Sparse Heteroscedastic Regression via Semiparametric Variational Inference

LSE Research Showcase (April 7 - 8, 2025)

Dr. Giulia Livieri The London School of Economics and Political Science (e-mail: g.livieri@lse.ac.uk)

joint work with: Prof. Mauro Bernardi (Universy of Padova), Dr. Luca Maestrini (The Australian National University)

A brief introduction to Variational Inference

- Lots of modern statistical inference problems, under the Bayesian framework, depends on the calculation of the posterior, via the famous Bayes Rule.

$$p(\boldsymbol{z}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{y})},$$

  where $\boldsymbol{y} = (y_1, \dots, y_n)$ denotes a set of observations and $\boldsymbol{z} = (z_1, \dots, z_m)$ denotes a set of latent variables (or parameters); $n, m \in \mathbb{N}_{>0}$

- The evidence, i.e., $p(\boldsymbol{y})$, can be difficult to calculate and intractable in some cases as it involves integrating over $\boldsymbol{z}$ (in discrete probability space it would be a summation).

$$p(\boldsymbol{y}) = \int_{\boldsymbol{z}} p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})\, d\boldsymbol{z}.$$

- Solution 1: observe that the evidence is only a function of the data points, and so

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}).$$

  For certain combination of the likelihood $p(\boldsymbol{y}|\boldsymbol{z})$ and the prior $p(\boldsymbol{z})$ pair, the posterior can be identified directly by pattern-matching with the product on the RHS (e.g., conjugate priors).

- Solution 2: further the above proportion and realize that:

$$\frac{p(\boldsymbol{z}_0|\boldsymbol{y})}{p(\boldsymbol{z}_1|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{z}_0)p(\boldsymbol{z}_0)}{p(\boldsymbol{y}|\boldsymbol{z}_1)p(\boldsymbol{z}_1)}$$

  This equation is exact and tells us the "relative entropy" in the posterior. This is actually an important factor within the "acceptance probability" in MCMC methods like M-H algorithm.

- Solution 3: Variational Inference (VI).
  Basic Idea: turn the approximation problem into an optimization problem.

  $$q^\star(\boldsymbol{z}) = \operatorname{argmin}_{q(\boldsymbol{z})\in \mathsf{Q}} \mathsf{L}(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{y}))$$

  Picking the distribution inside the variational family Q that minimizes the "difference" between it an the posterior density measured by a loss function L.
  - Q: the complexity of this "family" is related to the tradeoff between the accuracy of $q^\star$ and the difficulty to optimize.
  - L: often we use the Kullback-Leibler (KL) divergence; we will go along with KL divergence.

    $$\mathrm{KL}(p(\boldsymbol{x})||q(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{x})}[\log q(\boldsymbol{x})]$$

  - Apply to the VI optimization objective function:

    $$\mathrm{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{y})) = \mathbb{E}_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{y})] + \log p(\boldsymbol{y}).$$

  Question: Can we calculate these quantities exactly? Unfortunately the answer is No, because the log-evidence $\log p(\boldsymbol{y})$ is intractable.

- Define the evidence lower bound (ELBO) as

$$\begin{aligned}\mathrm{ELBO}(q(\boldsymbol{z})) &= \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{y})] - \mathbb{E}_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \\ &= \log p(\boldsymbol{y}) - \mathrm{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{y})),\end{aligned}$$

  where the equation above shows how ELBO is smaller or equal to the log-evidence (because $\mathrm{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{y})) \geq 0$)).
- Minimizing KL-divergence between $q(\boldsymbol{z})$ and $p(\boldsymbol{z}|\boldsymbol{y})$ is equivalent to maximizing the ELBO since the log-evidence is fixed.
- Rewriting the ELBO gives us additional intuitions:

$$\mathrm{ELBO}(q(\boldsymbol{z})) = \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{y}|\boldsymbol{z})] - \mathrm{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}))$$

  - To maximize the ELBO, we want to maximize the first term and minimize the second term on the RHS.
  - Max first term: we want to place densities $q(\cdot)$ on regions where the log-likelihood is high.
  - Min second term: we want to make $q(\cdot)$ similar to the prior.
  - We thus see the familiar likelihood-prior tradeoff in Bayesian statistics.

