Efficient sparse regression using ℓ_0-norm regularization

Hilbert J Kappen[1] and Vicenç Gómez[2]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen (The Netherlands),
[2] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona (Spain)

Abstract

- **Sparse linear regression** is widely used in biomedical data analysis. We analyze the variational method for sparse regression using ℓ_0-norm regularization, which we refer to as the Variational Garrote (VG) [1].
- The VG finds correct solutions when the lasso (ℓ_1-norm) solution is inconsistent due to large input correlations.
- The computational cost scales cubic in the number of samples, but close to linear in the number of features.
- We show the performance of VG on input data obtained from a genetic domain, where inputs denote single nucleotide polymorphisms (SNPs).

Introduction

The Linear Regression Problem

- Input data: x_i^μ (n-dimensional), $i = 1, \dots, n$ and $\mu = 1, \dots, p$
- Output data: y^μ (1-dimensional)
- Find weights w_i, w_0 that best describe the relation

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

ξ^μ models uncertainty as zero-mean noise with inverse variance $\beta = 1/\sigma_P^2$.

Simplest solution : Ordinary Least Squares

OLS solution minimizes sum of squares error:

$$\mathbf{w} = \chi^{-1}\mathbf{b}$$
$$w_0 = \bar{y} - \sum_i w_i \bar{x}_i$$

where χ is the input covariance matrix $\chi_{ij} = \frac{1}{p}\sum_\mu x_i^\mu x_j^\mu$
$\mathbf{b}$ is the vector of input-output covariances $\mathbf{b} = \frac{1}{p}\sum_\mu x_i^\mu y^\mu$ and $\bar{x}_i, \bar{y}$ are the mean values.

- **Problem** : if dimension n is very large and the number of samples is very small p
  $$n \gg p$$
  inverse of χ is not well defined!

Penalized linear regression

Solution: Penalize undesirable solutions in the objective function

$$\mathcal{L} = \underbrace{\frac{1}{2}\sum_{\mu=1}^{p}\left(y^\mu - \sum_{i=1}^{n} w_i x_i^\mu\right)^2}_{\text{Sum of Squares}} + \underbrace{\lambda \sum_{i=1}^{n} |w_i|^q}_{\text{Penalty term}}$$

where $\lambda > 0$ determines how much we penalize and $q \geq 0$.

- **Ridge regression**: $q = 2$.
  - Penalizes the ℓ_2-norm of the weight vector $\mathbf{w}$
  - Replaces the input covariance matrix χ with $\chi + \lambda I$, that can be invertible
  - Improves prediction accuracy, but not the interpretability
- **Lasso**: $q = 1$.
  - Penalizes the ℓ_1-norm of the weight vector (sum of the absolute values)
  - Favors sparse solutions by setting certain coefficients to zero and shrinking the rest
  - Preserves the convexity (tractability) of the optimization problem
  - Good compromise between prediction accuracy, interpretability and tractability
- **ℓ_0 norm**: $q = 0$.
  - Penalizes the ℓ_0 norm (number of non-zeros α_i)
  - Improves the selection of relevant variables, resulting in more interpretable solutions.
  - Prevents over-shrinkage of the regression coefficients.
  - For $q < 1$, non-convex optimization problem: more difficult.

The Variational Garrote

Variable Selection: ℓ_0-norm penalty

Introduce additional binary variables $s_i = \{0, 1\}$ that indicate if predictor i is active ($s_i = 1$) or inactive ($s_i = 0$). The regression model becomes:

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

Bayesian Inference: : Probability distribution over parameters ($\boldsymbol{w}, \boldsymbol{s}, \beta$) given the data D

