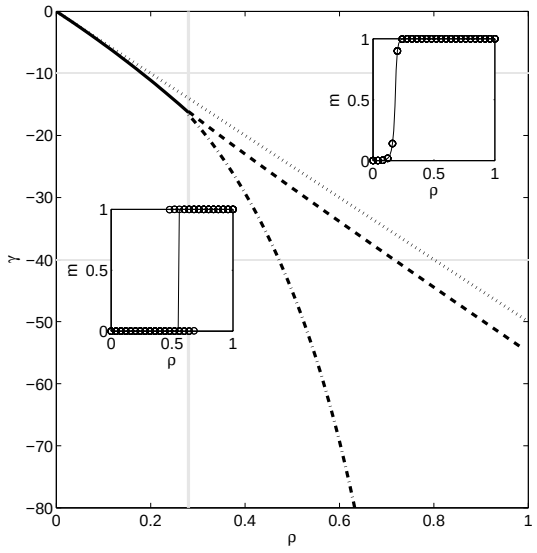

Fig. 1 Phase plot ρ, γ for $p = 100$ giving the different solutions for m. Dashed and dot-dashed lines for $\rho > \rho^* = 0.28$ are from Eq. 18 where two solutions for m exist. Solid line for $\rho < \rho^*$ is the solution for γ when $m = 1/2$, to indicate the transition from the unique solution $m \approx 0$ to the unique solution $m \approx 1$. Dotted line is the exact transition from $s = 0$ to $s = 1$ from Eq. 13. Insets indicate solutions for m versus ρ for $\gamma = -10, p = 100$ (top-right) and for $\gamma = -40, p = 100$ (bottom-left). In the lower left corner of the insets, the unique solution $m \approx 0$ is found. In the top right corner, the unique solution $m \approx 1$ is found. Between the dot-dashed and the dashed line, the two variational solutions $m \approx 0$ and $m \approx 1$ co-exist.

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 $w_{\rm ols} \approx \langle xy \rangle = w$, 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 $w_{\rm ridge} \approx \lambda w$, with $0 < \lambda < 1$ depending on the ridge prior. The Lasso solution (for non-negative w) is given by $w_{\rm lasso} = (w - \gamma)^+$ [2], with γ depending on the L_1 constraint. Breiman's Garrote solution is given by $w_{\rm garrote} = (1 - \frac{\gamma}{w^2})^+ w$ [2], with γ depending on the L_1 constraint. The VG solution is given by $w_{\rm vg} = mw$, with m the solution of Eq. 14. Note, that the VG solution depends, in addition to w, γ, on the unexplained variance σ_y^2 and the number of samples p, whereas the other methods do not.

The qualitative difference of the solutions is shown in fig. 2. 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.

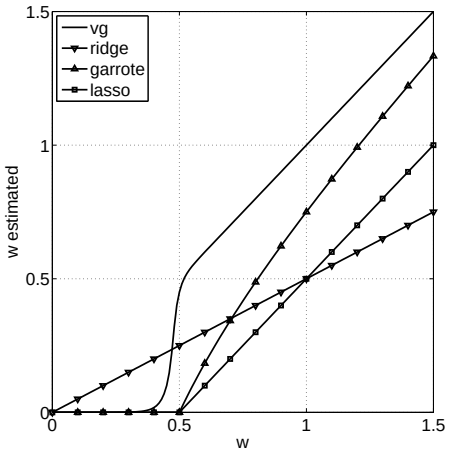

Fig. 2 Uni-variate solution for different regression methods. All methods yield a shrinked solution (deviation from diagonal line). Variational Garrote (VG) with $\gamma = -10, p = 100$ and $\sigma_y^2 = 1$. Ridge regression with $\lambda = 0.5$. Garrote with $\gamma = 1/4$. Lasso with $\gamma = 1/2$.

The qualitative nature of the phase plot fig. 1 and the input-output behavior fig. 2 extends to the multi-variate orthogonal case. The symmetry breaking of feature i is independent of all other features, except for the term $\delta = \sum_{j \neq i} b_j^2 m_j$ that enters through β. If we increase γ, δ increases in steps each time that one of the features j switches from $m_j \approx 0$ to $m_j \approx 1$. Thus δ is constant almost always, except at the step points. Since the critical values of ρ and γ depend in a simple way on δ, the phase plot for the multivariate orthogonal case is qualitatively the same as for the uni-variate case.

4 Numerical examples

In the following examples, we compare the VG with Lasso, ridge regression and in some cases, with the paired mean field approach (PMF) [1].

For most of the examples, we generate a training set, a validation set and a test set. Inputs are generated from a zero mean 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 $d\xi^\mu \in \mathcal{N}(0, \hat{\sigma})$ and $\hat{w}_i$ depending on the problem.

For VG, ridge regression and Lasso, we optimize the model parameters on the training set and, when necessary, optimize the hyper parameters (γ in the case of VG, λ in the case of ridge regression and Lasso) that minimize the quadratic error on the validation set. For the Lasso, we used the method described in [3] [4].

[4] `http://www-stat.stanford.edu/~tibs/glmnet-matlab/`.

Comparison with PMF is performed using the software available online for the regression case with one-dimensional output [5]. Since PMF optimizes hyperparameters as well, we merge both training and validation sets and the resulting dataset is used as input for the PMF method. This ensures that all methods use the same data for parameter estimation.

We define the solution vector for a given method as $\mathbf{v}$. For VG, the components are $v_i \equiv m_i w_i$. In the case of PMF, m_i corresponds to the spike-and-slab variational posterior and w_i to the variational mean for the weights [6]. For Ridge and Lasso $v_i \equiv w_i$.

4.1 Small Example 1

In the first example, we take independent inputs $x_i^\mu \in \mathcal{N}(0,1)$ and a teacher weight vector with only one non-zero entry: $\hat{w} = (1, 0, \ldots, 0)$, $n = 100$ and $\hat{\sigma} = 1$. 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. 12, $\gamma_{\max} = 0.02\gamma_{\min}, \Delta\gamma = -0.02\gamma_{\min}$ (see Algorithm 1 for details).