Variational Families

- Mean Field Variational Family

$$z = \{z_1, \dots, z_m\}, \quad q(z) = \prod_{i=1}^{m} q_i(z_i).$$

  - The latent variables are mutually independent and each governed by its own variational density $q_i(z_i)$. Easy to optimize with coordinate ascent variational inference (CAVI) algorithm.

Algorithm 1: CAVI

1: **procedure**
 Input: a model $p(\boldsymbol{y}, \boldsymbol{z})$, a dataset $\boldsymbol{y}$
 Output: variational density $q(\boldsymbol{z}) = \prod_{i=1}^{m} q_i(z_i)$.
 Initialize: variational factors $q_i(z_i)$.
 While: the ELBO has not converged **do**
 for $j \in 1, \dots, m\}$ **do**
  $\log q_j(z_j) \propto \mathbb{E}_{-j}[\log p(\boldsymbol{y}, \boldsymbol{z})]$
 end while
2: **end procedure**

- Sem-parametric Mean Field Variational Family

$$q(\boldsymbol{z}) = \prod_{i=1}^{m} q_i(z_i) = q_1(z_1) \cdot q_2(z_2) \cdots \tilde{q}_j(z_j; \text{par}) \cdots q_m(z_m).$$

  - Some density functions (e.g., $\tilde{q}_j(z_j; \text{par})$) in the product density restriction are pre-specified to be members of convenient parametric families (e.g., Gaussian) that can be conveniently chosen for reasons of tractability.

- Main Reference: Rohde, D. and Wand, M. P. (2016). Semiparametric mean field variational bayes: general principles and numerical issues. The Journal of Machine Learning Research, 17(1):5975-6021.

- Important: Despite its popularity, the convergence of CAVI remains poorly understood. Research Opportunities.

In this work we have introduced a new Skew-Normal Variational Family.

Heteroscedastic Regression
Why a new Skew-Normal Variational Family? Insights from a toy model

- Consider the following heteroscedastic regression model

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\varrho}, \sigma^2 \sim \mathrm{N}\big(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{diag}\{\exp(\boldsymbol{Z}\boldsymbol{\varrho})\}\big),$$

for a given $n \times 1$ vector of responses $\boldsymbol{y}$, and matrices of covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively of size $n \times p$ and $n \times q$.

- Assume an hierarchical Normal-Gamma prior for $(\boldsymbol{\beta}, \sigma^2)$, e.g.:

$$\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}(0, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \qquad \sigma^2 \sim \mathrm{IG}(\nu, \lambda), \quad \nu, \lambda > 0.$$

- The conditional joint distribution of the model parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\varrho})^\top$ is given by the following expression, where $\boldsymbol{\Omega}_{\boldsymbol{\varrho}} = \mathrm{diag}\{\exp(-\boldsymbol{Z}\boldsymbol{\varrho})\}$

$$\begin{aligned}
&p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\varrho}|\boldsymbol{y}, \boldsymbol{X}) \\
&\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}1_n^\top \boldsymbol{Z}\boldsymbol{\varrho}\right) \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}_{\boldsymbol{\varrho}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} \\
&\times (\sigma^2)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\beta})\right\} \times (\sigma^2)^{-(\nu+1)} \exp\left(-\frac{\lambda}{\sigma^2}\right). \quad (1)
\end{aligned}$$

- The full conditional density of $\boldsymbol{\beta} \in \mathbb{R}^p$ is

$$p(\beta|\boldsymbol{y}, \boldsymbol{X}, \sigma^2, \boldsymbol{\varrho}) \sim \mathrm{N}(\widehat{\boldsymbol{\Sigma}}_n \boldsymbol{X}^\top \boldsymbol{\Omega}_{\boldsymbol{\varrho}} \boldsymbol{y}, \sigma^2 \widehat{\boldsymbol{\Sigma}}_n).$$

