File S1

Supporting Information

S.1 Notation

We briefly recall here the notation that was used along the paper. Moreover we introduce some new notation to easy the the illustration of the MCMC scheme.

Let $\boldsymbol{Y}$ and $\boldsymbol{X}$ the $n \times q$ and $n \times p$ matrix of the responses and predictors, respectively. Let $\boldsymbol{\Gamma} = \{\gamma_{lkj}, 1 \leq l \leq L, 1 \leq k \leq q, 1 \leq j \leq p\}$ the matrix of latent binary values, where L is the number of simulated chains, q is the number of responses and p is the number of predictors and let $\boldsymbol{\Gamma}_k = (\boldsymbol{\gamma}_{1k}, \ldots, \boldsymbol{\gamma}_{lk}, \ldots, \boldsymbol{\gamma}_{Lk})^T$ the $L \times p$ latent binary matrix for the kth response in expanded state-space, where $\boldsymbol{\gamma}_{lk} = (\gamma_{lk1}, \ldots, \gamma_{lkj}, \ldots, \gamma_{lkp})^T$. Similarly let $\boldsymbol{\Omega} = \{\omega_{lkj}, 1 \leq l \leq L, 1 \leq k \leq q, 1 \leq j \leq p\}$ the matrix of selection probability with $\omega_{lkj} = \omega_{lk} \times \rho_{lj}$ and let $\boldsymbol{\Omega}_k = (\boldsymbol{\omega}_{1k}, \ldots, \boldsymbol{\omega}_{lk}, \ldots, \boldsymbol{\omega}_{Lk})^T$ the $L \times p$ selection matrix for the kth response in expanded state-space, where $\boldsymbol{\omega}_{lk} = (\omega_{lk1}, \ldots, \omega_{lkj}, \ldots, \omega_{lkp})^T$. For a given chain l, let $\boldsymbol{\omega}_l = (\omega_{l1}, \ldots, \omega_{lk}, \ldots, \omega_{lq})^T$ and $\boldsymbol{\rho}_l = (\rho_{l1}, \ldots, \rho_{lj}, \ldots, \rho_{lp})^T$ the 'row' and the 'column' effect, respectively. Finally the temperature ladder for each regression equation k is denoted by $\boldsymbol{t}_k = (t_{1k}, \ldots, t_{lk}, \ldots, t_{Lk})^T$ with $1 = t_{1k} < t_{2k} < \ldots < t_{Lk}$.

S.2 Technical details of MCMC implementation

S.2.1 Full conditionals

Given (6), to sample the binary latent value γ_{lkj}, the selection probability $\omega_{lkj} = \omega_{lk} \times \rho_{lj}$ and the scaling coefficient τ, the tempered full conditionals in the expanded state-space are:

- $p(\boldsymbol{\gamma}_{lk}|\cdots) \propto p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau)^{1/t_{lk}} \prod_{j=1}^{p} p(\gamma_{lkj}|\omega_{lkj})^{1/t_{lk}}$
- $p(\omega_{lk}|\cdots) \propto p(\omega_{lk})^{1/t_{lk}} \prod_{j=1}^{p} p(\gamma_{lkj}|\omega_{lkj})^{1/t_{lk}}$
- $p(\rho_{lj}|\cdots) \propto p(\rho_{lj}) \prod_{k=1}^{q} p(\gamma_{lkj}|\omega_{lkj})^{1/t_{lk}}$
- $p(\tau|\cdots) \propto p\left(\tau\right) \prod_{l=1}^{L} \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau)^{1/t_{lk}}$

Note that in the full conditional $p(\rho_{lj}|\cdots)$ the prior density $p(\rho_{lj})$ is not tempered and the reason will be explained in Supporting Information S.2.3.

S.2.2 $\boldsymbol{\Gamma}$ update

The update of the elements of the $q \times p$ latent binary matrix $\boldsymbol{\Gamma}$ is of paramount importance and efficient algorithms are required in order to visit the very large model space $(2^p)^q$ and to escape from local modes. In the following we provide some technical details omitted from the main text of the local and global moves that we found useful to implement. At each sweep of the algorithm each/both of moves can be applied to *all* the q regression equations or to a random without replacement subgroup of them (see Richardson et al. (2011) for alternative subgroup selection with adaptive probability).