Results for a single run of the VG are shown in fig. 3. In fig. 3a, we plot the minimal variational free energy F versus γ for both the forward and backward run. Note, the hysteresis effect due to the local minima. For each γ, we use the solution with the lowest F. In fig. 3b, we plot the training error and validation error versus γ. The optimal $\gamma \approx -21$ is denoted by a star and the corresponding $\sigma = 1/\sqrt{\beta} = 1.05$. In fig. 3c, we plot the non-zero component $v_1 = m_1 w_1$ and the maximum absolute value of the remaining components versus γ. Note the robustness of the VG solution in the sense of the large range of γ values for which the correct solution is found. In fig. 3d, we plot the optimal solution $v_i = m_i w_i$ 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 λ minimizes the validation error (star). In fig. 4b,c we see that the Lasso selects a number of incorrect features as well. Fig. 4b also shows that the Lasso solution with a larger λ in the range $0.45 < \lambda < 0.95$ could select the single correct feature, but would then estimate $\hat{w}_1$ 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. Note from fig. 4e, that ridge regression yields a non-sparse solution for all values of λ.

Table 1 shows 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 this simple example, there is no significant difference in the prediction error of Lasso, PMF and VG, but the Lasso solution is significantly less sparse. There is no significant difference between the solutions found by PMF and VG.

[5] `http://www.well.ox.ac.uk/~mtitsias/software.html`.

[6] The notation in [1] uses $\tilde{w}_i$ for w_i and γ_i for m_i.

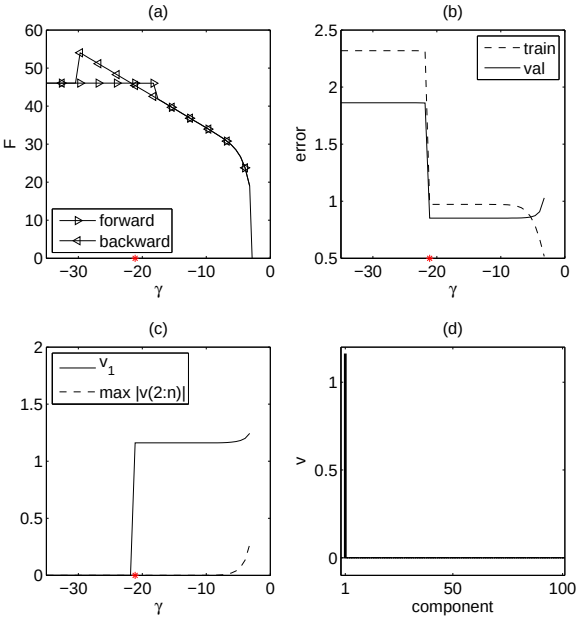

Fig. 3 Top left (a): Minimal variational free energy versus γ. The two curves correspond to warm start solution from small to large γ ('forward') and from large to small γ ('backward') (see also Algorithm 1). Top right (b): Training and validation error versus γ. The optimal γ minimizes the validation error. Bottom left (c): Solution $v_1 = m_1 w_1$ and $\max_{i=2:n} |m_i w_i|$. The correct solution is found in the range $\gamma \approx -20$ to $\gamma \approx -5$. Bottom right (d): Optimal solution $v_i = w_i m_i$ versus i.

	Train	Val	Test	# non-zero	$\|\delta\mathbf{v}\|_1$
Ridge	0.60 ± 0.43	1.72 ± 0.39	1.80 ± 0.12	–	3.97 ± 1.23
Lasso	0.78 ± 0.26	1.07 ± 0.20	1.17 ± 0.20	8.65 ± 6.75	0.80 ± 0.57
PMF	–	–	1.02 ± 0.10	1.5 ± 1.19	0.33 ± 0.37
VG	0.85 ± 0.22	0.96 ± 0.17	1.01 ± 0.10	1.20 ± 0.52	0.31 ± 0.30
True	0.93 ± 0.14	0.87 ± 0.20	0.98 ± 0.04	1	0

Table 1 Results for Example 1 averaged over 20 instances. Train is mean squared error (MSE) on the training set. Val is MSE on the validation set. Test is MSE on the test set. # non-zero is the number of non-zero elements in the Lasso solution and $\sum_{i=1}^{n}(m_i > 0.5)$ for VG and PMF. $\|\delta\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i - \hat{w}_i|$.

4.2 Small Example 2

In the second example, 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 $\chi_{ij} = \zeta^{|i-j|}$, with $\zeta = 0.5$. 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 $n = 100, \hat{\sigma} = 1$ and $p/p_v/p_t = 50/50/400$. In table 2 we compare the performance of the VG, Lasso, ridge regression and PMF on

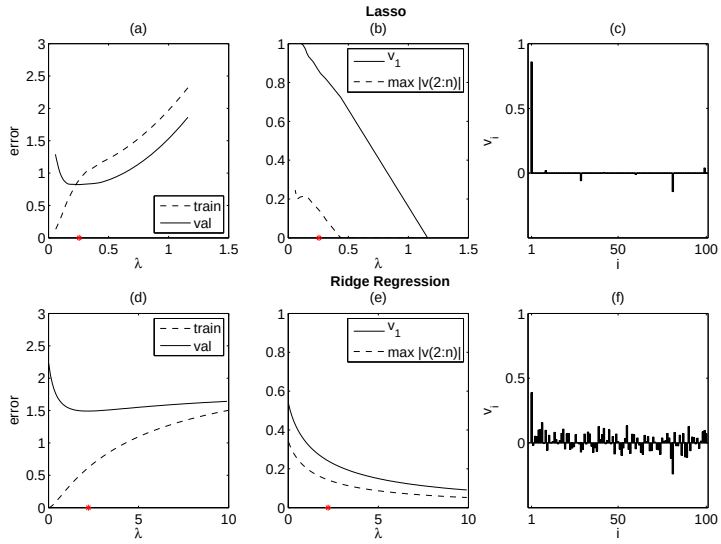

Fig. 4 Regression solution for Lasso and ridge regression for same data set as in fig. 3. Top row (a,b,c): Lasso. Bottom row (d,e,f): Ridge regression. Left column (a,d): training and validation errors versus λ. Middle column (b,e): Solution for the non-zero feature v_1 and the zero-features $\max_{i=2:n} |v_i|$. Right column (c,f): Optimal Lasso and ridge regression solution v_i versus i.

