Transdimensional Markov Chain Monte Carlo Using Hyperplane Inflation in Locally Nested Spaces

Giovanni Petris
University of Arkansas

Luca Tardella
Università di Roma "La Sapienza"

January 18, 2006

Giovanni Petris is Associate Professor, Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701 (email: GPetris@uark.edu). Luca Tardella is Associate Professor, Department of Statistics, Probability and Applied Statistics, University of Rome "*La Sapienza*", Rome, Italy (email: luca.tardella@uniroma1.it).

Abstract

We present some new results that extend the geometric approach to trans-dimensional Markov chain Monte Carlo simulations originally proposed in Petris and Tardella (2003a). These provide a black-box method to generate a sample from a Markov chain with a prescribed stationary distribution on a disjoint union of Euclidean spaces not necessarily of the same dimension. The only requirement is that the support spaces of different dimensions have to be locally nested and the corresponding densities of the target distribution have to be known up to a normalizing constant. Empirical evidence of effectiveness of the proposed method is provided by a controlled experiment of variable selection in a general regression context as well as by an original approach to mixture of normal models.

Key Words: Bayesian inference; Model selection; Model averaging; Variable selection in regression models; Mixture models.

1 Introduction

It is by now widely acknowledged in many applied fields that uncertainty about the fitted model has to be accounted for in any statistical data analysis. The books by Burnham and Anderson (2002) and Koop (2003) provide two examples, in biology and econometrics, of the increasing interest in multimodel inference. The first is entirely devoted to the issues of model selection and model averaging, while the latter has a full chapter on model averaging. Within the Bayesian approach, model uncertainty can be quantified through posterior model probabilities which can be used both for model/variable selection and for improving predictions of future observations by weighing and combining model-specific predictions. Averaging predictions from different models has been shown in many practical cases (Chatfield, 1995) to provide predictions with better statistical properties. Given the importance of Bayesian multimodel inference, it is very important to devise simple techniques and easy-to-use algorithms to obtain posterior inferences in a multimodel setting. The present paper is intended to be a contribution in this direction.

Reversible Jump (RJ) (Green, 1995), which currently seems to be the transdimensional sampler most often used in applied work, although very flexible, often requires careful specification of jump proposals. A few methods to assist the user in finding efficient jump proposals have recently become available (see for example Brooks, Giudici and Roberts, 2003; Ehlers and Brooks, 2003; Green, 2003), but they are still of limited use. In the paper we suggest a way of reducing the problem of transdimensional sampling to one of ordinary, fixed-dimensional, Markov chain Monte Carlo (MCMC), which is in principle simpler and for which automatic methods exist. Compared with the similar-in-spirit product-space approach proposed in Carlin and Chib (1995), our method is more efficient in terms of keeping as low as possible the dimensions of the simulation spaces and does not require the specification of well-tuned pseudo-priors. For a current extensive review of transdimensional Markov chain methods, see Sisson (2005).

Based on a simple geometric intuition Petris and Tardella (2003a) introduced a new method to facilitate sampling from a distribution whose support is comprised of a set of nested hyperplanes. This is the case, for example, of the posterior distribution of the parameters of a polynomial regression model $E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p$, when the maximum order is

set to p but there is a positive probability that the true regression is in fact polynomial of order k, strictly less than p. In the simplest setting of a distribution supported just on two hyperplanes – the full Euclidean space and one proper hyperplane – the basic tool developed by the authors is a simple transformation of the original distribution viewed as a mixture of two mutually singular components into a single absolutely continuous distribution supported on the largest space. In the general case of several nested hyperplanes, the same goal, constructing a continuous distribution which is in some sense equivalent to the original one, is achieved through an iterative use of the basic tool for two supporting hyperplanes. One of the advantages is that in general understanding and approximating a sample from an absolutely continuous distribution in a (Markov Chain) Monte Carlo technique is much simpler than understanding and approximating the original distribution, supported by nested hyperplanes. In fact the new approach has been already successfully tested in many frequently used nested settings such as regenerative MCMC (Petris and Tardella, 2003a; Petris and Tardella, 2003b), model selection/ averaging for nested linear models (Petris and Tardella, 2003a), autoregressive models (Petris and Tardella, 2001). With the exception of regenerative MCMC, the target distribution is always a posterior distribution supported by a set of nested models, as in the polynomial regression example mentioned at the beginning of this section. In fact in all the abovementioned applications the target distribution is supported by a set of nested hyperplanes and although the class of nested models is large and contains many useful models, being able to treat the nested case only is definitely a limitation of the approach. In the present paper we show how the geometric approach can be exploited to deal also with the much more general transdimensional MCMC case when the different models are only locally nested. (A precise definition of locally nested is given in Section 3.)

In Section 2 after briefly reviewing the results of Petris and Tardella (2003a) we introduce some extensions that will be needed in Section 3, where we prove the main theoretical result needed to apply the new approach to transdimensional MCMC in the general locally nested case. An application to variable selection for linear regression models is included in Section 4 Section 5 contains a practical example where we show how to apply the proposed approach to a Bayesian analysis of a mixture model with an unknown number of components. Section 6 concludes the paper.