Local move

We first introduce the single chain sampling scheme and then we extend the results for multiple chains. There are many ways to update locally $\boldsymbol{\gamma}_k$, but we found useful to apply an extension of Bottolo and Richardson (2010) proposal, where traditional samplers used in Bayesian variable selection (*i.e.* MC^3, Gibbs sampler and Reversible Jump) are replaced by a Metropolis-within-Gibbs sampler known as Fast Scan Metropolis-Hastings (FSMH). Let $L_{k(j=1)} = p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{k(j=1)}, \tau)$ and $L_{k(j=0)} = p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{k(j=0)}, \tau)$ with $\boldsymbol{\gamma}_{k(j=1)} = (\gamma_{k1}, \ldots, \gamma_{kj} = 1, \ldots, \gamma_{kp})^T$ and $\boldsymbol{\gamma}_{k(j=0)} = (\gamma_{k1}, \ldots, \gamma_{kj} = 0, \ldots, \gamma_{kp})^T$ the marginal likelihood once the regression coefficients $\boldsymbol{\beta}_k$ and the residual error variance σ_k^2 are integrated out. Moreover let $p(\gamma_{kj} = 1|\omega_{kj}) = \omega_{kj}$ and $p(\gamma_{kj} = 0|\omega_{kj}) = 1 - \omega_{kj}$. If a Gibbs sampler update is performed, a new value of γ_{kj} is drawn from a Bernoulli distribution with probability

$$\theta_{kj} = \frac{\omega_{kj} L_{k(j=1)}}{(1 - \omega_{kj}) L_{k(j=0)} + \omega_{kj} L_{k(j=1)}} \tag{S.1}$$

if, in the previous iteration, $\gamma_{kj} = 0$ since by independence $p(\gamma_{kj} = 1|\boldsymbol{\gamma}_{k\setminus j}, \boldsymbol{\omega}_k) = p(\gamma_{kj} = 1|\omega_{kj})$ (with an obvious modification if $\gamma_{kj} = 1$ in the previous iteration). However in a sparse framework, where $p_{\boldsymbol{\gamma}_k} \ll p$, this probability is dominated by ω_{kj} and if ω_{kj} is small (because for instance ω_k or ρ_j or both are small) also θ_{kj} will be small. For instance, it easy to show that when $p_{\boldsymbol{\gamma}_k} \ll p$ and therefore by Kohn et al. (2001) $a_k \ll b_k$, the sampled value of ω_k is, on average, very small

$$E(\omega_k|\boldsymbol{y}_k) = \frac{p_{\boldsymbol{\gamma}_k} + a_k}{p + a_k + b_k}.$$

It turns out that, if $\gamma_{kj} = 0$, it is likely that also the new sampled value will be zero. Kohn et al. (2001) propose to split the acceptance probability of the Metropolised version of (S.1) (to add a new covariate in the regression)

$$1 \wedge \frac{\omega_{kj} L_{k(j=1)}}{(1-\omega_{kj}) L_{k(j=0)}} \frac{Q_{kj}(1 \to 0)}{Q_{kj}(0 \to 1)},$$

where $Q_{kj}(\cdot \to \cdot)$ is the proposal density, into two parts: firstly, sampling a proposed value of γ_{kj}, γ^*_{kj}, from a Bernoulli distribution with probability ω_{kj} and then, if $\gamma^*_{kj} \neq \gamma_{kj}$, accept the new value with probability

$$1 \wedge \frac{L_{k(j=1)}}{L_{k(j=0)}}$$

since $Q_{kj}(0 \to 1) = \omega_{kj}$ and $Q_{kj}(1 \to 0) = 1 - \omega_{kj}$, with an obvious modification if a deletion is proposed. The advantage of this scheme is that the time consuming evaluation of the marginal likelihood L_{kj} is limited to the set of variables where $\gamma^*_{kj} \neq \gamma_{kj}$.

The same sampling scheme can be extended to a parallel tempering set-up as illustrated in Bottolo and Richardson (2010). In this case the Metropolis-within-Gibbs acceptance probability of the jth predictor in the kth regression and the lth chain is

$$1 \wedge \frac{L^{1/t_{lk}}_{lk(j=1)}}{L^{1/t_{lk}}_{lk(j=0)}},$$

where $L^{1/t_{lk}}_{lk(j=1)} = [p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk(j=1)}, \tau)]^{1/t_{lk}}$ and similarly for $L^{1/t_{lk}}_{lk(j=0)}$, since adding (deleting) a covariate in the regression equation is proposed with probability $Q_{lkj}(0 \to 1|t_{lk}) = \tilde{\omega}_{lkj}(t_{lk})$ $(Q_{lkj}(1 \to 0|t_{lk}) = 1 - \tilde{\omega}_{lkj}(t_{lk}))$, with

$$\tilde{\omega}_{lkj}(t_{lk}) = \frac{\omega^{1/t_{lk}}_{lkj}}{\omega^{1/t_{lk}}_{lkj} + (1-\omega_{lkj})^{1/t_{lk}}}$$

the renormalised probability $[p(\gamma_{lkj} = 1|\omega_{lkj})]^{1/t_{lk}} = \omega^{1/t_{lk}}_{lkj}$ and t_{lk} the temperature attached to the kth regression in the lth chain. Further discussion and advantages of this sampling scheme over MC^3, Reversible Jump and Gibbs sampler in a multiple chain set-up when the number of predictors is very large with respect to the number of truly associated variables are presented in Bottolo and Richardson (2010).