	Train	Val	Test	# non-zero	$\|\delta \mathbf{v}\|_1$
Ridge	0.32 ± 0.27	3.30 ± 0.67	3.46 ± 0.31	–	11.09 ± 0.93
Lasso	0.75 ± 0.37	1.39 ± 0.37	1.48 ± 0.29	16.30 ± 6.60	2.08 ± 0.87
PMF	–	–	1.06 ± 0.11	5.15 ± 0.49	0.67 ± 0.35
VG	0.80 ± 0.25	1.13 ± 0.31	1.15 ± 0.21	5.05 ± 0.51	0.83 ± 0.54
True	0.93 ± 0.14	0.87 ± 0.20	0.98 ± 0.04	5	0

Table 2 Results for Example 2. For definitions see caption of Table 1 above.

20 random instances. We see that the VG and PMF significantly outperform the Lasso method and ridge regression both in terms of prediction error and accuracy of the estimation of the parameters. Again, there is no significant difference between PMF and VG.

4.3 Effect of the noise

In this subsection we show the accuracy VG, Lasso and PMF as a function of the noise $\hat{\sigma}^2$. We generate data with $n = 100, p = 100, p_v = 20$ and $\hat{w}_i = 1$ for 20 randomly chosen components i. We vary $\hat{\sigma}^2$ in the range 10^{-4} to 10 for for two values of the correlation strength in the inputs $\zeta = 0.5, 0.95$.

For weakly correlated inputs, Fig. 5a., we distinguish three noise domains: for large noise all methods produce errors of $\mathcal{O}(1)$ and fail to find the predictive

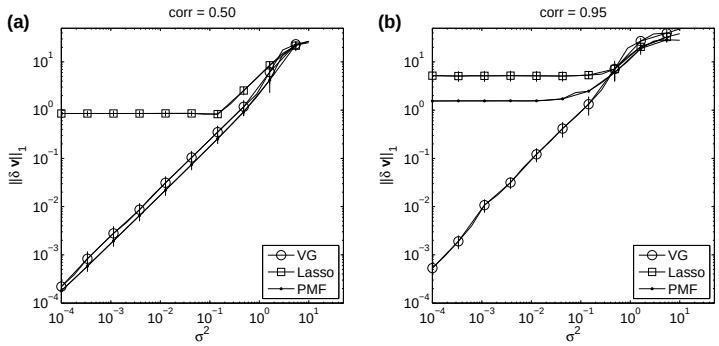

Fig. 5 Accuracy of VG, Lasso and PMF as a function of the noise. Errorbars of $\| \delta \mathbf{v} \|_1$ for 10 different random instances. Data is generated using $n = 100, p = 100, p_v = 20$ and $\hat{w}_i = 1$ for 20 randomly chosen components i. We consider two values of the correlation strength in the inputs: **(a)** weakly correlated inputs $\zeta = 0.5$ and **(b)** strongly correlated inputs $\zeta = 0.95$. For PMF we choose the best solution (the one with highest value of the bound) for 10 different random initializations for each of the 10 instances.

features. For intermediate and low noise levels, VG and PMF are significantly better than Lasso. In the limit of zero noise, the error of VG and PMF keeps on decreasing whereas the Lasso error saturates to a constant value.

For strongly correlated inputs, Fig. 5b., we observe that whereas the error of VG scales approximately as before, PMF gets stuck in local minima in some instances, yielding worse average performance than VG. See section 5 for a further discussion of this point.

4.4 Analysis of consistency: VG vs Lasso

It is well-known that the Lasso method may yield inconsistent results when input variables are correlated. In [17], necessary and sufficient conditions for consistency are derived. 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, ξ, e are independent and Normal distributed random variables, $x_3 = 2/3x_1 + 2/3x_2 + \xi$ and $y = \sum_{i=1}^{3} \hat{w}_i x_i + e$, $p = 1000$. When $\hat{w} = (-2, 3, 0)$ (Example b) this example is consistent, but when $\hat{w} = (2, 3, 0)$ (Example a) this example violates the consistency condition. The Lasso and VG solution for Example a for different values of λ and γ are shown in fig. 6a,b, respectively. The VG solution $v_i = m_i w_i$ in terms of m_i and w_i is shown in fig. 6c,d. The average results over 100 instances for Example a and Example b are shown in table 3. We see that the VG does not suffer from inconsistency and always finds the correct solution. This is remarkable as one might have feared that the non-convexity of the VG would result in sub-optimal local minima.