2 Hyperplane Inflation: the nested case

We briefly recall the main results of Petris and Tardella (2003a). The basic idea is more easily grasped when the target distribution is supported by only two hyperplanes, both improper. Let η_k and δ_k denote two measures in $\mathbb{R}^k$ corresponding to the Lebesgue measure and the Dirac measure concentrated at the origin and consider the target distribution (possibly not normalized)

$$\mu(dx) = f_0(x)\eta_k(dx) + f_k\delta_k(dx)$$

which is a mixture distribution of two components, absolutely continuous the former and degenerate on the origin the latter. In order to transform μ into a single absolutely continuous distribution, one can move the density $f_0(x)$ of the continuous component away from the origin using the one-to-one transformation

$$x \longmapsto x\,(\|x\|^k + r^k)^{1/k}\,\|x\|^{-1}, \tag{1}$$

where r is a parameter to be appropriately set. With this transformation there is no density defined in a region corresponding to ball of radius r around the origin. Hence one can obtain an auxiliary density by extending the definition of the transformed density with a constant density over that ball, more precisely, spreading uniformly the mass f_k on the ball. If z is a draw from the continuous distribution corresponding to this auxiliary density, it is easy to transform z into a draw x from the original target distribution, essentially by undoing (1). In fact, if $\|z\| > r$, then the inverse of (1) is

$$z \longmapsto x = z\,(\|z\|^k - r^k)^{1/k}\,\|z\|^{-1}; \tag{2}$$

while, if $\|z\| < r$, the corresponding x in the original uninflated space is the origin. The boundary region $\{z : \|z\| = r\}$ is clearly irrelevant since it has probability zero.

The map (1) suggested the name Hyperplane Inflation (HI), since it simply *inflates* the origin into a ball. The fact that (1) preserves Lebesgue measure avoids the need for computing complicated Jacobians.

The case of distribution with two components, one of which supported by a proper hyperplane, can be treated similarly: essentially by applying the inflation and deflation maps to some of the coordinates only, in an appropriate coordinate system. To obtain the general case with $k > 2$ nested hyperplanes one can iteratively combine the continuous component of the

distribution with a degenerate one, transforming the two into a new continuous component until there are no degenerate components left (for details, see Petris and Tardella, 2003a). All the described procedures can be carried out in an automatic way starting from the (possibly not normalized) densities of all the degenerate components. The R package HI, developed by the authors and available from CRAN[1], provides functions that, given the original degenerate densities, evaluate the density of the auxiliary absolutely continuous distribution and the function that transform a draw from it into a draw from the original distribution.

We now give a result that generalizes Theorem 1 and Theorem 2 of Petris and Tardella (2003a) so that one can realize a transdimensional Markov chain sampler working for general locally nested subspaces by using the basic idea of reformulating a two component mixture distribution into a single distribution.

To formalize the setting described above, we have a target distribution μ on a space $(X, \mathcal{S}_X)$ (representing the original mixture defined in terms of μ), a distribution τ on another space $(Z, \mathcal{S}_Z)$ (corresponding to the continuous distribution on the inflated space), and a measurable map $\phi\colon Z \to X$ such that $\mu = \tau\phi^{-1}$. The running assumption is that drawing a sample from τ, or, more generally, from a Markov chain for which τ is the limiting distribution, is easy, while drawing a sample directly from μ is not. If one wants to apply the HI approach just described within a Gibbs sampler type of simulation, where the whole parameter space is a product of $(X, \mathcal{S}_X)$ with some other subspace(s) then μ represents in fact a full conditional, hence μ (as well as τ and ϕ) changes at any iteration. In this case, one has to map x_0 (part of the current state of the chain) to a point z_0 in Z, draw a z_1 from an appropriate Markov transition kernel, and map back z_1 to $x_1 = \phi(z_1)$. Theorem 2 in Petris and Tardella (2003a) shows that if τ is invariant for the transition kernel used to sample z_1 then, subject to an intuitive condition on the mapping $x_0 \mapsto z_0$, μ is invariant for the resulting transition kernel in X. We prove below a similar result stated in terms of reversibility, instead of invariance. Remember that, by definition, if μ is a probability and H is a transition kernel, H and μ are in detailed balance if, for any two measurable sets A, B,

$$\int_{A\times B} \mu(dx)H(x, dy) = \int_{B\times A} \mu(dx)H(x, dy).$$

[1]http://cran.r-project.org/mirrors.html

Theorem 1. *Let μ and τ be probabilities on $(X, \mathcal{S}_X)$ and $(Z, \mathcal{S}_Z)$, respectively, and let $\phi\colon Z \to X$ be a measurable function such that $\mu = \tau\phi^{-1}$ (i.e., ϕ is a random element of X defined on Z, whose distribution is μ). Suppose that K is a transition kernel on $(Z, \mathcal{S}_Z)$ which is in detailed balance with τ, and J is a transition kernel from X to Z such that, for every $B \in \mathcal{S}_Z$, $J(x, B)$ is a version of $\tau(B|\phi = x)$. Define a transition kernel H on X by setting*

$$H(x, A) = \int_Z J(x, dz)K(z, \phi^{-1}A), \qquad x \in X, A \in \mathcal{S}_X.$$

Then H and μ are in detailed balance.

