Transport Reversible Jump Proposals - Follow Along!

[a] [b] [c]

Figure: (a) Website (b) Preprint (c) Slides

Transport Reversible Jump Proposals

Using Normalizing Flows

L. Davies[1] R. Salomone[2] M. Sutton[1] C. Drovandi[1]

[1]School of Mathematical Sciences
Queensland University of Technology

[2]School of Computer Science
Queensland University of Technology

Workshop on Statistical Deep Learning, October 2022

Table of Contents

Table of Contents

Problem: Sampling a Transdimensional Space

The problem of intrest is sampling probability distribution π on

$$\mathcal{X} = \bigcup_{k \in \mathcal{K}} (\{k\} \times \Theta_k), \quad (1)$$

with *Parameters* $\boldsymbol{\theta}_k \in \Theta_k \subseteq \mathbb{R}^{n_k}$ and *Model index* (or indicator) $k \in \mathcal{K}$. We want to make inference on the joint distribution (or conditional factorization)

$$\pi(k, \boldsymbol{\theta}_k) = \pi(k)\pi(\boldsymbol{\theta}_k | k).$$

When data $\boldsymbol{y}$ is introduced this is $\pi(k, \boldsymbol{\theta}_k | \boldsymbol{y}) = \pi(k|\boldsymbol{y})\pi(\boldsymbol{\theta}_k | k, \boldsymbol{y})$.
Notation. Denote $\boldsymbol{x} = (k, \boldsymbol{\theta}_k)$, ϕ_n is n-dimensional standard normal, $\phi_{\boldsymbol{\Sigma}_n}$ is an n-dimensional normal with mean $\mathbf{0}$ and $\boldsymbol{\Sigma}_n$ covariance, $|J_f(\boldsymbol{\theta})|$ denotes absolute determinant of Jacobian matrix of function $f : \mathbb{R}^n \to \mathbb{R}^n$, and $\otimes_n \nu$ is $\underbrace{\nu \otimes \cdots \otimes \nu}_{n \text{ times}}$ for a distribution ν.

Reversible Jump Markov Chain Monte Carlo

We want to propose from point $\boldsymbol{x}$ to point $\boldsymbol{x}'$, noting $\boldsymbol{\theta}_k$, $\boldsymbol{\theta}_{k'}$ have dimensions n_k, $n_{k'}$ respectively.

- Require dimensions match: introduce auxiliary variables $\boldsymbol{u}_k \in \mathcal{U}_{k,k'} \subseteq \mathbb{R}^{w_k}$ and $\boldsymbol{u}_{k'} \in \mathcal{U}_{k',k} \subseteq \mathbb{R}^{w_{k'}}$ such that $n_k + w_k = n_{k'} + w_{k'}$.
- **Choose** a diffeomorphism e. $\boldsymbol{\theta}_{k'}, \boldsymbol{v} = h_{k,k'}(\boldsymbol{\theta}_k, \boldsymbol{u})$.

A (simplified) RJMCMC Algorithm when $n_{k'} > n_k$ is:

1. Propose model index $k' \sim j_k(\,\cdot\,)$
2. Propose auxiliary variables $u_k \sim g_{k,k'}(\,\cdot\,)$
3. Accept with probability

$$\alpha(\boldsymbol{x}, \boldsymbol{x}') = 1 \wedge \frac{\pi(\boldsymbol{x}') j_{k'}(k) g_{k',k}(\boldsymbol{u}'_{k'})}{\pi(\boldsymbol{x}) j_k(k') g_{k,k'}(\boldsymbol{u}_k)} \left| J_{h_{k,k'}}(\boldsymbol{\theta}_k, \boldsymbol{u}_k) \right|. \tag{2}$$

Transport Maps and Normalizing Flows

Consider random vectors $\boldsymbol{\theta} \sim \mu_{\boldsymbol{\theta}}$, $\boldsymbol{Z} \sim \mu_{\boldsymbol{z}}$, s.t. their distributions $\mu_{\boldsymbol{\theta}}$, $\mu_{\boldsymbol{z}}$ are absolutely continuous w.r.t. the n-dimensional Lebesgue measure.