Global moves

We recall that global moves are bold moves that try to swap part or the whole state of two randomly selected chains among the population of chains (Liang and Wong, 2000). In the following we present the accepted probability of crossover operator (partial swap), exchange operator and all-exchange operator (full swap).

Suppose that in the kth regression two new latent binary vectors $\boldsymbol{\gamma}_{lk}^{*}$ and $\boldsymbol{\gamma}_{rk}^{*}$ are generated from two preselected chains, l and r, according to some crossover operator (Liang and Wong, 2000; Bottolo and Richardson, 2010). The proposed population of chains in the kth regression $\boldsymbol{\Gamma}_k^{*} = (\boldsymbol{\gamma}_{1k}, \dots, \boldsymbol{\gamma}_{lk}^{*}, \dots, \boldsymbol{\gamma}_{rk}^{*} \dots, \boldsymbol{\gamma}_{Lk})^T$ is accepted with probability

$$1 \wedge \frac{\exp\left\{f(\boldsymbol{\gamma}_{lk}^{*}|\boldsymbol{\omega}_{lk}, \tau)/t_{lk} + f(\boldsymbol{\gamma}_{rk}^{*}|\boldsymbol{\omega}_{rk}, \tau)/t_{rk}\right\}}{\exp\left\{f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk}, \tau)/t_{lk} + f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{rk}, \tau)/t_{rk}\right\}} \frac{Q_k(\boldsymbol{\Gamma}_k^{*}, \boldsymbol{\Gamma}_k|\boldsymbol{\Omega}_k, \tau, \boldsymbol{t}_k)}{Q_k(\boldsymbol{\Gamma}_k, \boldsymbol{\Gamma}_k^{*}|\boldsymbol{\Omega}_k, \tau, \boldsymbol{t}_k)},$$

where $f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk}, \tau) = \log(p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau)) + \sum_j \log(p(\gamma_{lkj}|\omega_{lkj}))$ and $Q_k(\boldsymbol{\Gamma}_k, \cdot|\boldsymbol{\Omega}_k, \tau, \boldsymbol{t}_k)$ is the proposal density which is defined as the product of the selection probability and the crossover operator probability (Liang and Wong, 2000). The transition density depends on the selection probabilities $\boldsymbol{\Omega}_k$ in the kth regression, the scaling coefficient τ and the kth regression temperature ladder $\boldsymbol{t}_k$.

The exchange operator can be seen as special case of the crossover operator where the whole information contained in the two preselected chains with uniform probability l and r are tentatively swapped with probability

$$1 \wedge \frac{\exp\left\{f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{lk}, \tau)/t_{lk} + f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{rk}, \tau)/t_{rk}\right\}}{\exp\left\{f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk}, \tau)/t_{lk} + f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{rk}, \tau)/t_{rk}\right\}}$$

since $Q_k(\mathbf{\Gamma}_k, \mathbf{\Gamma}_k^*|\mathbf{\Omega}_k, \tau, \boldsymbol{t}_k) = Q_k(\mathbf{\Gamma}_k^*, \mathbf{\Gamma}_k|\mathbf{\Omega}_k, \tau, \boldsymbol{t}_k)$ because the selection probability is uniform over the L chains (random selection without replacement).

Finally, in the all-exchange operator the chains whose states are swopped are selected at random with probability equal to

$$p_{hk} = \frac{\tilde{p}_{hk}}{\sum_{h=1}^{1+L(L-1)/2} \tilde{p}_{hk}}, \tag{S.2}$$

where in (S.2) each pair $(l, r < l)$ is denoted by a single number h, $\tilde{p}_{hk} = \tilde{p}_{(l,r)k}$, including the rejection move, $h = 1$ with $\tilde{p}_{(l,r)k} = \exp\{(f(\boldsymbol{\gamma}_{rk}|\boldsymbol{\omega}_{rk}, \tau) - f(\boldsymbol{\gamma}_{lk}|\boldsymbol{\omega}_{lk}, \tau))(1/t_{lk} - 1/t_{rk})\}$.