Proof. By definition of conditional probability, for any $A \in \mathcal{S}_X$ and $E \in \mathcal{S}_Z$,

$$\int_A \mu(dx)J(x, E) = \tau(\phi^{-1}A \cap E).$$

It follows, using standard arguments, that for any bounded measurable $f\colon Z \to \mathbb{R}$ and for any $A \in \mathcal{S}_X$,

$$\int_A \mu(dx) \int_Z J(x, dz)f(z) = \int_{\phi^{-1}A} \tau(dz)f(z). \tag{3}$$

Consider now sets $A, B \in \mathcal{S}_X$. One has,

$$\begin{aligned}
\int_{A\times B} \mu(dx)H(x, dy) &= \int_A \mu(dx)H(x, B) \\
&= \int_A \mu(dx) \int_Z J(x, dz)K(z, \phi^{-1}B) \\
&= \int_{\phi^{-1}A} \tau(dz)K(z, \phi^{-1}B) \qquad \text{(by (3))} \\
&= \int_{\phi^{-1}B} \tau(dz)K(z, \phi^{-1}A) \qquad \text{(by reversibility of } K) \\
&= \int_B \mu(dx) \int_Z J(x, dz)K(z, \phi^{-1}A) \qquad \text{(by (3))} \\
&= \int_B \mu(dx)H(x, A) = \int_{B\times A} \mu(dx)H(x, dy),
\end{aligned}$$

which shows that H and μ are in detailed balance. □

To clarify the content of the theorem in the context of HI when the target μ has a continuous component and a point mass we point out that the kernel J is such that if $x \neq 0$ then it just maps x back to $z = \phi^{-1}(x)$ deterministically (ϕ is in this case given by (2) and is one-to-one), while if $x = 0$ it draws a point z at random in the ball of radius r. Furthermore, the above theorem shows that, if one uses ARMS along a randomly selected line through the current point as the transition kernel K to sample from the continuous distribution τ (as advocated in Petris and Tardella, 2003a), then the resulting sampler is reversible with respect to μ. This is a stronger conclusion than the one of Theorem 2 in Petris and Tardella (2003a), where it is only shown that μ is the stationary distribution of the sampler. Similar considerations hold when the HI approach is applied to a target with more than two nested components.

3 Hyperplane Inflation: locally nested models

In transdimensional MCMC the goal is to generate a Markov chain having a prescribed limiting distribution on a space $X = \cup_k X_k$, which is a disjoint countable union of spaces usually identified as models in a Bayesian setting. In most applications the models X_k are *locally nested*, in the sense that for any distinct X_j and X_k there is a sequence $i_0, \ldots, i_n$ with $i_0 = j$, $i_n = k$, such that either $X_{i_{s-1}} \subset X_{i_s}$ or $X_{i_{s-1}} \supset X_{i_s}$, for $s = 1, \ldots, n$. A typical example is provided by the context of variable selection for a linear regression function where p covariates are available. Any two distinct subsets of covariates locally nested in the sense they are both nested in at least the full model with all covariates.

Having built an all-purpose sampler which is in detailed balance on pairs of nested models, we can now try to use it locally in this more general setting. For example, from the current state $x \in X_j$ of the chain, one can select at random another model X_k so that the two models X_j and X_k are nested, and use a Markov kernel H on $X_j \cup X_k$ as described in the previous section, to move on the subset $X_j \cup X_k$ of the entire space X. The condition of having a locally nested family of models ensures that the resulting sampler can be irreducible, i.e. any point in X can be reached from any other point in a finite number of transitions. Most RJ samplers use the same philosophy of

moving between nested models when proposing the addition or removal of one "dimension" of the model. This extra dimension may be a component for mixture models, a change point for change point models, an explanatory variable for regression models, and so on. In RJ one has to design a tailored joint proposal distribution for the move type and the proposal state, and detailed balance with the target distribution is achieved through the introduction of an accept/reject step involving both the move type selection mechanism and the mechanism used to generate the proposed new state of the chain. Following the approach proposed here, on the other hand, it is enough to consider an acceptance probability depending on the move type only, since HI provides a black-box kernel already in equilibrium with the target distribution.

We now formalize our approach, starting with some definitions. Let $(V, \mathcal{S}_V)$ be a measurable space indexing the set of available *move types*; for example, HI in conjunction with ARMS on $X_j \cup X_k$ may be a move type (note that this can be considered a transition kernel on the entire X by defining it to coincide with the identity kernel $(x, A) \mapsto I_A(x)$ outside $X_j \cup X_k$). Let ν be an integral transition kernel from X to V, with density f with respect to a base measure β: when the chain is at x, a move type $v \in E \in \mathcal{S}_V$ is selected with probability $\nu(x, E) = \int_E f(x, v)\, \beta(dv)$. In practice V is always finite and a convenient choice for $\nu(x, \cdot)$ is the uniform distribution on the set of move types available from x. Finally, for each $v \in V$, let K_v be a transition kernel on X, which is in detailed balance with the target distribution μ. We assume that for every $A \in \mathcal{S}_X$ the function $K_v(x, A)$ is measurable in (v, x). Define the acceptance probability

$$\alpha_v(x, y) = \frac{f(y, v)}{f(x, v)} \wedge 1, \qquad v \in V,\ x, y \in X,$$

and the transition kernel on X

$$H(x, A) = \int_V \beta(dv) f(x, v) \Big\{ \int_A K_v(x, dy) \alpha_v(x, y) + \delta_x(A) \int_X K_v(x, dy) \big(1 - \alpha_v(x, y)\big) \Big\}, \qquad x \in X, ; A \in \mathcal{S}_X.$$