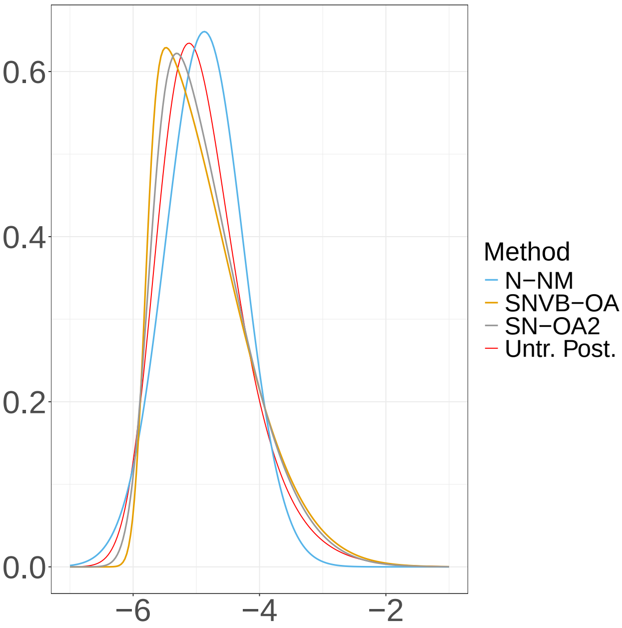
where $\widehat{\boldsymbol{\Sigma}}_n = \left(\boldsymbol{X}^\top \boldsymbol{\Omega}_{\boldsymbol{\varrho}} \boldsymbol{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)^{-1}$.

Proposition (BLM 2025)

Let $\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\varrho}, \sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{diag}\{\exp(\boldsymbol{Z}\boldsymbol{\varrho})\})$ for a given $n \times 1$ vector of responses $\boldsymbol{y}$, and matrices of covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively of size $n \times p$ and $n \times q$, and assume an hierarchical Normal-Gamma prior for $(\boldsymbol{\beta}, \sigma^2)$ such that $\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}(0, \sigma^2 \boldsymbol{\Sigma}_\beta)$ and $\sigma^2 \sim \mathrm{IG}(\nu, \lambda)$, where $\nu, \lambda > 0$. Then, the marginal likelihood is proportional to

$$\ell(\boldsymbol{\varrho}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) \propto \exp\left(-\frac{1}{2} 1_n^\top \boldsymbol{Z}\boldsymbol{\varrho}\right) |\boldsymbol{\Sigma}_\beta|^{-\frac{1}{2}} |\widehat{\boldsymbol{\Sigma}}_n|^{\frac{1}{2}} \left(\lambda + \frac{S^2}{2}\right)^{-(\nu + n/2)}, \tag{2}$$

where $S^2 = \boldsymbol{y}^\top \boldsymbol{H}_{\boldsymbol{\varrho}} \boldsymbol{y}$, with $\boldsymbol{H}_\varrho = (\boldsymbol{\Omega}_{\boldsymbol{\varrho}}^{-1} + \boldsymbol{X}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{X}^\top)^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_n = \left(\boldsymbol{X}^\top \boldsymbol{\Omega}_{\boldsymbol{\varrho}} \boldsymbol{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)^{-1}$.


0.6
0.4
0.2
0.0
−6
−4
−2
Method
N−NM
SNVB−OA
SN−OA2
Untr. Post.