Transport Map (TM)

A function $T : \mathbb{R}^n \to \mathbb{R}^n$ is called a *transport map* from $\mu_{\boldsymbol{\theta}}$ to $\mu_{\boldsymbol{z}}$ if $\mu_{\boldsymbol{z}} = T\sharp\mu_{\boldsymbol{\theta}}$.

Normalizing Flows (NF) and Flow-Based Models

Let $\{T_{\boldsymbol{\psi}}\}$ be a $\boldsymbol{\psi} \in \Psi$-parameterized family of diffeomorphisms with domain on the support of some arbitrary *base* distribution $\mu_{\boldsymbol{z}}$. Then, for fixed $\boldsymbol{\psi}$, the PDF of the random vector $\boldsymbol{\zeta} = T_{\boldsymbol{\psi}}(\boldsymbol{Z})$ is

$$\mu_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}; \boldsymbol{\psi}) = \mu_{\boldsymbol{z}}(T_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\zeta}))|J_{T_{\boldsymbol{\psi}}^{-1}}(\boldsymbol{\zeta})|, \ \boldsymbol{\zeta} \in R^n. \tag{3}$$

Distributions $\mu_{\boldsymbol{\zeta}}$ are *flow-based models*, where $\{T_{\boldsymbol{\psi}}\}$ are the *normalizing flows*.

With finite samples $\boldsymbol{s} \sim \pi$, we obtain an *approximate* TM $\hat{T}$ via density estimation, minimising the KLD from $\{\boldsymbol{s}\}$ to $\mu_{\boldsymbol{\zeta}}$.

Table of Contents

Proposed Method: Transport Reversible Jump Proposals

First, let ν be a univariate analytic distribution e.g. ϕ.

Proposed Method: Transport Reversible Jump Proposals

First, let ν be a univariate analytic distribution e.g. ϕ.

For each model $k \in \mathcal{M}$:

Proposed Method: Transport Reversible Jump Proposals

First, let ν be a univariate analytic distribution e.g. ϕ.
For each model $k \in \mathcal{M}$:

- Define reference distributions $\bigotimes_{n_k} \nu$.

Proposed Method: Transport Reversible Jump Proposals

First, let ν be a univariate analytic distribution e.g. ϕ.

For each model $k \in \mathcal{M}$:

- Define reference distributions $\bigotimes_{n_k} \nu$.
- Train flows on samples $\boldsymbol{s}_k \sim \pi(\boldsymbol{\theta}_k | k)$ to obtain $\hat{T}_k$.

Proposed Method: Transport Reversible Jump Proposals

First, let ν be a univariate analytic distribution e.g. ϕ.
For each model $k \in \mathcal{M}$:

- Define reference distributions $\bigotimes_{n_k} \nu$.
- Train flows on samples $\boldsymbol{s}_k \sim \pi(\boldsymbol{\theta}_k | k)$ to obtain $\hat{T}_k$.

Then, a transdimensional proposal where $n_{k'} > n_k$ is

$$\begin{aligned} \boldsymbol{z}_k &\leftarrow T_k(\boldsymbol{\theta}_k), \\ \boldsymbol{z}'_{k'} &\leftarrow \bar{h}_{k,k'}(\boldsymbol{z}_k, \boldsymbol{u}_k), \\ \boldsymbol{\theta}_{k'} &\leftarrow T_{k'}^{-1}(\boldsymbol{z}'_{k'}), \end{aligned} \tag{4}$$

Proposed Method: Transport Reversible Jump Proposals

First, let ν be a univariate analytic distribution e.g. ϕ.

For each model $k \in \mathcal{M}$:

- Define reference distributions $\bigotimes_{n_k} \nu$.
- Train flows on samples $\boldsymbol{s}_k \sim \pi(\boldsymbol{\theta}_k | k)$ to obtain $\hat{T}_k$.