In summary, the kernel H can be roughly described by the following steps:

1. From the current point $x \in X$, select a move type v according to $\nu(x, \cdot)$.

2. Draw a proposal $y \in X$ from the distribution $K_v(x, \cdot)$.

3. Accept y with probability $\alpha_v(x, y)$, otherwise stay at x.

Theorem 2. *With the previous definitions, H and μ are in detailed balance.*

Proof. By assumption, the measure $\mu(dx)K_v(x, dy)$ is symmetric for each v. Moreover, for every v, the function $f(x, v)\alpha_v(x, y) = f(x, v) \wedge f(y, v)$ is symmetric in x and y. Therefore the measure (on $X \times X$) $\mu(dx)K_v(x, dy)f(x, v) \wedge f(y, v)$ is symmetric as well. Consider now two sets $A, B \in \mathcal{S}_X$. We only need to show detailed balance for the substochastic kernel given by the first term in the definition of H above, i.e. the kernel arising when a proposal is accepted.

$$\begin{aligned}
&\int_A \mu(dx) \int_V \beta(dv) f(x, v) \int_B K_v(x, dy)\alpha_v(x, y) \\
&\quad = \int_V \beta(dv) \int_A \mu(dx) \int_B K_v(x, dy) f(x, v)\alpha_v(x, y) \\
&\quad = \int_V \beta(dv) \int_{A\times B} \mu(dx)K_v(x, dy)\big(f(x, v) \wedge f(y, v)\big) \\
&\quad = \int_V \beta(dv) \int_{B\times A} \mu(dx)K_v(x, dy)\big(f(x, v) \wedge f(y, v)\big) \\
&\quad = \int_B \mu(dx) \int_V \beta(dv) f(x, v) \int_A K_v(x, dy)\alpha_v(x, y)
\end{aligned}$$

□

Since we have assumed that the kernels K_v are all in detailed balance with μ, one may wonder whether the randomized acceptance step is really needed or can be avoided altogether. The following example, in a very simple setting, shows that the move type selection process may destroy the reversibility property enjoyed by all the K_v.

Example. Consider $X = \{1, 2, 3, 4\}$, $V = \{0, 1\}$ and suppose the target distribution μ is uniform on X. Let K_v be defined by the transition matrices:

$$K_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad K_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Note that both K_0 and K_1 are symmetric, which implies that they are in detailed balance with the uniform distribution. Let p be a number in $(1/2, 1)$, and let ν be the transition kernel from X to V defined by

$$\nu = \begin{bmatrix} 1-p & p \\ 1-p & p \\ p & 1-p \\ p & 1-p \end{bmatrix}.$$

The transition kernel on X obtained by first selecting v according to ν and then moving according to the dynamics determined by K_v is

$$K(x, y) = \int_V \nu(x, dv) K_v(x, y).$$

Simple arithmetics shows that this corresponds to the transition matrix

$$K = \begin{bmatrix} 0 & 1-p & p & 0 \\ 1-p & 0 & 0 & p \\ 1-p & 0 & 0 & p \\ 0 & 1-p & p & 0 \end{bmatrix}.$$

Since K is not symmetric, the resulting Markov kernel is not in equilibrium with the uniform distribution. However, consider using K_v as a proposal kernel only, introducing the acceptance step described above: if, from the current x, move type v is selected and y is drawn from $K_v(x, \cdot)$, the chain moves to y with probability $\alpha_v(x, y) = 1 \wedge \big(\nu(y, v)/\nu(x, v)\big)$. Straightforward calculations show that in this case the resulting kernel on X has transition matrix

$$\begin{bmatrix} 2p-1 & 1-p & 1-p & 0 \\ 1-p & 2p-1 & 0 & 1-p \\ 1-p & 0 & 0 & p \\ 0 & 1-p & p & 0 \end{bmatrix},$$

which is symmetric and hence in equilibrium with the target uniform distribution. □