- Let assume $\tilde{q}(\boldsymbol{\varrho}, \boldsymbol{\xi})$ is a general approximating density for the unnormalized intractable posterior distribution in Equation (2).
- The ELBO is (Rohde, D. and Wand, M. P. (2016), Equation (13)):

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y}, \tilde{q}) &= \int \tilde{q}\big(\boldsymbol{\varrho}; \boldsymbol{\xi}\big) \log\{\ell\big(\boldsymbol{\varrho}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}\big)\} d\boldsymbol{\varrho} + \text{Entropy}\{\tilde{q}\big(\boldsymbol{\varrho}; \boldsymbol{\xi}\big)\} \\
&= \overline{H}_{\tilde{q}(\boldsymbol{\varrho})}(\boldsymbol{\varrho}, \boldsymbol{\xi}) + H_{\tilde{q}(\boldsymbol{\varrho})}(\boldsymbol{\varrho}, \boldsymbol{\xi}) \\
H_{\tilde{q}(\boldsymbol{\varrho})}(\boldsymbol{\varrho}, \boldsymbol{\xi}) &= -\mathrm{E}_{\tilde{q}(\boldsymbol{\varrho})}\big[\log q(\boldsymbol{\varrho}, \boldsymbol{\xi})\big] \\
\overline{H}_{\tilde{q}(\boldsymbol{\varrho})}(\boldsymbol{\varrho}, \boldsymbol{\xi}) &= -\frac{1}{2} 1_n^\top \boldsymbol{Z} \mathrm{E}_{\tilde{q}(\boldsymbol{\varrho})}(\boldsymbol{\varrho}) - \frac{1}{2} \log |\boldsymbol{\Sigma}_\beta| \\
&\quad + \frac{1}{2} \mathrm{E}_{\tilde{q}(\boldsymbol{\varrho})}\Big[\log |\widehat{\boldsymbol{\Sigma}}_n|\Big] - (\nu + n/2) \mathrm{E}_{\tilde{q}(\boldsymbol{\varrho})}\Big[\log\Big(\lambda + \frac{S^2}{2}\Big)\Big].
\end{aligned}
$$

- Main problems:
  - How to select the approximating distribution;
  - The approximating distribution depends on additional parameters, that should be calibrated; e.g., the mean and variance-covariance matrix for the Gaussian distribution.

A new multivariate Skew-Normal distribution

- The standard multivariate Skew-Normal distribution $\boldsymbol{X} \overset{d}{\sim} \mathrm{SN}_d(0, \boldsymbol{I}_d, \boldsymbol{\alpha})$, with $\boldsymbol{\alpha}$ the skewness parameter, has p.d.f. given by $p_{\boldsymbol{X}}(\boldsymbol{x}) = 2\varphi_d(\boldsymbol{x})\Phi_1(\boldsymbol{\alpha}^\top \boldsymbol{x})$, where $\varphi_d(\cdot)$ is the p.d.f. of a d-dimensional Gaussian distribution and $\Phi_1(\cdot)$ is the c.d.f. of a 1-dimensional Gaussian distribution.

Proposition (BLM 2025)

$\boldsymbol{X} \overset{d}{\sim} \mathrm{SN}_d(0, \boldsymbol{I}_d, \boldsymbol{\alpha})$, then the affine transformation $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{X}$ where $\boldsymbol{\Sigma}$ is a proper variance-covariance matrix such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^\top$, is $\boldsymbol{Y} \overset{d}{\sim} \mathrm{SN}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ with p.d.f

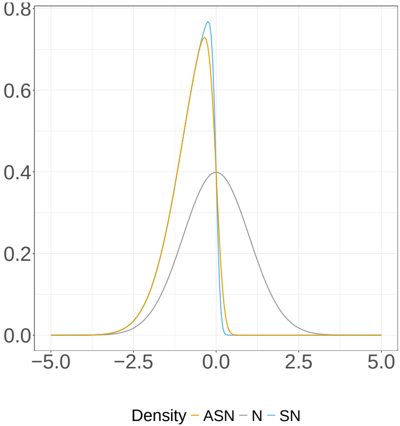
$$p_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = 2\varphi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})\Phi_1(\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\mu})),$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is a d-dimensional vector of location parameters, $\boldsymbol{\Sigma}$ is a positive definite square matrix of dimension d, $\boldsymbol{\alpha} \in \mathbb{R}^d$ is the skewness parameter. Important: In Azzalini, A. and Capitano, A. (2013) $\boldsymbol{Y} = \boldsymbol{\mu} + \mathrm{diag}(\boldsymbol{\Sigma})\boldsymbol{X}$: relevant problems with the optimization.