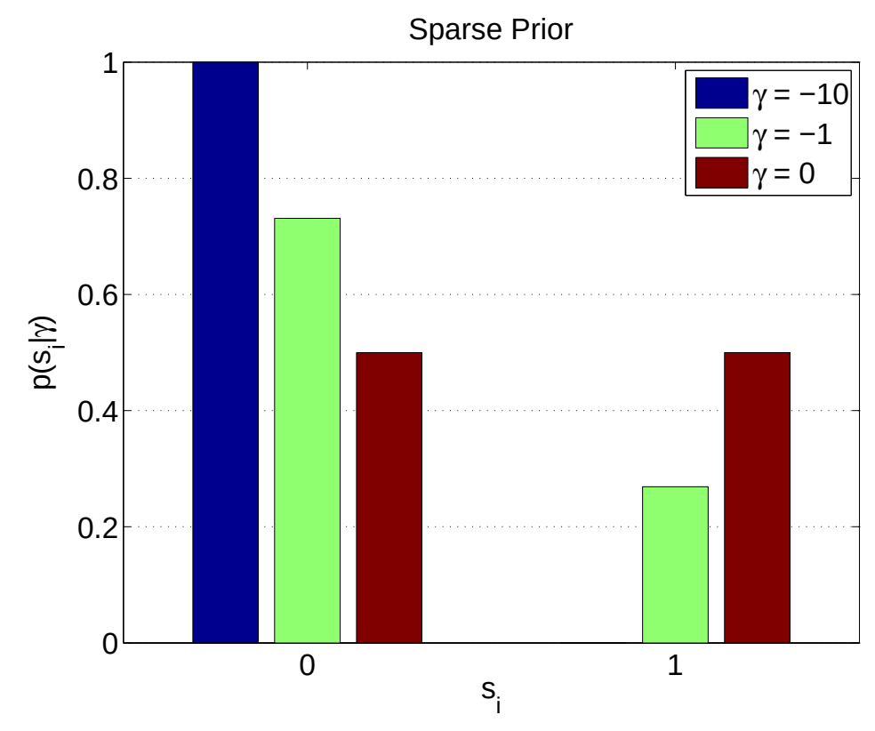
$$p(\mathbf{s}, \mathbf{w}, \beta | D, \gamma) = \frac{p(\mathbf{w}, \beta) p(\mathbf{s}|\gamma) p(D|\mathbf{s}, \mathbf{w}, \beta)}{p(D|\gamma)}$$

Sparse prior distribution for $\boldsymbol{s}$:

$$p(\boldsymbol{s}|\gamma) = \prod_{i=1}^{N} p(s_i|\gamma), \qquad p(s_i|\gamma) = \frac{\exp(\gamma s_i)}{1 + \exp(\gamma)},$$

where γ (similar to λ before) determines the sparsity of the solution:

- $\gamma \ll 0$ favors sparse solutions
- $\gamma \approx 0$ indicates bias towards dense solutions



The Variational Garrote is an approximated method:

1. Performs variational approximation to the marginal posterior $p(\boldsymbol{w}, \beta | D, \gamma)$
2. Computes Maximum-a-Posteriori (MAP) solution with respect to $\boldsymbol{w}, \beta$

Variational (Mean-Field) Approximation

The marginal posterior is approximated with the following variational bound:

$$p(\boldsymbol{w}, \beta | D, \gamma) \propto \sum_{\boldsymbol{s}} p(\boldsymbol{s}|\gamma) p(D|\boldsymbol{s}, \boldsymbol{w}, \beta)$$
$$\geq \exp\left(-\sum_{\boldsymbol{s}} q(\boldsymbol{s}) \log \frac{q(\boldsymbol{s})}{p(\boldsymbol{s}|\gamma) p(D|\boldsymbol{s}, \boldsymbol{w}, \beta)}\right),$$

Mean-Field approximation

$$q(\boldsymbol{s}) = \prod_{i=1}^{N} (m_i s_i + (1 - m_i)(1 - s_i))$$

Allows to specify q with only the expected values $m_i = q_i(s_i = 1)$.

Fixed Point Equations

For a given γ, expected values $\boldsymbol{m}$ of $\boldsymbol{s}$ and $\boldsymbol{w}, \beta$ are found iteratively