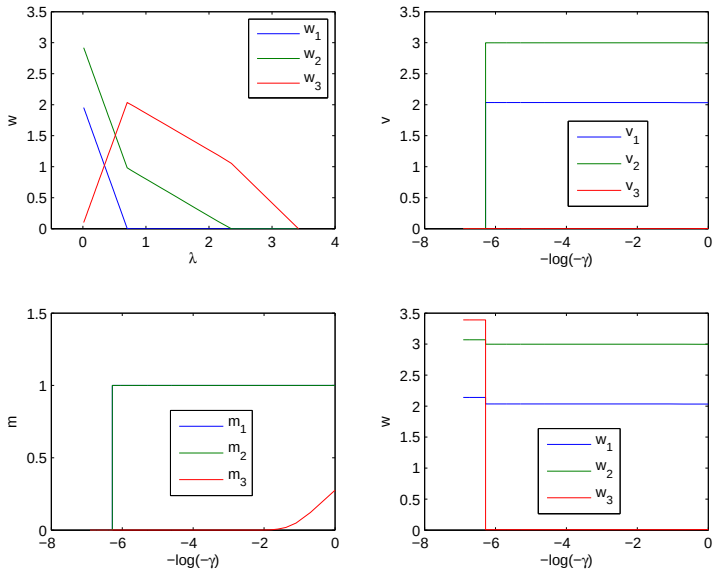

Fig. 6 (Color online) Lasso and VG solution for the inconsistent Example a of [17]. Top left: Lasso solution versus λ is called inconsistent because it does not contain a λ for which the correct sparsity ($w_{1,2} \neq 0, w_3 = 0$) is obtained. Top right: the VG solution for $\mathbf{v}$ versus γ contains large range of γ for which the correct solution is obtained. Bottom left: VG solution for $\mathbf{m}$ (curves for $m_{1,2}$ are identical). Bottom right: VG solution for $\mathbf{w}$.

	Example a		Example b	
	$\|\delta\mathbf{v}\|_1$	$\max(\|v_3\|)$	$\|\delta\mathbf{v}\|_1$	$\max(\|v_3\|)$
Ridge	0.64 ± 0.18	0.48	0.02 ± 0.02	0.27
Lasso	0.19 ± 0.14	0.30	0.00 ± 0.00	0.00
VG	0.05 ± 0.03	0.00	0.00 ± 0.00	0.00

Table 3 Accuracy of Ridge, Lasso and VG for Example 1a,b from [17]. $p = p_v = 1000$. Parameters λ (Ridge and Lasso) and γ (VG) optimized through cross validation. $\|\delta\mathbf{v}\|_1$ as before, $\max(|v_3|)$ is maximum over 100 trials of the absolute value of v_3. Example a is inconsistent for Lasso and yields much larger errors than the VG. Example b is consistent and the quality of the Lasso and VG are similar. Ridge regression is bad for both examples.

4.5 Boston-housing dataset: VG vs PMF

We now focus on comparing in more detail the performance of VG with PMF. In [1], the Boston-housing dataset[7] is used to test the accuracy of the PMF approximation.

This is a linear regression problem that consists of 456 training examples with one-dimensional response variable y and 13 predictors that include housing values. We use here the same setup as in [1] to compare VG with PMF. For PMF, hyperparameters were fixed to values $\sigma = 0.1 \times \mathrm{var}(y), \pi = 0.25, \sigma_w = 1$

[7] `http://archive.ics.uci.edu/ml/datasets/Housing`

	soft-error	*extreme*-error
PMF [1]	0.208 [0.002, 0.454]	0.204 [0.002, 0.454]
PMF	0.237 [0.001, 0.454]	0.209 [0.001, 0.454]
VG	0.006 [0.006, 0.006]	0.006 [0.006, 0.006]

Table 4 Comparison of VG and PMF in the Boston-housing dataset in terms of approximating the ground-truth $\hat{w}$. Average errors $\|\delta \mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i - \hat{w}_i|$, with v_i the approximation of VG or PMF, together with 95% confidence intervals (given by percentiles) obtained after 300 random initializations for both soft and extreme initializations.

where var(y) denotes the output variance. For the VG, we use $\beta = 1/\sigma^2$, $\gamma = \log(\pi/(1-\pi))$ and hyperparameter σ_w is implicitly equal to ∞ in the VG (see Appendix C for details of how both models compare). Since γ and β are given, the VG algorithm reduces to iterate eqs. (8) and (9) starting from a random $\mathbf{m}$. Similarly, the PMF reduces to perform an E-step given the fixed hyperparameter values.

As in [1], we use random initial values for the variational parameters *between* 0 and 1 (*soft* initialization) and random values *equal to* 0 or 1 (*hard* initialization). We considered as ground truth $\hat{w} \equiv \mathbf{w}^{\mathrm{tr}}$ the result of the efficient paired Gibbs sampler developed in [1].

Table 4 shows the results. The first and second rows show the errors reported in [1] and the errors that we obtain using their software, respectively. We observe a small discrepancy in the average errors. However, if we consider the percentiles, the results are consistent. In practice, what we observe is that PMF finds two local optima depending on the initialization: one is the correct solution (error $\approx 10^{-3}$) whereas the other has error 0.454. These two solutions are found equally often for both soft or hard initializations, showing no dependence on the type of initialization, in agreement with [1].

The results of VG are shown on the third row. Contrary to PMF, the VG shows no dependence on the initialization and always finds a solution with an error of order 10^{-3}. These results give evidence that the combined variational/MAP approach of VG can be better than PMF avoiding local minima.

4.6 Dependence on the Number of Samples

We now analyze the performance of all considered methods as a function of the proportion of samples available. We first analyze the case when inputs are not correlated and then consider correlations of practical relevance that appear in genetic datasets.

For these experiments, we generate the data for dimension $n = 500$ and noise level $\beta = 1$. We explore two scenarios: very sparse problems with only 10% of active predictors and denser problems with 25% of active predictors. The weights of the active predictors are set to 1.