Main advantages:

- It relies on $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ in both $\varphi_d(\cdot)$ and $\Phi_1(\cdot)$.
- It retains the same properties of the parametrization by Azzalini, A. and Capitano, A. (2013).
- It is enough flexible to account the Gaussian case for $\boldsymbol{\alpha} = 0$ as well as skewness with just one additional parameter;
- Right compromise between accounting skewness and complexity of the resulting algorithm: the skewness parameter $\boldsymbol{\alpha}$ only enters $\Phi_1(\cdot)$;
- In our Variational Inference algorithm it requires numerical integration of just a scalar parameter.
- Important: In our Variational Inference algorithm it eliminates the need for a Newton-Raphson update for $\boldsymbol{\Sigma}_{\tilde{q}}$.

Main Reference: Azzalini, A. and Capitanio, A. (2013). *The Skew-Normal and Related Families.* Institute of Mathematical Statistics Monographs. Cambridge University Press.



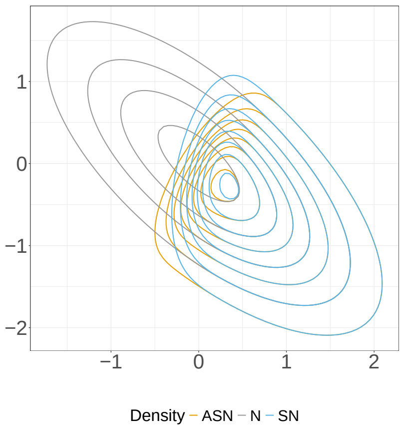


Figure: Contour plots of two different definitions of Skew-Normal distributions. The parameter values are set as follows: $\boldsymbol{\mu} = (0,0)^\top$, $\boldsymbol{\alpha} = (-5,2)^\top$, $\boldsymbol{\Sigma}_{11} = 2$, $\boldsymbol{\Sigma}_{22} = 4$, $\boldsymbol{\Sigma}_{12} = -2$, $\rho = -0.7$. The gray lines represent the density and contour plots of the corresponding bivariate Gaussian distribution.

Variable Selection (so far, numerical results only for the homoscedastic case)

Spike-and-Slab Lasso prior

$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}), \quad \sigma^2|\alpha \sim \text{Inverse-}\chi^2(1,1/\alpha), \quad \alpha \sim \text{Inverse-}\chi^2(1,1/s_\sigma^2),$$
$$\beta_j \,|\, \sigma^2,\omega_{0j},\omega_{1j},\gamma_j \sim N\big(0,\sigma^2\left\{(1-\gamma_j)\omega_{0j}+\gamma_j\omega_{1j}\right\}^{-1}\big),$$
$$\omega_{0j} \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(2,\lambda_0), \quad \omega_{1j}|\lambda_1 \overset{\text{ind.}}{\sim} \text{Inverse-}\chi^2(2,\lambda_1), \quad \gamma_j|\vartheta \overset{\text{ind.}}{\sim} \text{Bernoulli}(\vartheta),$$
$$j=1,\ldots,p,$$
$$\lambda_1 \sim \text{Gamma}(d_{\lambda_1},r_{\lambda_1}), \quad \vartheta \sim \text{Beta}(a_\vartheta,b_\vartheta),$$

where $\boldsymbol{y}$ is an output vector of length n, $\boldsymbol{X}$ is a design matrix of size $n \times p$, $s_\sigma, d_{\lambda_1}, r_{\lambda_1}, a_\vartheta, b_\vartheta > 0$ are user-specified hyperparameters.

Intuition: If λ is large, then the density of ω is very peaked around zero; If λ is small it is diffusive.

We typically set $\lambda_0 \gg \lambda_1$, so that ω_0 is the spike and ω_1 the slab; γ indexes the 2^p possible models, and ϑ is a mixing proportion.