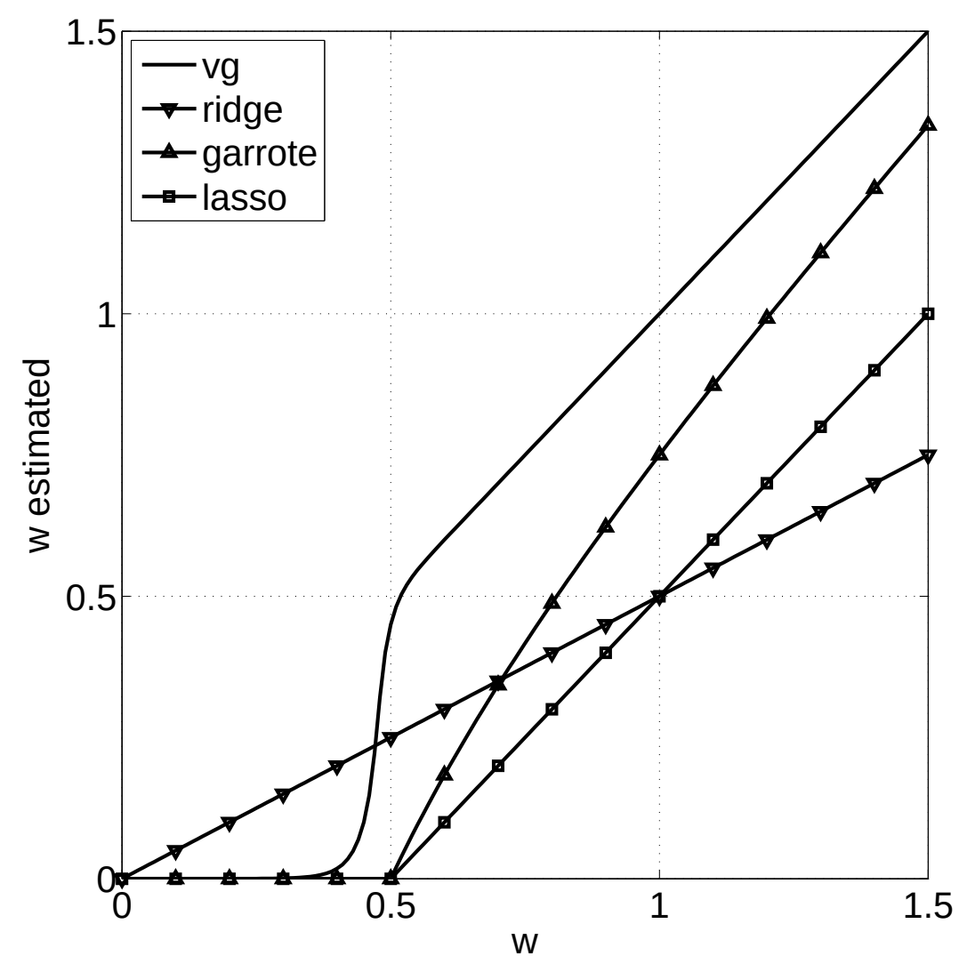
$$m_i = \sigma\left(\gamma + \frac{\beta p}{2} w_i^2 \chi_{ii}\right)$$
$$\boldsymbol{w} = (\chi')^{-1}\boldsymbol{b}, \qquad \chi'_{ij} = \chi_{ij} m_j + (1 - m_j)\chi_{jj}\delta_{ij}$$
$$\frac{1}{\beta} = \sigma_y^2 - \sum_{i=1}^{n} m_i w_i b_i$$

where $\sigma(x) = (1 + \exp(-x))^{-1}$ and $\sigma_y^2 = \frac{1}{p}\sum_\mu (y^\mu)^2$

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

- Similar to ridge regression, but with diagonal term depending on i and is dynamically adjusted depending on the solution for $\boldsymbol{m}$.
- The size of m_i (and thus the rank of χ') is controlled by γ