Then, a transdimensional proposal where $n_{k'} > n_k$ is

$$\begin{aligned} \boldsymbol{z}_k &\leftarrow T_k(\boldsymbol{\theta}_k), \\ \boldsymbol{z}'_{k'} &\leftarrow \bar{h}_{k,k'}(\boldsymbol{z}_k, \boldsymbol{u}_k), \\ \boldsymbol{\theta}_{k'} &\leftarrow T_{k'}^{-1}(\boldsymbol{z}'_{k'}), \end{aligned} \tag{4}$$

where $\bar{h}_{k,k'}$ is a *volume-preserving* diffeomorphism on $\bigotimes_{n_k} \nu$.

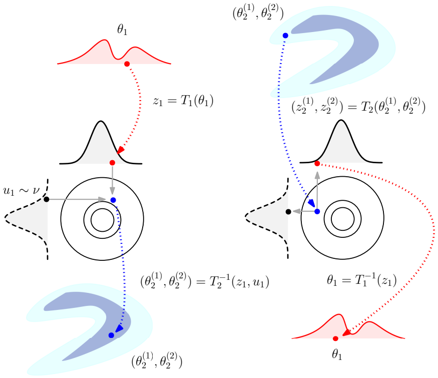


Figure: Illustration of the proposal class. Here, the reference ν is Gaussian. The diffeomorphisms ($\bar{h}$) on the reference distributions simply concatenate or extract coordinates as required.

Proposition: RJMCMC with Exact TMs

Proposition 1

Suppose that RJMCMC proposals are of the form described in (4), and for each $k \in \mathcal{K}$, satisfy $T_k \sharp \pi_k = \otimes_{n_k} \nu$. Then, (2) reduces to

$$\alpha(\boldsymbol{x}, \boldsymbol{x}') = 1 \wedge \frac{\pi(k')}{\pi(k)} \frac{j_{k'}(k)}{j_k(k')}. \tag{5}$$

Corollary

Provided the conditions of Proposition 1 are satisfied, choosing $\{j_k\}$ such that

$$\pi(k') j_{k'}(k) = \pi(k) j_k(k'), \quad \forall k, k' \in \mathcal{K}, \tag{6}$$

leads to a rejection-free proposal.

Table of Contents

Sinh Arcsinh 1D 2D Example

As an illustrative example with known TMs, we use the (element-wise) inverse sinh-arcsinh transformation of [Jones and Pewsey, 2009]

$$S_{\boldsymbol{\epsilon},\boldsymbol{\delta}}(\cdot) = \sinh(\boldsymbol{\delta}^{-1} \odot (\sinh^{-1}(\cdot) + \boldsymbol{\epsilon})),$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^n, \boldsymbol{\delta} \in \mathbb{R}^n_+$. For an $n \times n$ matrix L, define a transform $T(\boldsymbol{Z})$ where

$$T(\boldsymbol{Z}) = S_{\boldsymbol{\epsilon},\boldsymbol{\delta}}(L\boldsymbol{Z}), \tag{7}$$

and $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}_n, I_{n\times n})$. The probability density function for the transformed variable $\boldsymbol{\theta} = T(\boldsymbol{Z})$ is

$$p_{\boldsymbol{\epsilon},\boldsymbol{\delta},L}(\boldsymbol{\theta}) = \phi_{LL^\top}\left(S^{-1}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}(\boldsymbol{\theta})\right)|J_{S^{-1}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}}(\boldsymbol{\theta})|,$$

where $S^{-1}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}(\cdot) = \sinh(\boldsymbol{\delta} \odot \sinh^{-1}(\cdot) - \boldsymbol{\epsilon})$.