S.2.3 $\mathbf{\Omega}$ update

For each chain l, $l = 1, \ldots, L$, we update the elements of the $q \times p$ selection probability matrix $\mathbf{\Omega}$ by using a Metropolis-within-Gibbs sampler with adaptive proposals. Let ω_{lk}^* and ρ_{lj}^* the proposed new values of the kth row effect and jth column effect in the lth chain respectively. The acceptance probability of the two parameters is

$$1 \wedge \left[\frac{(\omega_{lk}^*)^{p_{\gamma_{lk}}}(1-\omega_{lk}^*)^{p-p_{\gamma_{lk}}}}{\omega_{lkj}^{p_{\gamma_{lk}}}(1-\omega_{lkj})^{p-p_{\gamma_{lk}}}} \frac{Beta(\omega_{lk}^*; a_{\omega_k}, b_{\omega_k})|J(\lambda^{-1}(\omega_{lk}^*))|}{Beta(\omega_{lk}; a_{\omega_k}, b_{\omega_k})|J(\lambda^{-1}(\omega_{lk}))|} \right]^{1/t_{lk}} \frac{Q_{lk}(\lambda_{lk}^*, \lambda_{lk})}{Q_{lk}(\lambda_{lk}, \lambda_{lk}^*)} \tag{S.3}$$

and

$$1 \wedge \frac{Ga(\rho_{lj}^*; c_{\rho_j}, d_{\rho_j})|J(\varphi^{-1}(\rho_{lj}^*))|}{Ga(\rho_{lj}; c_{\rho_j}, d_{\rho_j})|J(\varphi^{-1}(\rho_{lj}))|} \prod_{k=1}^{q} \left[\frac{(\omega_{lkj}^*)^{\gamma_{lkj}}(1-\omega_{lkj}^*)^{1-\gamma_{lkj}}}{\omega_{lkj}^{\gamma_{lkj}}(1-\omega_{lkj})^{1-\gamma_{lkj}}} \right]^{1/t_{lk}} \frac{Q_{lj}(\varphi_{lj}^*, \varphi_{lj})}{Q_{lj}(\varphi_{lj}, \varphi_{lj}^*)}, \tag{S.4}$$

where in (S.3) $p_{\gamma_{lk}} = \boldsymbol{\gamma}_{lk}^T \mathbf{1}_p$, $\lambda_{lk} = \text{logit}(\omega_{lk})$, $J(\lambda^{-1}(\omega_{lk}))$ is the Jacobian of the inverse transformation evaluated in ω_{lk} and $Beta(\cdot)$ is the beta density function, while in (S.4) $J(\varphi^{-1}(\rho_{lj}))$ is the Jacobian of the inverse transformation evaluated in ρ_{lj}, $\omega_{lkj}^* = \omega_{lk} \times \rho_{lj}^*$, and $Ga(\cdot)$ is the gamma density function. As a technical point, since the prior density $p(\rho_{lj})$ cannot be indexed by k, in order to write the acceptance probability (S.4), in our model the prior for ρ_{lj} is not tempered.

We sample the proposed new values ω_{lk}^* and ρ_{lj}^* after suitable transformation from $Q_{lk}(\lambda_{lk}, \cdot) = \phi(\lambda_{lk}, s_{lk}^2(b))$ and $Q_{lj}(\varphi_{lj}, \cdot) = \phi(\varphi_{lj}, s_{lj}^2(b))$, respectively, where $s_{lk}(b)$ and $s_{lk}(b)$ are the adaptive proposals' standard deviations at batch b and $\phi(\cdot)$ is the normal density function. Following Roberts and Rosenthal (2009), *asymptotic convergence* is obtained enforcing the *diminishing adaptation condition* and imposing the *bounded convergence condition*. For the former condition, after the batch bth of 50 sweeps, say, the proposals' standard deviation are updated as follow: $s_{lk}(b+1) = s_{lk}(b) \pm \delta_s(b)$ and $s_{lj}(b+1) = s_{lj}(b) \pm \delta_s(b)$, where we add (subtract) to the current values $s_{lk}(b)$ and $s_{lj}(b)$ the quantity $\delta_s(b) = \min\{0.01, b^{-1/2}\}$ if the acceptance frequency of (S.3) and (S.4) are higher (lower) than the optimal acceptance rate (0.44), respectively. The latter condition is fulfilled assuming that $L_\lambda < s_{lk} < U_\lambda$ and $L_\varphi < s_{lj} < U_\varphi$ for some large positive (negative) values of U_λ and U_φ (L_λ and L_φ).