Comparison with Existing Methods



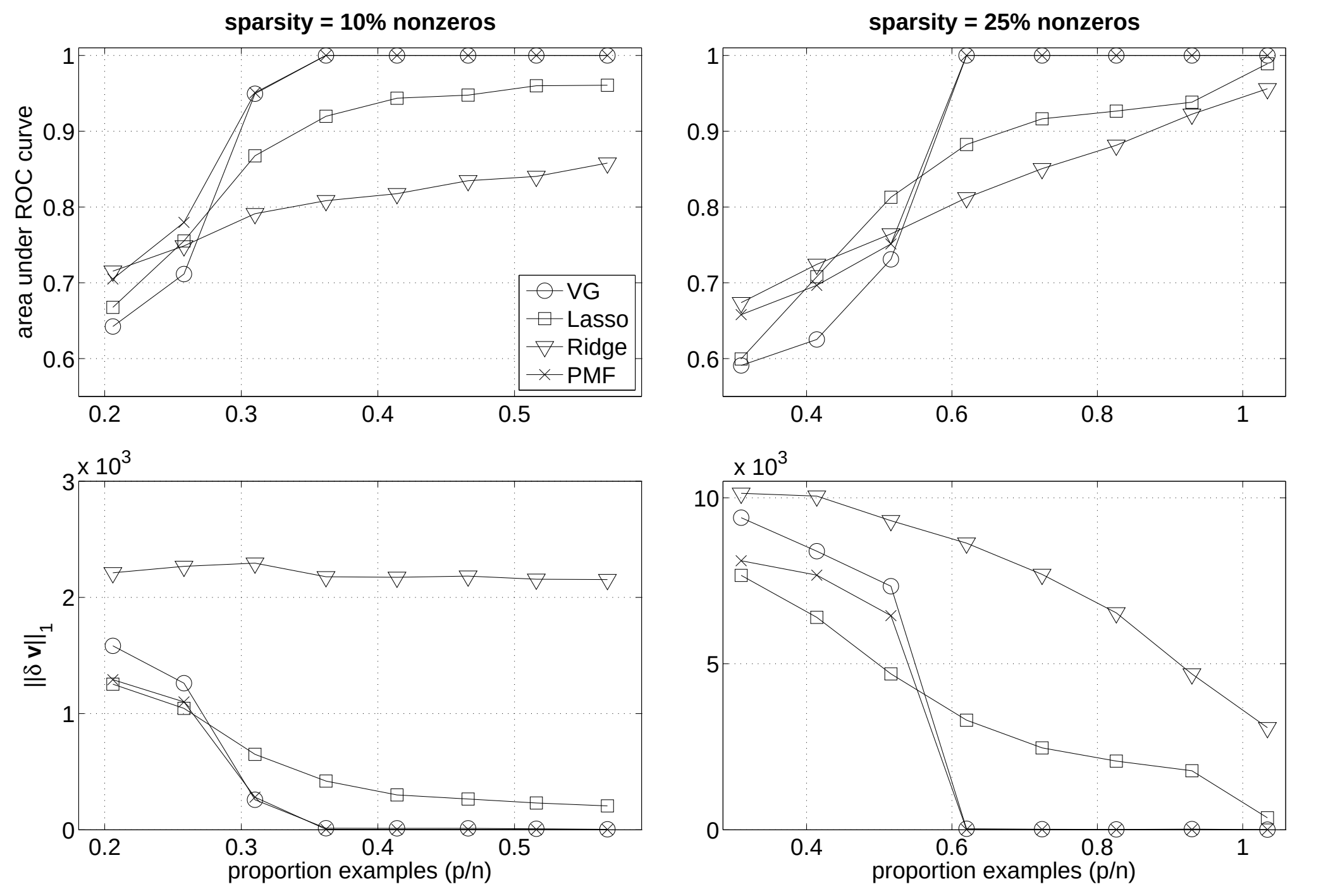
The VG gives an almost ideal behavior and can be interpreted as a soft version of variable selection:

- For small w, the solution is close to zero and the variable is ignored
- Above a threshold it is identical to the OLS solution.

Results on a Synthetic Data

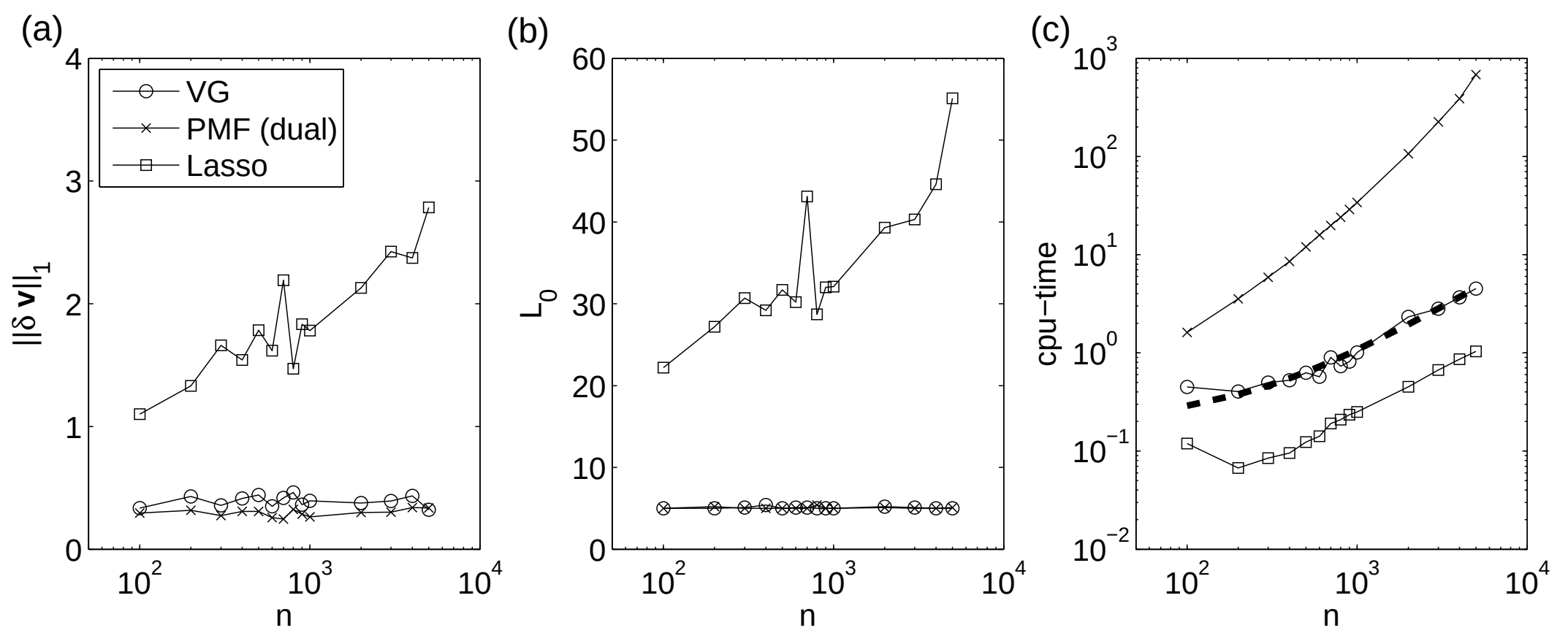
Performance as a function of the available training samples