4 Application to model selection

We illustrate the effectiveness of the advances in exploring distributions supported on subspaces which are not nested but only locally nested through a

classical example of variable selection in a regression context. A controlled experiment can be performed as follows. A data set is produced by fixing a number of observations, n, and fixing corresponding values of a certain number of covariates $Z_i = (Z_{i1}, ..., Z_{ip})$ for $i = 1, ..., n$. Given those covariates one generates

$$Y_i \sim N(Z_i\beta, \sigma^2 I)$$

where $\beta = (\beta_1, ..., \beta_p)$ is a known fixed vector of regression coefficients with possibly some null component so that only some of the covariates are effective in explaining the data. When making inference in this context the main parameter space of interest to consider is the locally nested space of regression coefficients that can be written as

$$X = \bigcup_{\lambda \in \{0,1\}^p} X_\lambda$$

where $\lambda = (\lambda_1, ..., \lambda_r, ..., \lambda_p)$ is a binary label indexing all 2^p possible submodels which take into consideration the presence of only a subset of non-null regressors so that the corresponding labelled subspace is

$$X_\lambda = \{(\beta_1, ..., \beta_r, ..., \beta_p) \in \mathbb{R}^p \mid \beta_r = 0 \text{ if } \lambda_r = 0\}.$$

In fact, one is interested in exploring the full parameter space $X \times (0, \infty)$ for both regression parameters β and σ^2. We adopt the following conjugate prior for the unknown parameters:

$$\lambda \sim \text{Unif}(\{0,1\}^p),$$

$$\beta_r \mid \lambda \sim \mathcal{N}(0, \rho\sigma^2), \quad \text{if } \lambda_r = 1,$$

$$\sigma^{-2} \sim \mathcal{G}\left(\frac{\xi}{2}, \frac{\psi}{2}\right).$$

(Here the Gamma distribution $\mathcal{G}(a, b)$ has mean a/b). With this prior it is possible to compute the exact values of posterior probabilities of all 2^p submodels and hence compare the estimated probabilities based on our proposed transdimensional MCMC sampler.

The only transdimensional component of the implemented MCMC scheme is performed through sampling with the aid of the HI procedure combining two competing models $X_j = X_{\lambda_{(1)}}$ and $X_k = X_{\lambda_{(2)}}$ which differs only for one

dimension, i.e. for the presence/absence of a regressors. Hence we considered moves which combine two subsets of regressors which are the same except for the presence/absence of one regressor

$$V = \left\{ v = (\lambda_{(1)}, \lambda_{(2)}) : \sum_{r=1}^{p} \left| \lambda_{(1)r} - \lambda_{(2)r} \right| = 1 \right\}.$$

We have designed a very simple move type distribution $f(\beta, v)$, with $v \in V$, that at each iteration proposes sequentially

1. to add or drop with equal probability one regressor (i.e. leaving it free to be non-null) with the obvious exceptions when β is such that there is no regressor to add or drop (all non-null regressors already in β or just one non-null regressor in β);

2. randomly choosing which one among all available non-null regressors is the one to add or drop.

In order to keep the chain in detailed balance we have to adjust for move type unbalance. In fact, if β_{prop} contains say $k+1$ non-null regressors and β_{curr} contains say k non-null regressors $f(\beta_{cur}, v)$ corresponds to the probability of adding one non-null component move while $f(\beta_{prop}, v)$ corresponds to the probability of dropping one non-null component. Of course one has to realize that the two probabilities are not the same, since when adding from a model with k regressors the one which is in β_{prop} this is drawn with probability $(p-k)^{-1}$ while when dropping from a set of $k+1$ non-null regressors the one corresponding to β_{curr} this happens with probability $(k+1)^{-1}$. Obvious adjustments have to be made in case either β_{curr} (or β_{prop}) contains no regressor or all regressors. Also, in order to get a faster mixing and better accuracy, instead of the original densities on the corresponding submodel space we used a scaled-shifted transformation exploiting simple least-square estimates. All the remaining simulation tasks in the inflated subspaces or in the σ^2 space have been performed through the ARMS routine, after writing the appropriate un-normalized full-conditional density to draw from. In Figure 1 one can see the traces of the estimated probabilities of the two most probable models, while in Figure 2 we display the discrepancies between simulated and exact values of posterior probabilities for the most probable ten submodels when $p = 10$, $\beta = (6, 5, 4, 3, 2, 1, 0, 0, 0, 0, 0)'$, $\rho = 100$, $\xi = \psi = 0.01$.

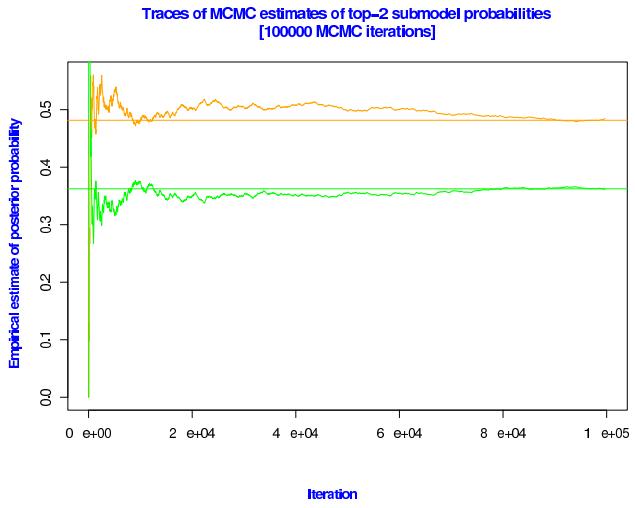

Figure 1: Traces of the MCMC estimated probabilities of the two most probable submodels