Sinh Arcsinh 1D 2D Example

The target of interest for this example, where $\boldsymbol{\theta}_1 = (\theta_1^{(1)})$ and $\boldsymbol{\theta}_2 = (\theta_2^{(1)}, \theta_2^{(2)})$, is

$$\pi(k, \boldsymbol{\theta}_k) = \begin{cases} \frac{1}{4} p_{\epsilon_1, \delta_1, 1}\big(\boldsymbol{\theta}_1\big), & k = 1, \\ \frac{3}{4} p_{\boldsymbol{\epsilon}_2, \boldsymbol{\delta}_2, L}\big(\boldsymbol{\theta}_2\big), & k = 2, \end{cases} \tag{8}$$

where

$$\begin{aligned} \epsilon_1 &= -2, & \delta_1 &= 1, \\ \boldsymbol{\epsilon}_2 &= (1.5, -2), \quad \text{and} & \boldsymbol{\delta}_2 &= (1, 1.5), \end{aligned} \tag{9}$$

and L is a lower-triangular matrix such that

$$LL^\top = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}.$$

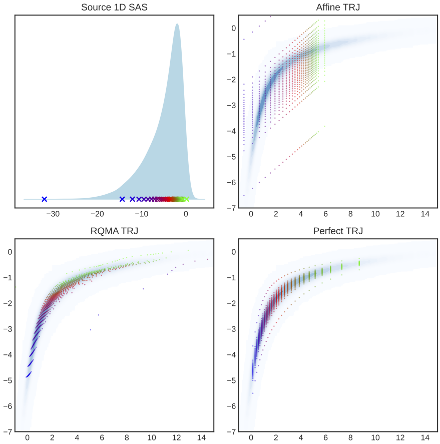
By construction, for chosen reference distributions ϕ_n, $n_k = k$, the exact transport is given by the function

$$T^{-1}(\cdot) = L^{-1} S^{-1}_{\boldsymbol{\epsilon}, \boldsymbol{\delta}}(\cdot). \tag{10}$$

Systematic draws from conditional target $\pi(x_1|k=1)$ of (8) are transported from $(1, \theta_1) \in \mathcal{K} \times \mathbb{R}^1$ (*top left*) to $(2, (\theta_1, \theta_2)) \in \mathcal{K} \times \mathbb{R}^2$ via TRJ proposals using:

- Top right: Approximate affine,
- Bottom left: Approx RQMA-NF,
- Bottom right: Perfect TM.

The auxilliary variables in the proposals are also drawn systematically (30 for each point in the source distribution).



Example: Sinh Arcsinh Target with Transport RJ Proposal

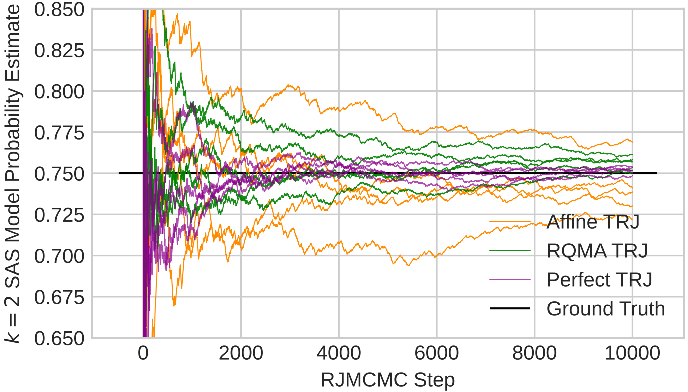


Figure: Running estimates of the model probabilities for the $k = 2$ component of the Sinh-Arcsinh target. Proposal are all TRJ with input TMs (1) Affine, (2) RQMA-NF, (3) Perfect. Ten chains on each proposal type are depicted, where alternating within-model proposals are a simple normal random walk.

Table of Contents

Modified Bartolucci Bridge Sampling Estimator

For an RJMCMC chain, [Bartolucci et al., 2006] showed that the Bayes factor $B_{k,k'}$ (ratio of marginal likelihoods) is estimated via

$$\hat{B}_{k,k'} = \frac{N_{k'}^{-1} \sum_{i=1}^{N_{k'}} \alpha_i'}{N_k^{-1} \sum_{i=1}^{N_k} \alpha_i}, \tag{11}$$

where $N_{k'}$ and N_k are the number of proposed moves from model k' to k, and from k to k', respectively in the run of the chain.
In the special case when prior model probabilities are uniform, convert estimators of Bayes factors to estimators of model probabilities [Bartolucci et al., 2006,] via

$$\hat{\pi}(k) = \hat{B}_{j,k}^{-1}\left(1 + \sum_{i \in \mathcal{K}\setminus\{j\}} \hat{B}_{i,j}\right)^{-1}, \text{ for arbitrary } j \in \mathcal{K}. \tag{12}$$

The **Modified Bartolucci Estimator** (MBE) simply adopts the above for proposals from *samples* of the conditional targets.