Homoscedastic data

- In detail, let m_1 and n two integers parameters and suppose the existence of two groups of diets with $n/2$ subjects in each group. Then, $m_1 + 1$ explanatory variables are generated in the following way.
- First, a binary diet indicator z is generated for each subject $i = 1, \ldots, n$, $z_i = 1_{\{i > n/2\}} - 1_{\{i < n/2\}}$.
- Then, we generate $\mathrm{x}_k = [x_{1,k}, \ldots, x_{n,k}]^\top$, $k = 1, \ldots, m_1$, such that $x_{ik} = u_{ik} + z_i v_k$, where u_{ik} are independent uniform $U([0,1])$ random variables, and $v_1, \ldots, v_{0.75 m_1}$ are independent uniform $U([0.25, 0.75])$ random variables, and $v_{0.75 m_1 + 1}, \ldots, v_{0.75 m_1}$ are identically zero. In this way, there are m_1 variables $x_1, \ldots, x_{m_1}$ where the first 75% of the $x's$ depend on z.
- In the end, the response vector is generated as follows:

$$\boldsymbol{y} = \beta_1 z + \beta_2 \mathrm{x}_1 + \beta_3 \mathrm{x}_2 + \beta_4 \mathrm{x}_3 + \sum_{k=5}^{m_1} \beta_k \mathrm{x}_{k-1} + \beta_{m_1+1} \mathrm{x}_{m_1} + \varepsilon, \quad (3)$$

where ε is normally distributed with mean zero and covariance $\sigma^2 \mathrm{I}$.

- We use the following values for the parameters: $m_1 = 40$, $n = 80$, $\sigma^2 = 1$ and $\beta = (1 - (\kappa - 1)/12) \times (4.5,\ 3,\ -3,\ -3,\ 0^\top,\ 3)$ where $0^\top$ is an $(m_1 - 4)-$dimensional vector of zeros and κ is a simulation parameter.
- The data $x_1, \ldots, x_{m_1}$ are generated according to four distinct categories with a well defined interpretation. In particular, correlations for the first $0.75\, m_1$ variables are around 0.8 in absolute magnitude. The remaining variables are independent from each other and the first $0.75\, m_1$ variables.
- We generate 100 independent data sets for each value of κ in the set $\{1, 2, 3, 4, 5, 6, 7\}$ and apply each of the variable selection procedure we consider. The results are discussed in the following.

- MSE measure comparison between:
  - least absolute shrinkage and selection operator (LASSO).
  - smoothly clipped absolute deviation (SCAD) penalty.
  - minimax concave penalty (MCP) through the R package *ncvreg*.
  - expectation maximization variable selection (EMVS).
  - Bayesian model selection (BMS).

Method	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$	$\kappa = 6$	$\kappa = 7$
lasso	6.613 (0.151)	5.821 (0.084)	4.845 (0.062)	4.017 (0.038)	3.271 (0.028)	2.635 (0.025)	2.021 (0.018)
scad	3.44 (0.303)	3.306 (0.227)	2.811 (0.205)	2.357 (0.17)	2.076 (0.119)	1.863 (0.083)	1.524 (0.057)
mcp	2.636 (0.322)	2.368 (0.287)	1.965 (0.235)	1.754 (0.197)	1.375 (0.153)	1.397 (0.118)	1.387 (0.075)
emvs	8.296 (0.322)	7.131 (0.257)	6.583 (0.203)	6.062 (0.122)	5.468 (0.103)	3.717 (0.095)	3.543 (0.051)
bms	2.772 (0.058)	2.266 (0.031)	2.064 (0.057)	1.89 (0.074)	1.582 (0.057)	1.475 (0.07)	1.34 (0.073)
vbl	3.321 (0.024)	2.904 (0.022)	2.501 (0.021)	2.066 (0.017)	1.728 (0.013)	1.395 (0.011)	1.074 (0.009)
vbss	1.815 (0.137)	1.63 (0.152)	1.515 (0.169)	1.326 (0.133)	1.244 (0.085)	1.121 (0.035)	0.931 (0.017)
vbssl	1.941 (0.2)	1.967 (0.212)	1.849 (0.152)	1.626 (0.088)	1.424 (0.039)	1.171 (0.014)	0.912 (0.012)
gsl	3.077 (0.029)	2.715 (0.025)	2.363 (0.023)	1.969 (0.018)	1.665 (0.014)	1.359 (0.012)	1.053 (0.01)
gsss	1.923 (0.113)	1.677 (0.084)	1.621 (0.122)	1.672 (0.147)	1.797 (0.114)	1.601 (0.051)	1.272 (0.023)
gsssl	1.864 (0.06)	1.631 (0.041)	1.44 (0.053)	1.208 (0.047)	1.075 (0.047)	0.913 (0.042)	0.739 (0.039)