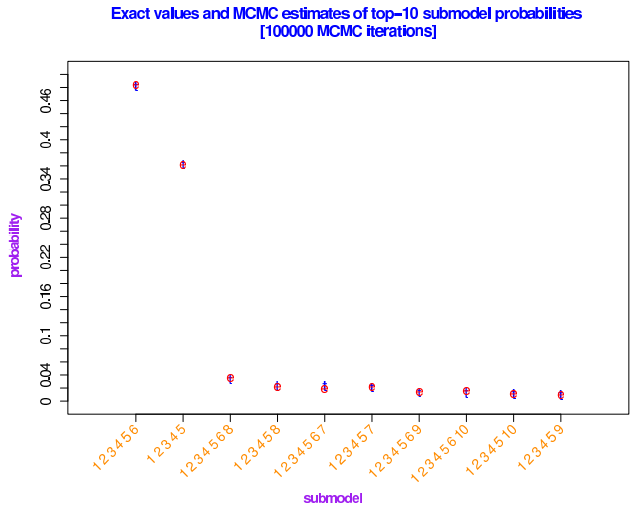

Figure 2: Comparison between true exact posterior probabilities (t) and MCMC estimated probabilities (e) of the most probable ten submodels

It can be seen from Figure 2 that the MCMC estimates of the posterior probabilities of the different models are very close to the true posterior probabilities.

5 Application to mixture models

We have also applied the approach described in the previous section to the Bayesian analysis of mixture models. One of the peculiar and intriguing feature of this application is that our method proved itself effective even in a context where the different dimensional spaces are not usual Euclidean spaces like $\mathbb{R}^k$ but consists of simplexes. More specifically, we consider a mixture of an unknown number of normal distributions. Let $K+1$, for a fixed K, be the maximum number of components that one is willing to allow in the mixture, and let $\varphi(y;\mu,\sigma^2)$ denote the Gaussian density with mean μ and variance σ^2. The model assumes that we have independent observations from the density

$$f(y) = \sum_{i=0}^{K} \pi_i \varphi(y;\mu_i,\sigma_i^2).$$

For the unknown means and variances of the individual components we assume the following prior distribution.

$$\begin{aligned}
\mu_0,\ldots,\mu_K &\overset{\text{iid}}{\sim} \mathcal{N}(\mu,\sigma^2),\\
\mu &\sim \mathcal{N}(\bar{\mu},\rho\sigma^2),\\
\sigma^{-2} &\sim \mathcal{G}(\alpha_\mu,\beta_\mu),\\
p(\sigma_i^2) &\propto \left(\sigma_i^2\right)^{(\epsilon-1)/2} I_{(\psi,\psi+\Delta)}(\sigma_i^2), \qquad i=0,\ldots,K \quad \text{(independent)},\\
\psi &\sim \mathcal{G}(\alpha_\psi,\beta_\psi),\\
\Delta &\sim \text{Pareto}(\alpha_\Delta,\beta_\Delta).
\end{aligned}$$

(Here Pareto(α,β) has density $\beta\alpha^\beta x^{-\beta-1} I_{(\alpha,+\infty)}(x)$.) For the weights $\pi_0,\ldots,\pi_K$ we adopt the following prior. First of all, in order to enforce identifiability, we assume that $\pi_0 \geq \pi_1 \geq \cdots \geq \pi_K$. Second, we want to give a positive prior probability to each of the regions

$$S_k^0 = \left\{(\pi_0,\ldots,\pi_K) : \sum \pi_i = 1, \pi_0 \geq \cdots \geq \pi_K, \pi_k > 0, \pi_{k+1} = 0\right\},$$

for $k = 0, \ldots, K$, so we set a priori $P(S_k^0) = w_k$ and, given S_k^0, $(\pi_0, \ldots, \pi_K) \sim \text{Unif}(S_k^0)$ $(k = 0, \ldots, K)$. For convenience we follow the standard procedure to introduce for each observation an indicator variable z so that

$$\begin{aligned} f(y|z) &= \varphi(y; \mu_z, \sigma_z^2), \\ P(z = i) &= \pi_i \qquad i = 0, \ldots, K. \end{aligned}$$

Clearly, integrating out the indicator variable z, one gets back the original mixture model for the observation y. Assuming that we have a sample of size n, we will denote the observations by $y_1, \ldots, y_n$ and the corresponding indicator variables by $z_1, \ldots, z_n$. Moreover, we will denote by J_i^z the set of indices j such that $z_j = i$ and by n_i the cardinality of J_i^z. If we consider the whole set of parameters subdivided into the following blocks, $(z_1, \ldots, z_n)$, $(\mu_0, \ldots, \mu_K)$, $(\sigma_0^2, \ldots, \sigma_K^2)$, μ, σ^2, ψ, Δ and $(\pi_0, \ldots, \pi_K)$, we can think of building up a Gibbs sampling scheme such that drawing from the full conditional distribution of any block but the last one is straightforward. So we focus on how to use HI to draw from the full conditional distribution of $(\pi_1, \ldots, \pi_K)$ over the space $X = \cup_{k=1}^K S_k$, where S_k denotes the projection of S_k^0 onto the last $K - 1$ coordinates, i.e.