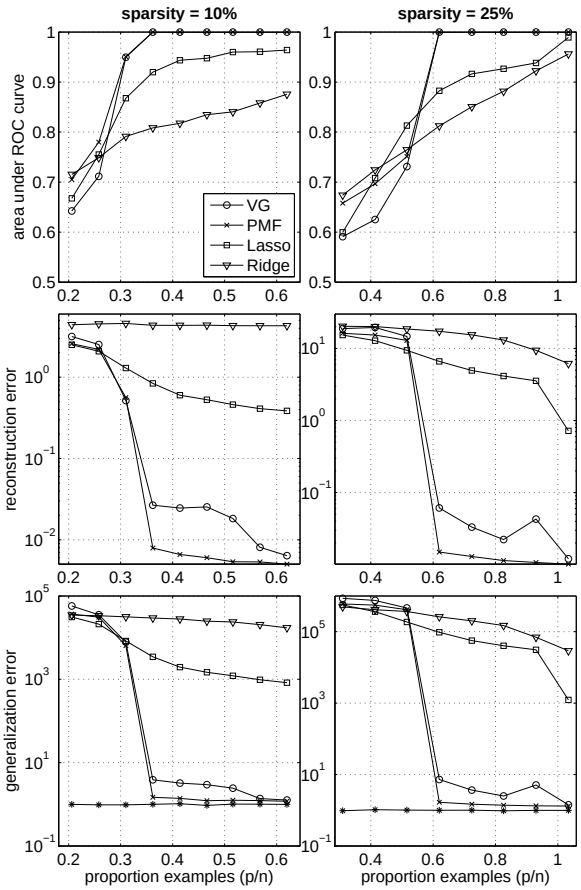

Fig. 7 Uncorrelated case: Performance as a function of number of training samples p for two levels of sparsity (10% and 25% of non-zero entries). For each value averages over 20 runs are plotted. **Top:** area under the ROC curves (see text for definition). **Middle:** reconstruction error, defined as $\|\delta \mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i - \hat{w}_i|$. **Bottom:** generalization error, defined as the MSE in the test set. For all methods except for PMF, train set size is p and validation sets size $p_v = p/30$. For PMF the training set has size $p + p_v$. Lowest curve shows theoretically optimal generalization error obtained by using the target weights from which the data is generated.

4.6.1 Uncorrelated Case

Figure 7 shows results of performance for uncorrelated inputs. Top plots show the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is calculated by thresholding the weight estimates. Those weights that lie above (below) the threshold are considered as active (inactive) predictors. The ROC curve plots the fraction of true positives versus the fraction of false positives for all threshold values. The area under the curve measures the ability of the method to correctly classify those predictors that are and are not

active. A value of 1 for the area represents a perfect classification whereas 0.5 represents random classification. The ROC is plotted as a function of the fraction of samples relative to the number of inputs: p/n.

For both VG and PMF, we observe in all performance measures a transition from a regime where solutions are poor to a regime with almost perfect recovery. This transition, not noticeable in the other (convex) methods, occurs at around 35% of examples for 10% of sparsity (left column) and shifts to higher values for denser problems ($\approx 60\%$ for 25% of sparsity, right column).

If we compare VG with PMF we see that in the regime where both methods perform well, PMF performs slightly better than VG in terms of reconstruction error but their performance is identical in terms of the area under the ROC curve. The difference between VG and PMF is slightly more pronounced for denser problems. We also see that Lasso performs better than Ridge regression, but the difference between both methods tends to be smaller for denser problems. Both Lasso and ridge regression are significantly worse than VG and PMF.

4.6.2 Correlated case: Genetic dataset

We now consider correlated inputs. We use input data obtained from a genetic domain, where inputs x_i denote single nucleotide polymorphisms (SNPs) that have values $x_i = \{0, 1, 2\}$. SNPs typically show correlations structured in blocks, where nearby SNPs are highly correlated, but show no dependence on distant SNPs. An example of such correlation matrix can be seen in Figure 8 (left). The output data are generated as above.

Figure 8 (right) shows the results. Contrary to the uncorrelated case, the existence of strong correlations between some of the predictors prevents a clear distinction between solution regimes as a function of training set size. We observe, as before, that both VG and PMF are the preferable methods for sufficiently large training set size. In the three performance measures considered, VG performs better or comparable to PMF. Interestingly, the difference between VG and PMF becomes more significant for denser problems, when we expect more difficulty due to more presence of local minima.

4.7 Scaling with dimension n

We conclude our empirical study by analyzing how the methods scale, both in terms of the quality of the solution as in terms of CPU times, as a function of the number of features n for a constant number of samples. We use the data as in Example 2 above, with uncorrelated inputs.

Figure 9 shows the results for VG, PMF and Lasso. For the VG, we use the dual method described in the appendix B. Fig. 9a shows that the VG and PMF have constant quality in terms of the error $\|\delta \mathbf{v}\|_1$, whereas the quality of the Lasso deteriorates with n. Fig. 9b shows that the VG and PMF have close to optimal norms $L_0 = 5$ and that the L_0 norm of the Lasso