- Comparison with Ridge Regression, Lasso and Paired-Mean field (PMF), another variational approximation
- **Top:** area under the ROC curves
- **Bottom:** reconstruction error, defined as $\|\delta\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |m_i w_i - \hat{w}_i|$



- **The VG shows better or comparable performance than any other method considered**

Performance as a function of the number of features n



(a): Error of the solution vector
(b): ℓ_0 of the solution vector
(c): Computational time in seconds

VG is much more efficient than methods that perform similarly

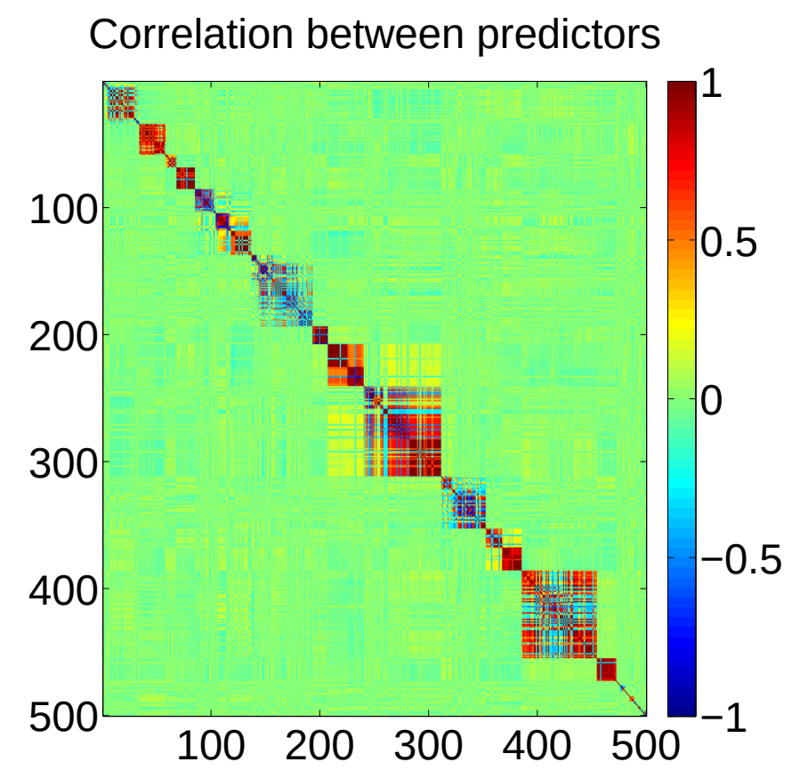
Results on a genetic dataset

Single Nucleotide Polymorphisms (SNPs)