$$S_k = \big\{(\pi_1, \ldots, \pi_K) : \textstyle\sum \pi_i \leq 1,\ 1 - \sum \pi_i \geq \pi_1 \geq \cdots \geq \pi_K,\ \pi_k > 0,\ \pi_{k+1} = 0\big\}.$$

(In the notation of Section 3, $X_k = S_k$.) In fact, the prior density on the mixture weight parameters $\pi = (\pi_1, \ldots, \pi_K)$ can be viewed as a density

$$g_{\text{prior}}(\pi) = \sum_{k=0}^{K} w_k g_k(\pi_1, \ldots, \pi_k) I_{S_k}(\pi),$$

with respect to the measure γ defined by

$$d\gamma = \sum_{k=0}^{K} d\eta_k d\delta_{K-k}$$

on the largest K-dimensional simplex

$$\big\{(\pi_1, \ldots, \pi_K) : \pi_i \geq 0, i = 1, \ldots, K;\ \textstyle\sum \pi_i \leq 1\big\}$$

where $g_0 = 1$ and $g_k(\pi_1, \ldots, \pi_k) = k!(k+1)!$ for $k = 1, \ldots, K$. A direct application of Bayes theorem shows that the full conditional distribution of

π has density, w.r.t. γ,

$$g(\pi) \propto \sum_{k=0}^{K} w_k g_k(\pi_1, \ldots, \pi_k) I_{S_k}(\pi) \prod_{h=0}^{K} \pi_h^{n_h},$$

with $\pi_0 = 1 - \sum_{i=1\ldots K} \pi_i$ and with the convention $0^0 = 1$. Hence we are in the same context of Section 3, so we just need to introduce a move type distribution $\nu(\pi, \cdot)$ over

$$V = \{v = (i, j) : 0 \leq i < j \leq K, j - i = 1\},$$

and, once the move v has been selected, draw from the auxiliary density evaluated by the HI package. The distribution $\nu(\pi, \cdot)$ selecting the move type works as follows: if the current value of π within the Gibbs sampler belongs to S_k, then move type $v = (k, k+1)$ or $v = (k, k-1)$ is selected, each with equal probability. Clearly, if $k = 0$ (or $k = K$), then there is only one available move $v = (k, k+1)$ (or, respectively, $v = (k-1, k)$); similarly, if $n_k > 0$, then S_{k-1} has probability zero and the only effectively available move is $v = (k, k+1)$. The cases $k = 0$ and $k = K$ are treated similarly, with the obvious modifications. In all cases, the acceptance probability $\alpha_v(\pi_{current}, \pi_{prop})$ of a proposed point π_{prop} is either one or one half and it can be computed in a straightforward manner. Hence, to draw a new candidate π_{prop} one can use HI and ARMS to draw from the auxiliary density corresponding to the restriction of $g(\pi)\gamma(d\pi)$ over $S_i \cup S_j$, if $v = (i, j)$.

We tried our algorithm with the prior specified above on a data set containing the velocity (in 10^3 km/sec) of 82 galaxies, previously analyzed by several authors, including Roeder (1990) and Richardson and Green (1997). We allowed for a maximum of $K + 1 = 20$ components in the mixture, with $w_k \propto 1$, specifying the following values for the parameters of the prior:

$\bar{\mu} = 21$	$r = 100$	$\epsilon = 10^{-3}$
$\alpha_\mu = 2$	$\alpha_\psi = 0.5$	$\alpha_\Delta = 10$
$\beta_\mu = 10$	$\beta_\psi = 0.5$	$\beta_\Delta = 0.1$

The sampler was run for 100000 iterations, including a burn in of 10000 iterations. Let us focus on posterior model probabilities, i.e. the probability of the data coming from a mixture of normals with a specific number of components. Figure 3 shows the trace of the visited models. It is apparent

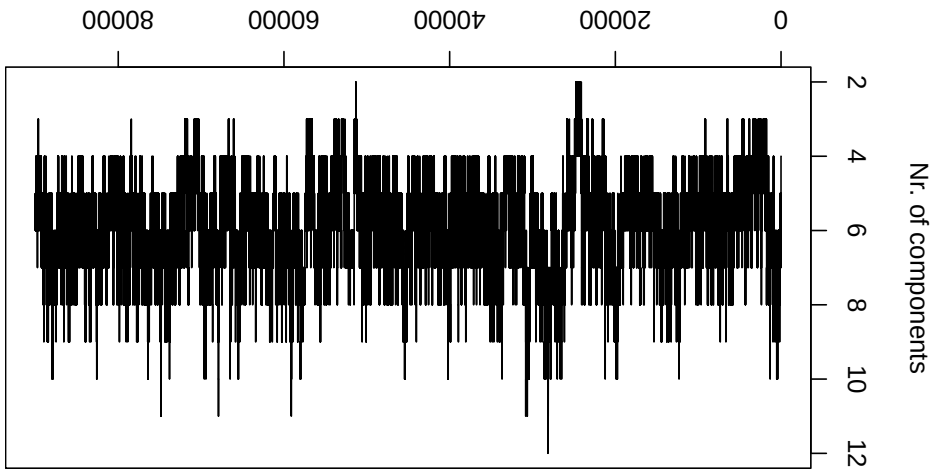

Figure 3: *Trace of visited models*

that the sampler moved fairly often among different models, which is a sign of good mixing properties. The MCMC estimates of the posterior model probabilities are given in Table 1, together with their Monte Carlo standard error, computed using Sokal's method (Sokal, 1989). Models having posterior probability below 1% are not included in the table.

