Submitted to the Annals of Applied Statistics

SUPPLEMENTARY MATERIAL FOR THE PAPER "INCORPORATING BIOLOGICAL INFORMATION INTO LINEAR MODELS: A BAYESIAN APPROACH TO THE SELECTION OF PATHWAYS AND GENES"

BY FRANCESCO C. STINGO[†], YIAN A. CHEN[‡], MAHLET G. TADESSE[§] AND MARINA VANNUCCI[†],

Rice University[†], Moffitt Cancer Center[‡] and Georgetown University[§]

MCMC scheme for sampling $(\boldsymbol{\theta}, \boldsymbol{\gamma})$. We now describe the MCMC steps for $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ in more detail. As previously described, no empty pathways or orphan genes are proposed during sampling and, for identifiability, selecting the same set of genes for different pathways is not allowed. At each iteration, only one pathway and/or a gene are proposed to be added or removed.

(1) Change inclusion status of both gene and pathway – randomly choose between addition (move 1.i) or removal (move 1.ii):

(1.i) Add a pathway and a gene:
First select a pathway that is not included in the model and has none of its member genes in the model ($\theta_k^o = 0$ and $p_{k\gamma}^o = 0$). Randomly choose one gene from the pathway ($\gamma_j^o = 0$) and propose including both the pathway and the gene, i.e., set $\theta_k^p = 1,\ \gamma_j^p = 1$. The move is accepted with probability

$$\min\left\{1, \frac{f(\boldsymbol{\theta}^p, \boldsymbol{\gamma}^p|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^o, \boldsymbol{\gamma}^o|\boldsymbol{T}, Y)} \cdot \frac{p_k \cdot \sum_{r=1}^K I\{\theta_r^o = 0, p_{r\gamma}^o = 0\}}{\sum_{r=1}^K I\{\theta_r^p = 1, p_{r\gamma}^p = 1, cond1, condId1\}}\right\}, \tag{1}$$

where *cond1* and *condId1* are explained in move type (1.ii) below.

(1.ii) Remove a pathway and a gene:
This move is the reverse of (1.i) described above. We first select a pathway that is included in the model and has only one of its member genes in the model ($\theta_k^o = 1$ and $p_{k\gamma}^o = 1$). In addition, this included gene ($\gamma_j^o = 1$) may not be the sole representative for other included pathways, to ensure that no empty pathway is created. Furthermore, identical sets of genes from different selected pathways cannot be created.

*Marina Vannucci is partially supported by NIH grant R01-HG0033190-05 and NSF grant DMS1007871.

These constraints correspond, respectively, to *cond1* and *condId1* in the proposal ratios (1) and (2). We attempt to remove both the pathway and the gene, i.e., set $\theta_k^p = 0,\ \gamma_j^p = 0$ and accept the move with probability

(2)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^p, \boldsymbol{\gamma}^p|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^o, \boldsymbol{\gamma}^o|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^K I\{\theta_r^o = 1, p_{r\gamma}^o = 1, cond1, condId1\}}{p_k \cdot \sum_{r=1}^K I\{\theta_r^p = 0, p_{r\gamma}^p = 0\}}\right\},$$

(2) Change the inclusion status of gene but not pathway – randomly choose between addition (2.i) or removal (2.ii):

(2.i) Add a gene in an already included pathway:
First select a pathway already included in the model and that has some member genes that could potentially be added ($\theta_k^o = 1$ and $p_k > p_{k\gamma}^o$). Let G be the set of pathways that satisfy these conditions. Choose one of the non-included genes from this pathway ($\gamma_j^o = 0$) and attempt to add it, i.e, set $\theta_k^p = \theta_k^o = 1,\ \gamma^p = 1$. The proposal is accepted with probability

(3)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^p, \boldsymbol{\gamma}^p|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^o, \boldsymbol{\gamma}^o|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^K I\{\theta_r^o = 1, p_r > p_{r\gamma}^o\} \cdot \sum_{r\in G} \frac{1}{p_{r\gamma}^{p(cond2\gamma, condId2\gamma)}}}{\sum_{r=1}^K I\{\theta_r^p = 1, p_{r\gamma}^p > 1, cond2_\theta, condId2_\theta\} \cdot \sum_{r\in G} \frac{1}{p_r - p_{r\gamma}^o}}\right\},$$

where '$cond2_\theta$', '$cond2_\gamma$', '$condId2_\theta$' and '$condId2_\gamma$' are explained in move type (2.ii) below.