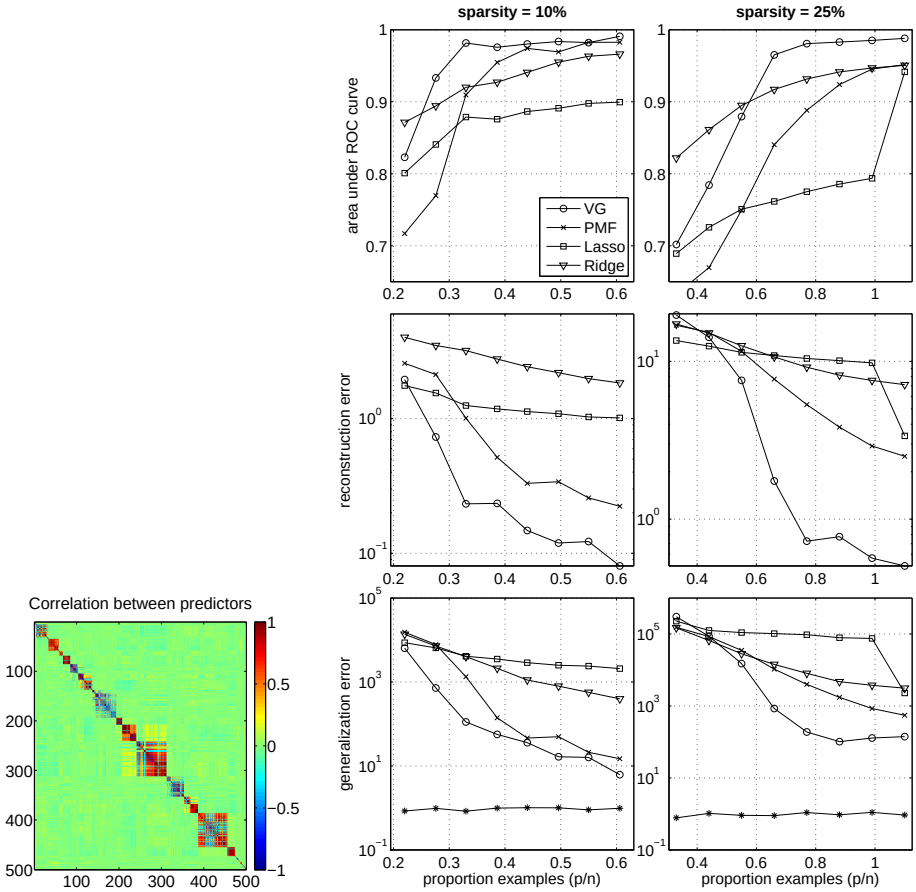

Fig. 8 Correlated case: **(LEFT)** (Color online) Example of input correlation matrix. **(RIGHT)** Performance as a function of number of training samples p for two levels of sparsity (10% and 25% of non-zero entries). For each value averages over 20 runs are plotted. **Top:** area under the ROC curves (see text for definition). **Middle:** reconstruction error, defined as $\|\delta \mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i - \hat{w}_i|$. **Bottom:** generalization error, defined as the MSE in the validation set. For all methods except for PMF, train set size is p and validation sets size $p_v = p/10$. For PMF the training set has size $p + p_v$.

deteriorates with n. Fig. 9c shows that the computation time of all methods scales approximately linear with n. Lasso is significantly faster than VG and PMF, and VG is significantly faster than PMF. Note, however that the VG and the PMF methods are implemented in Matlab whereas the Lasso method uses an optimized Fortran implementation.

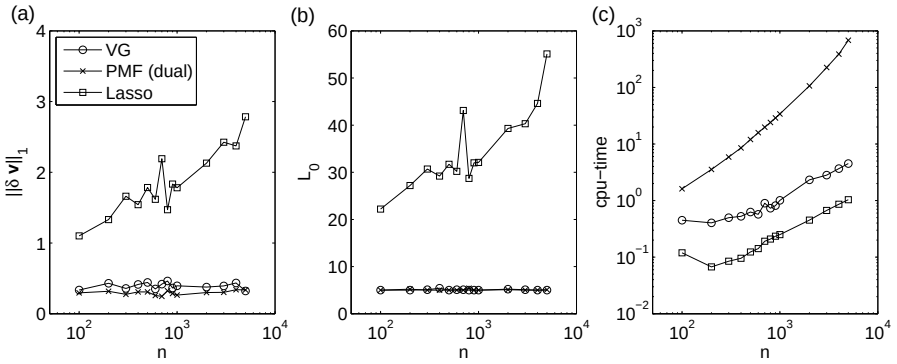

Fig. 9 Scaling with n: performance of VG, PMF and Lasso as a function of the number of features n. Data are generated as in Example 2. $p = 100, p_v = 100, \beta = 2, \zeta = 0$.

5 Discussion

In this paper, we have introduced a new variational method for sparse regression using L_0 penalty. We have presented a minimal version of the model with no (hierarchical) prior distributions to highlight some important features: the variational ridge term that dynamically regularizes the regression; the input-output behavior as a smoothed version of hard feature selection; a phase plot that shows when the variational solution is unique in the orthogonal design case for different p, ρ, γ. We have also shown numerically that the VG is efficient and accurate and yields results that significantly outperform other considered methods.

The VG suffers from local minima as can be expected for any method that needs to solve a non-convex problem, like the PMF. From the numerical experiments we can conclude that VG is on average preferable to PMF in practical scenarios with strongly correlated inputs and/or moderately sparse problems, where the local minima problem is more severe. Although we have no principled solution for the local minima problem, we think that the combined variational/MAP approach together with the the annealing procedure that results from increasing γ, followed by a "heating" phase to detect hysteresis works well in practice, helping to avoid local minima. Another obvious approach is to use multiple restarts or using more powerful approximations, such as structured mean field approximation or belief propagation. We remark that the PMF in the general setting [1] includes an extra layer of flexibility that can be used to capture correlations between input variables. Such extensions can also be considered for VG.

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

We have seen that the performance of the VG is excellent in the zero noise limit. In this limit, the regression problem reduces to a compressed sensing

problem [18,19]. The performance of compressed sensing with L_q sparseness penalty was analyzed theoretically in [20], 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.

Our implementation uses parallel updating of Eqs. 8-10, or Eqs. 8,21, 24-27 when using the dual formulation. One may consider also a sequential updating. This was done successfully for the Lasso based on the idea of the Gauss-Seidel algorithm [3]. The advantage of such an approach is that each update is linear in both n and p, since only the non-zero components need to be updated. However, the number of updates to converge will be larger. The proof of convergence for such a coordinate descend method for the VG is likely to be more complex than for the Lasso due to non-convexity. As a result, a smoothing parameter $\eta \neq 1$ (see Algorithm 1) may still be required.

Acknowledgments

We would like to thank M. Titsias for providing the code of PMF and specially the Boston Housing files. We also thank Kevin Sharp and Wim Wiegenrick for useful discussions and anonymous reviewers for helping on improving the manuscript.