- F1 measure comparison between Lasso, SCAD, MCP, EMVS, BMS, VB and MCMC based on 100 simulations for the diet simulation example.

Method	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$	$\kappa = 6$	$\kappa = 7$
lasso	0.937 (0.028)	0.946 (0.019)	0.94 (0.022)	0.939 (0.023)	0.939 (0.022)	0.927 (0.096)	0.938 (0.019)
scad	0.945 (0.026)	0.947 (0.024)	0.953 (0.023)	0.955 (0.021)	0.954 (0.022)	0.954 (0.017)	0.953 (0.015)
mcp	0.959 (0.029)	0.965 (0.027)	0.967 (0.028)	0.968 (0.027)	0.975 (0.024)	0.969 (0.022)	0.959 (0.018)
emvs	0.966 (0.036)	0.962 (0.029)	0.961 (0.027)	0.929 (0.166)	0.839 (0.312)	0.632 (0.446)	0.275 (0.432)
bms	0.998 (0.006)	0.999 (0.004)	0.996 (0.009)	0.992 (0.012)	0.989 (0.012)	0.981 (0.016)	0.957 (0.098)
vbss	0.995 (0.008)	0.993 (0.01)	0.985 (0.015)	0.95 (0.168)	0.846 (0.329)	0.701 (0.429)	0.39 (0.471)
vbssl	0.922 (0.255)	0.715 (0.437)	0.581 (0.477)	0.309 (0.453)	0.154 (0.354)	0.105 (0.3)	0.029 (0.163)
gsss	0.98 (0.098)	0.983 (0.081)	0.934 (0.193)	0.828 (0.287)	0.583 (0.338)	0.416 (0.303)	0.255 (0.221)
gsssl	0.99 (0.019)	0.992 (0.014)	0.993 (0.012)	0.992 (0.015)	0.987 (0.016)	0.984 (0.015)	0.976 (0.017)

- Accuracy analysis.

Method	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$	$\kappa = 6$	$\kappa = 7$
vbl	84.234 (4.576)	85.252 (4.028)	86.462 (3.66)	87.471 (3.281)	88.204 (2.988)	88.794 (2.784)	89.382 (2.651)
vbss	86.448 (13.892)	86.113 (12.324)	80.059 (18.505)	67.744 (28.53)	45.944 (31.358)	34.942 (27.191)	30.597 (24.857)
vbssl	81.029 (10.887)	77.406 (15.344)	76.098 (16.352)	76.11 (16.933)	76.343 (16.667)	77.403 (15.409)	79.282 (13.45)
vbl	26.738 (3.542)	30.965 (3.982)	36.831 (5.257)	42.394 (6.697)	49.05 (8.536)	55.946 (10.27)	64.019 (9.188)
vbss	63.389 (18.531)	61.689 (18.475)	57.388 (21.197)	50.349 (26.47)	36.42 (30.249)	32.802 (22.563)	40.389 (19.304)
vbssl	64.189 (28.865)	39.231 (35.731)	27.312 (29.171)	19.889 (21.025)	19.726 (16.479)	25.256 (14.537)	35.38 (12.828)

THANK YOU!