(2.ii) Remove a gene from an already included pathway:
This move is the reverse of (2.i) described above. We first select a pathway already included in the model and that has more than one of its member genes included in the model ($\theta_k^o = 1, p_{k\gamma}^o > 1$). In addition, at least one of the included genes from this pathway may not be the sole representative for other included pathways and its removal would not create an identifiability problem – this corresponds to constraints '$cond2_\theta$' and '$condId2_\theta$'in the proposal ratios of (3) and (4). Once the pathway is selected, choose a gene among the eligible candidates, that is, an included member gene ($\gamma_j^o = 1$) which is not the sole representative for other included pathways and whose removal does not create an identifiability problem – this corresponds to constraints '$cond2_\gamma$' and '$condId2_\gamma$'. Constraints '$cond2_\theta$' for pathways, and '$cond2_\gamma$' for genes, will ensure that no empty pathways are created after the proposed move. Leave the pathway status unchanged and attempt to re-

move the selected gene, i.e., set $\theta_k^p = \theta_k^o = 1,\ \gamma_j^p = 0$. The proposed move is accepted with probability

(4)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^p, \boldsymbol{\gamma}^p|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^o, \boldsymbol{\gamma}^o|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^K I\{\theta_r^o = 1, p_{r\gamma}^o > 1, cond2_\theta, condId2_\theta\} \cdot \sum_{r\in G} \frac{1}{p_r - p_{r\gamma}^p}}{\sum_{r=1}^K I\{\theta_r^p = 1, p_r > p_{r\gamma}^p\} \cdot \sum_{r\in G} \frac{1}{o(cond2_\gamma, condId2_\gamma)}}\right\}.$$

(3) Change inclusion status of pathway but not gene – randomly choose between addition (3.i) or removal (3.ii):

(3.i) Add a pathway but leave genes' status unchanged:
First select a pathway that is not included in the model but has some of its member genes included in the model through other pathways ($\theta_k^o = 0$ and $p_{k\gamma}^o \geq 1$). Attempt to add the pathway but leave the status of its member genes unchanged, i.e., set $\theta_k^p = 1$. The proposed move is accepted with probability

(5)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^p, \boldsymbol{\gamma}^p|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^o, \boldsymbol{\gamma}^o|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^K I\{\theta_r^o = 0, p_{r\gamma}^o \geq 1, condId3\}}{\sum_{r=1}^K I\{\theta_r^p = 1, p_{r\gamma}^p \geq 1, cond3\}}\right\},$$

where *condId3* means that it is not possible to select a pathway whose selected genes form the entire set of selected genes for another selected pathway, and *cond3* is explained in move type (3.ii) below.

(3.ii) Remove a pathway but leave genes' status unchanged:
This move is the reverse of (3.i) described above. First select a pathway included in the model that has all of its $p_{k\gamma}^o$ included member genes associated with other included pathways ($\theta_k^o = 1$ and 'cond3'). This will ensure that no orphan gene is created. Attempt to remove the pathway but leave the status of the genes unchanged, i.e., set $\theta_k^p = 0$ and accept the move with probability

(6)
$$\min\left\{1, \frac{f(\boldsymbol{\theta}^p, \boldsymbol{\gamma}^p|\boldsymbol{T}, Y)}{f(\boldsymbol{\theta}^o, \boldsymbol{\gamma}^o|\boldsymbol{T}, Y)} \cdot \frac{\sum_{r=1}^K I\{\theta_r^o = 1, p_{r\gamma}^o \geq 1, cond3\}}{\sum_{r=1}^K I\{\theta_r^p = 0, p_{r\gamma}^p \geq 1, condId3\}}\right\}.$$

It is easy to see that our Bayesian stochastic search variable selection kernel generates an ergodic Markov chain over the restricted space. First note that the chain produced by our MCMC has the following properties:

- It is aperiodic and has an invariant probability distribution (by definition of the M-H kernel);
- It is irreducible (noting that every move is equipped with its reverse, that it is possible to reach any valid configuration in the parameter space starting from the configuration where no pathways and no genes are selected, and that the probability of moving will never be zero, i.e., the probability of moving from one point to another in n steps is bigger then zero);
- Properties above imply that the chain is recurrent and therefore ergodic.

Let $\text{supp}\, q(\cdot|x)$ indicate the support of our proposal distribution, i.e., the set of possible configurations that can be generated from the previously visited configuration x. We say that the supports for two different x's are connected if they share at least one configuration. We need to check that the union of all connected supports is equal to the entire support of the posterior distribution, i.e,

$$\bigcup_{x \in \text{supp } f} \text{supp}\, q(\cdot|x) \supset \text{supp } f$$

where f is the target density and $q(\cdot|x)$ the proposal distribution (Robert & Casella 2004). It is easy to verify that our MCMC satisfies this condition because, starting from the configuration where no pathways or genes are selected, it is possible to reach every admissible configuration, and because the union of these points is exactly equal to the support of the target distribution.

References.

Robert, C. & Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer Verlag.

MOFFITT CANCER CENTER
TAMPA, FL 33612, USA.
E-MAIL: Ann.Chen@moffitt.org

DEPARTMENT OF STATISTICS
RICE UNIVERSITY
HOUSTON, TX 77005, USA.
E-MAIL: marina@rice.edu

DEPARTMENT OF MATHEMATICS & STATISTICS
GEORGETOWN UNIVERSITY
WASHINGTON, DC 20057, USA.
E-MAIL: mgt26@georgetown.edu