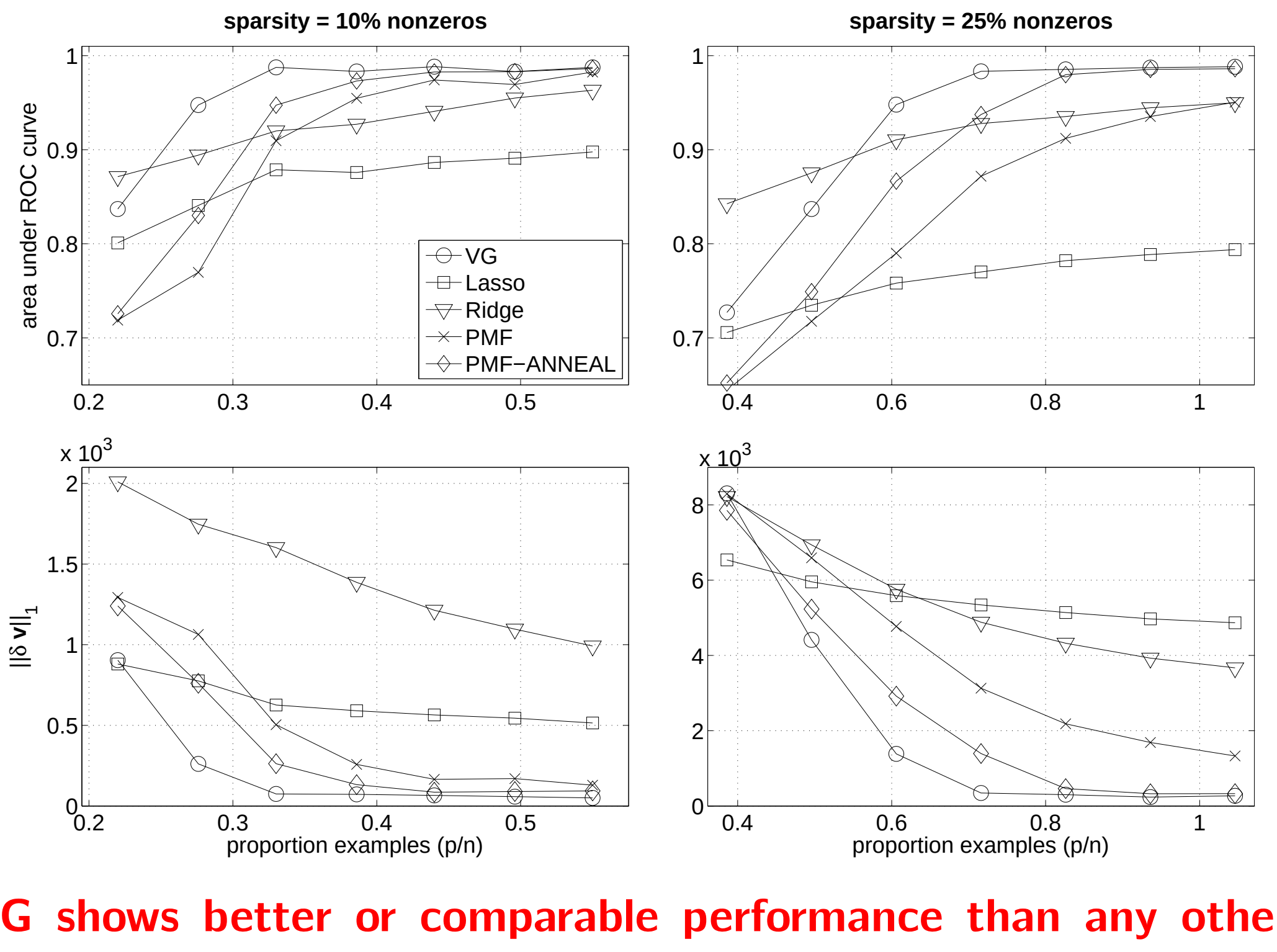
- **Input data** x_i: single nucleotide polymorphisms (SNPs) that have values $x_i = \{0, 1, 2\}$.
- **Output data**: generated artificially

The raw genetic dataset for that experiment included 928 samples of 2399 three-valued SNP predictors $\{0, 1, 2\}$.

SNPs show correlations structured in blocks, where nearby SNPs are highly correlated, but show no dependence on distant SNPs.



Performance as a function of the available training samples



VG shows better or comparable performance than any other method considered

[1] H. J. Kappen and V. Gómez. The variational garrote. *Machine Learning*, 96:269–294, 2014. **Matlab code available at**: `http://www.mbfys.ru.nl/staff/v.gomez/VG_1.0.tgz`.