S.2.4 τ updates

The variable scaling coefficient is common to *all* the q regression equations and to *all* L chains. A new value τ^* is obtained using a Metropolis-with-Gibbs with acceptance probability

$$1 \wedge \frac{Ga(\tau^*; 1/2, n/2)|J(\psi^{-1}(\tau^*))| \prod_{l=1}^{L} \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau^*)^{1/t_{lk}}}{Ga(\tau; 1/2, n/2)|J(\psi^{-1}(\tau))| \prod_{l=1}^{L} \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_{lk}, \tau)^{1/t_{lk}}} \frac{Q(\psi^*, \psi)}{Q(\psi, \psi^*)},$$

where $\psi = \log(\tau)$, $J(\psi^{-1}(\tau))$ is the Jacobian of the inverse transformation evaluated in τ, $Ga(\cdot)$ is the gamma density function and $Q(\psi, \cdot) = \phi(\psi, 1)$. As in (S.4), the prior density is not tempered since we are sampling a common value across the q regressions and the L chains. The rational of this choice, for a given k, is illustrated in detail in Bottolo and Richardson (2010).

S.2.5 Temperature placement

During the burn-in, for each regression equation k, we automatically tune the temperature ladder in order to reach a specified acceptance rate of the exchange operator. In particular we chose as temperature ladder the geometric scale, such that the ratio of two consecutive temperatures is constant, $t_{(l+1)k}/t_{lk} = r_k$. Then after batch bth, say 100 sweeps, we update r_k as follows: $r_k(b+1) = r_k(b) \pm \delta_r$, where we add (subtract) to the current values $r_k(b)$ the quantity δ_r if the acceptance frequency of the exchange operator are higher (lower) than the optimal acceptance rate (0.50). For details on how to fix the value of δ_r interested reader can refer to Bottolo and Richardson (2010). For a discussion of different temperature scales, see Atchadé et al. (2010).

S.3 Post-processing

For a fixed k,

$$p(\boldsymbol{\gamma}_k^{(t)}|\boldsymbol{y}_k) = \frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_k^{(t)}, \tau^{(s)})p(\tau^{(s)}) \prod_{j=1}^{p} p(\gamma_{kj}^{(t)}|\omega_{kj}^{(s)})p(\rho_j^{(s)})$$

is the model posterior probability for the kth regression, where $\boldsymbol{\gamma}_k^{(t)} = (\gamma_{k1}^{(t)}, \ldots, \gamma_{kq}^{(t)})^T$ is latent binary vector recorded at the tth sweep of the algorithm, $p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_k^{(t)}, \tau^{(s)})$ is the marginal likelihood and $\tau^{(s)}$, $\omega_{kj}^{(s)} = \omega_k^{(s)} \times \rho_j^{(s)}$ and $\rho_j^{(s)}$ are the values of the parameters recorded at the sth sweep.

When the q regressions are jointly considered, the configuration posterior probability is defined as

$$p(\mathbf{\Gamma}^{(t)}|\boldsymbol{Y}) = \frac{1}{S}\sum_{s=1}^{S} p(\tau^{(s)}) \prod_{k=1}^{q} p(\boldsymbol{y}_k|\boldsymbol{X}, \boldsymbol{\gamma}_k^{(t)}, \tau^{(s)}) p(\omega_k^{(s)}) \prod_{j=1}^{p} p(\gamma_{kj}^{(t)}|\omega_{kj}^{(s)}) p(\rho_j^{(s)})$$

with $\mathbf{\Gamma}^{(t)}$ the configuration of the latent binary matrix at sweep tth.

Further literature cited

Atchadé, Y. F., G. O. Roberts and J. S. Rosenthal (2010). Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Stat. Comput.*. To appear.

Bottolo, L. and S. Richardson (2010). Evolutionary Stochastic Search for Bayesian model exploration. *Bayesian Analysis 5*, 583–618.

Kohn, R., M. Smith and D. Chan (2001). Nonparametric regression using linear combinations of basis functions. *Stat. Comput. 11*, 313–322.

Liang, F. and W. H. Wong (2000). Evolutionary Monte Carlo: application to C_p model sampling and change point problem. *Stat. Sinica 10*, 317–342.

Richardson, S., L. Bottolo and J. S. Rosenthal (2011). Bayesian models for sparse regression analysis of high dimensional data (with discussion). In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics: Proc. 9th Int. Meeting*. Oxford University Press.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *J. Comput. Graph. Stat. 9*, 349–367.