References

1. M. Titsias and M. Lázaro-Gredilla. Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning. In *Advances in Neural Information Processing Systems 24*, pages 2339–2347, 2011.
2. R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.
3. J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010.
4. I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.
5. E. I. George and R. E. McCulloch. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.
6. B. A. Logsdon, G. E. Hoffman, and J. G. Mezey. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11:58, 2010.
7. R. Yoshida and M. West. Bayesian learning in sparse graphical factor models via variational mean-field annealing. *Journal of Machine Learning Research*, 99(Aug):1771–1798, 2010.
8. D. Hernández-Lobato, J. M. Hernández-Lobato, T. Helleputte, and P. Dupont. Expectation propagation for bayesian multi-task feature selection. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part I*, pages 522–537. Springer-Verlag, 2010.
9. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.
10. H. J. Kappen and J. J. Spanjers. Mean field theory for asymmetric neural networks. *Physical Review E*, 61:5658–5663, 2000.
11. K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy Belief Propagation for approximate inference: An empirical study. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, pages 467–475. Morgan Kaufmann Publishers, 1999.

12. M. Opper and D. Saad. *Advanced Mean Field Methods: Theory and Practice*. MIT press, 2001.
13. M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.
14. D. Barber and W. Wiegerinck. Tractable variational structures for approximating graphical models. In *Advances in Neural Information Processing Systems II*, pages 183–189. MIT Press, 1999.
15. L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.
16. T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):pp. 1023–1032, 1988.
17. P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7(Dec):2541–2563, 2006.
18. E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.
19. D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
20. 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*, 2009(09):L09003, 2009.

A : Phase plot computation for the orthogonal case

In the uni-variate case, $f(m)$ in Eq. 14 is an increasing function of m and crosses the line m either 1 or three times, depending on the values of p and γ (see fig. 10). In the multivariate orthogonal case, this is still true, since the influence of other features is only through β. We can thus write $\beta^{-1} = \sigma_y^2(1 - \rho m - \delta)$, where $0 \leq \delta < 1$ is a function of the variational parameters of the other features. Thus, there are regions of parameter space γ, p, ρ where

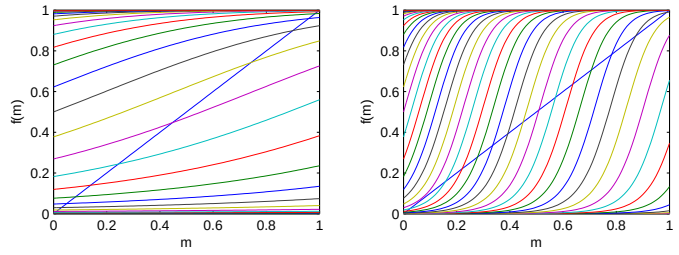

Fig. 10 (Color online) $f(m)$ vs m. Left (a): $p = 100, \gamma = -10$, different lines correspond to different values of $0 < \rho < 1$ (higher lines are higher ρ). The solution for m is given by the intersection f with the diagonal line. The solution for m is unique and increases with increasing ρ. Right (b): Same as left, but with $p = 100, \gamma = -30$. Depending on ρ, there are one or three solutions for m. The solutions close to $m \approx 0, 1$ correspond to local minima of F. The intermediate solution corresponds to a local maximum of F.

the uni-variate solution is unique and others for which there are two stable solutions.

The transition between these two regions is when $f'(m) = 1$ and $f(m) = m$. These two equations imply

$$\left(1 + \frac{p}{2}\right)\rho^2 m^2 - \left(2\rho(1-\delta) + \frac{p}{2}\rho^2\right)m + (1-\delta)^2 = 0 \tag{16}$$

This quadratic equation in m has either zero, one or two solutions, corresponding to no touching, touching once and touching twice, respectively. Denote $a = (1+\frac{p}{2})\rho^2, b = 2\rho(1-\delta)+\frac{p}{2}\rho^2$. The critical value for ρ, p is when Eq. 16 has one solution for m, which occurs when

$$D = b^2 - 4a(1-\delta)^2 = \frac{p}{2}\rho^2\left(\rho-\rho^*\right)\left(\frac{p}{2}\rho + 2(1-\delta) + 2(1-\delta)\sqrt{1+\frac{p}{2}}\right) = 0$$
$$\rho^* = \frac{4}{p}(1-\delta)\left(\sqrt{1+\frac{p}{2}}-1\right) \tag{17}$$

Thus, D is positive when $\rho > \rho^*$ and Eq. 16 has two solutions for m. We denote these solutions by $m_{1,2} = \frac{b\pm\sqrt{D}}{2a}$. Note, that the solutions in these critical points only depend on ρ, p. For each of these solutions we must find a γ such that $f(m) = m$, which is given by

$$\gamma_i = \log\frac{m_i}{1-m_i} - \frac{p}{2}\frac{\rho}{1-\rho m_i} \qquad i = 1,2 \tag{18}$$

It is easy to see that the smallest of these solutions $m_1 < m_2$ corresponds to a local maximum of the free energy and can be discarded. Thus, when $\rho > \rho^*$ and $\gamma_2 < \gamma < \gamma_1$ two stable variational solutions $m \approx 0, 1$ co-exist.

When $\rho < \rho^*$, Eq. 16 has no solutions for m. In this case the conditions $f'(m) = 1$ and $f(m) = m$ cannot be jointly satisfied and the variational solution is unique.

From Eq. 17 we see that ρ^* is a decreasing function of p and when $p \gg 1$, $\rho^* \approx 2\sqrt{\frac{2}{p}}$. In the critical point, where $\rho = \rho^*(p)$, $m = b/2a \approx \frac{1}{2}\left(1+\sqrt{\frac{2}{p}}\right)$ and

$$\gamma^* \approx -\sqrt{2p}(1-\delta) \tag{19}$$

When $\rho < \rho^*$ *or* $\gamma > \gamma^*$ the variational solution is unique. We illustrate the phase plot ρ, γ for $p = 100$ in fig. 1a.