Table 1: Posterior model probabilities

Nr. of comp.	3	4	5	6	7	8	9
Prob. (%)	2.8	13.2	28.7	29.2	17.4	6.2	1.6
Std. err. $\times 10^3$	4.23	8.35	8.78	5.47	4.99	2.18	5.36

6 Concluding remarks

We have shown how the transdimensional MCMC simulation scheme proposed by Petris and Tardella (2003a) can be extended from the nested model case to the more general situation of locally nested models. In fact, we have derived theoretical premises for a wider use of this all-purpose simulation scheme and we have shown the effectiveness of that approach on two of the most common problems in transdimensional MCMC, namely mixture models with unknown number of components and covariate selection in a general linear regression setting. Indeed, the new simulation scheme can be effective in

avoiding unnecessary simulations on a common maximal model space as was the case in the previous proposal of Petris and Tardella (2003a). Basically we have set up a transdimensional hybrid (compound) kernel which exploits an intermediate space where the dimension matching is performed through a geometrically intuitive inflation of the smaller submodel into the subspace of the larger one. We stress again the fact that one of the most appealing features of the proposed approach consists in avoiding the difficult evaluation of Jacobians usually required in RJ schemes to make effective split/merge steps. This is likely to facilitate routine implementation of transdimensional samplers by non-expert practitioners, possibly via stand-alone software packages such as the popular WinBUGS suite (Spiegelhalter, Thomas, Best and Gilks, 1994; Gilks, Thomas and Spiegelhalter, 1992). This remains true even for models in which the full conditional distribution of the within-model parameters is not available in closed form, so that the transdimensional Gibbs sampler of Gottardo and Raftery (2004) cannot be used.

A feature not yet fully explored that might further improve the efficiency of the simulation scheme is related to the possibility of making adaptive transformations of the component densities which contribute to define the auxiliary density in the subspace of the largest submodel. This will be the focus of future research.

References

Brooks, S., Giudici, P. and Roberts, G. (2003). Efficient construction of reversible-jump Markov chain Monte Carlo proposal distributions (with discussion), *Journal of the Royal Statistical Society, B* **65**: 3–39.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn, New York, NY: Springer-Verlag.

Carlin, B. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo, *Journal of the Royal Statistical Society, B* **57**: 473–484.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion), *Journal of the Royal Statistical Society, Series A: Statistics in Society* **158**: 419–466.

Ehlers, R. and Brooks, S. (2003). Constructing general efficient proposals for reversible-jump MCMC, *Technical report*, Federal University of Paraná, Dept. of Statistics.

Gilks, W. R., Thomas, D. and Spiegelhalter, D. (1992). Software for the gibbs sampler, *Computing Science and Statistics* **24**: 439–448.

Gottardo, R. and Raftery, A. (2004). Markov chain monte carlo with mixtures of singular distributions, *Technical Report 470*, Department of Statistics, University of Washington.

Green, P. (1995). Reversible-jump Markov chain Monte Carlo computations and Bayesian model determination, *Biometrika* **82**: 711–732.

Green, P. (2003). Trans-dimensional Markov chain Monte Carlo, *in* P. Green, N. Hjort and S. Richardson (eds), *Highly structured stochastic systems*, Oxford: Oxford University Press, pp. 179–198.

Koop, G. (2003). *Bayesian Econometrics*, New York, NY: Wiley.

Petris, G. and Tardella, L. (2001). Autoregressive model averaging: a new computational approach, *in* C. Provasi (ed.), *Modelli complessi e metodi computazionali intensivi per la stima e la previsione – S.CO. 2001.*

Petris, G. and Tardella, L. (2003a). A geometric approach to transdimensional Markov chain Monte Carlo, *The Canadian Journal of Statistics* **31**: 469–482.

Petris, G. and Tardella, L. (2003b). Regeneration techniques for MCMC, *in* C. Provasi (ed.), *Modelli complessi e metodi computazionali intensivi per la stima e la previsione – S.CO. 2003*, pp. 320–325.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society. Series B* **59**: 731–792.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies, *Journal of the American Statistical Association* **85**: 617–624.

Sisson, S. (2005). Transdimensional Markov chains: a decade of progress and future perspectives, *Journal of the Americal Statistical Association* **100**: 1077–1089.

Sokal, A. (1989). *Monte Carlo methods in statistical mechanics: foundations and new algorithms*, Cours de Troisiéme Cycle de la Physique en Suisse Romande, Lausanne.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1994). Bugs: Bayesian inference using gibbs sampling, version 0.50.