B : 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 $\chi' \mathbf{w} = \mathbf{b}$. 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 λ^μ:

$$\begin{aligned} F = &-\frac{p}{2}\log\frac{\beta}{2\pi} + \frac{\beta}{2}\sum_\mu^p (z^\mu - y^\mu)^2 + \frac{\beta p}{2}\sum_i m_i(1-m_i)w_i^2\chi_{ii} \\ &- \gamma\sum_{i=1}^n m_i + \sum_{i=1}^n \left(m_i\log m_i + (1-m_i)\log(1-m_i)\right) \\ &+ \sum_\mu \lambda^\mu (z^\mu - \sum_i m_i w_i x_i^\mu) \end{aligned} \tag{20}$$

We compute the derivatives of Eq. 20:

$$\begin{aligned}
\frac{\partial F}{\partial w_i} &= m_i \left(\beta p (1-m_i) \chi_{ii} w_i - \sum_\mu \lambda^\mu x_i^\mu \right) \\
\frac{\partial F}{\partial z^\mu} &= \beta (z^\mu - y^\mu) + \lambda^\mu \\
\frac{\partial F}{\partial \beta} &= -\frac{p}{2\beta} + \frac{1}{2} \sum_\mu^p (z^\mu - y^\mu)^2 + \frac{p}{2} \sum_i m_i (1-m_i) w_i^2 \chi_{ii} \\
\frac{\partial F}{\partial m_i} &= \frac{\beta p}{2} (1-2m_i) w_i^2 \chi_{ii} - \gamma + \sigma^{-1}(m_i) - \sum_\mu \lambda_\mu w_i x_i^\mu \\
\frac{\partial F}{\partial \lambda^\mu} &= z^\mu - \sum_i m_i w_i x_i^\mu
\end{aligned}$$

By setting $\frac{\partial F}{\partial w_i} = \frac{\partial F}{\partial z^\mu} = 0$ we obtain

$$w_i = \frac{1}{\beta p \chi_{ii}} \frac{1}{1-m_i} \sum_\mu \lambda^\mu x_i^\mu \tag{21}$$

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

$$\beta = \frac{1}{p} \sum_{\mu\nu} \lambda_\mu \lambda_\nu A_{\mu\nu} \tag{22}$$

$$\beta y^\mu = \sum_\nu A_{\mu\nu} \lambda^\nu \tag{23}$$

with $A_{\mu\nu}$ given by

$$A_{\mu\nu} = \delta_{\mu\nu} + \frac{1}{p} \sum_i \frac{m_i}{1-m_i} \frac{x_i^\mu x_i^\nu}{\chi_{ii}} \tag{24}$$

For given $A_{\mu\nu}$, let $\hat{y}$ denote the solution of

$$\sum_{\nu=1}^p A_{\mu\nu} \hat{y}^\nu = y^\mu \tag{25}$$

Then it is easy to verify that

$$\frac{1}{\beta} = \frac{1}{p} \sum_\mu \hat{y}^\mu y^\mu \tag{26}$$

$$\lambda^\mu = \beta \hat{y}^\mu \tag{27}$$

solve the system of Eqs. 22-23.

C : Relation with the Paired-Mean Field approximation

The VG shares many similarities with the recently proposed paired mean field (PMF) variational approach [1]. Here we relate both approaches in terms of three different aspects: the probabilistic model, the variational approximation and the optimization algorithm.

Model : The model considered for the PMF variational approximation is defined for multiple outputs and considers a linear combination of basis functions governed by a Gaussian process. To relate this model to the one presented in this work, we consider the one-dimensional output without the extra input layer.

The spike and slab model [5] considers a linear regression model of the form:

$$y = \sum_{i=0}^{n} \hat{v}_i x_i + \xi$$
$$\hat{v}_i \sim \pi \mathcal{N}(\hat{v}_i | 0, \sigma_w^2) + (1-\pi)\delta_0(\hat{v}_i), \qquad \forall i.$$

That is, the prior over the weights is factorized, with each weight distributed according to a mixture distribution: with probability π, each $\hat{v}_i$ is drawn from a Gaussian centered at zero with variance σ_w^2, and with probability $1-\pi$, each $\hat{v}_i$ is zero. The sparsity of the solution is controlled by π, either directly or by specifying a prior over π.

Observe that we can equivalently write $\hat{v}_i$ as the product of a Bernoulli random variable $s_i \sim \pi^{s_i}(1-\pi)^{1-s_i}$ and a Gaussian random variable $w_i \sim \mathcal{N}(w_i|0, \sigma_w^2)$, which is the reparameterization used in [1].

To relate the model used in the VG defined by Eqs. (2) and (4) to the previous one we make the following identifications:

- The prior on w_i is flat, which corresponds to setting $\sigma_w = \infty$ in [1].
- $\gamma = \log(\pi/(1-\pi))$.

Thus, the spike and slab model [5] and the model considered by [1] are identical and both models are identical to the model considered in this paper when a Gaussian prior is placed over the weights.

Variational approximation : The PMF variational distribution places each weight w_i and bit s_i in the same factor:

$$q(\mathbf{w}, \mathbf{s}) = \prod_{i=1}^{n} q_i(w_i, s_i). \tag{28}$$

On the contrary, the VG reduces to the classical factorized variational approach under the restriction that the posterior for the weight is a delta function.

Algorithm : The optimization in [1] uses an EM algorithm that alternates between expected values of the latent variables $\mathbf{w}, \mathbf{s}$ (E-Step) and optimization of hyperparameters $\{\sigma_y^2, \sigma_w^2, \pi\}$ (M-Step).

The VG method differs mainly in two points. The VG method:

- Computes expectation of $\mathbf{s}$ (denoted by $\mathbf{m}$) but finds MAP solution for $\mathbf{w}$.
- Searches the space of solutions using a forward and a backward sequential search over hyperparameter γ using a validation set. For a given γ, the rest of the parameters are optimized using a training set and initialized with a 'warm' solution from the previous step (see Algorithm 1).