Table of Contents

Bayesian Factor Analysis

We model monthly exchange rates of six currencies relative to the British pound, spanning January 1975 to December 1986 [West and Harrison, 1997,], denoted as $\boldsymbol{y}_i \in \mathbb{R}^6$ for $i = 1, ..., 143$, of the random vector $\boldsymbol{Y}$.

We assume $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}_6, \boldsymbol{\Sigma})$, where

- $\boldsymbol{\Sigma} = \boldsymbol{\beta}_k \boldsymbol{\beta}_k^\top + \boldsymbol{\Lambda}$,
- Λ is a 6×6 positive diagonal matrix,
- $\boldsymbol{\beta}_k$ is a $6 \times k$ lower-triangular matrix with a positive diagonal,
- k is the number of factors, $\boldsymbol{\theta}_k$ dimension $6(k+1) - k(k-1)/2$.

Bayesian Factor Analysis: Model Configuration

Following [Lopes and West, 2004], for each $\boldsymbol{\beta}_k = [\beta_{ij}]$ with $i = 1, \ldots, 6$, $j = 1, \ldots, k$, the priors are

$$\begin{aligned} \beta_{ij} &\sim \mathcal{N}(0,1), \quad i < j \\ \beta_{ii} &\sim \mathcal{N}_+(0,1), \\ \Lambda_{ii} &\sim \mathcal{IG}(1.1, 0.05), \end{aligned} \tag{13}$$

We are interested in the posterior probability of $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \boldsymbol{\Lambda})$ for $k = 2$ or 3 factors, with $\boldsymbol{\theta}_k$ dimensions 17 and 21 respectively. Via Bayes' Theorem the posterior is

$$\pi(k, \boldsymbol{\theta}_k | \boldsymbol{y}) \propto p(k) p(\boldsymbol{\beta}_k | k) p(\boldsymbol{\Lambda}) \prod_{i=1}^{143} \phi_{\boldsymbol{\beta}\boldsymbol{\beta}^\top + \boldsymbol{\Lambda}}(\boldsymbol{y}_i), \tag{14}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_{143})$.

Bayesian Factor Analysis: Proposal Design

Original [Lopes and West, 2004] Independence Proposal

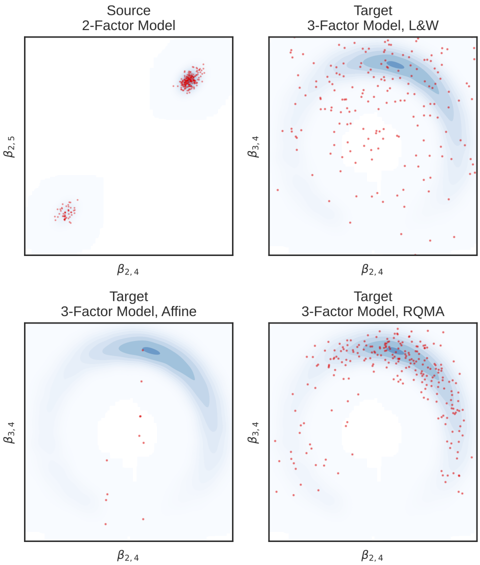
Write $\boldsymbol{\mu}_{\boldsymbol{\beta}_k}$, $\boldsymbol{B}_k$ as the posterior mean and covariance of $\boldsymbol{\beta}_k$. Denoting $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \boldsymbol{\Lambda})$, the independence proposal is

$$q_k(\boldsymbol{\theta}_k) = q_k(\boldsymbol{\beta}_k) \prod_{i=1}^{6} q_k(\Lambda_{ii}), \tag{15}$$

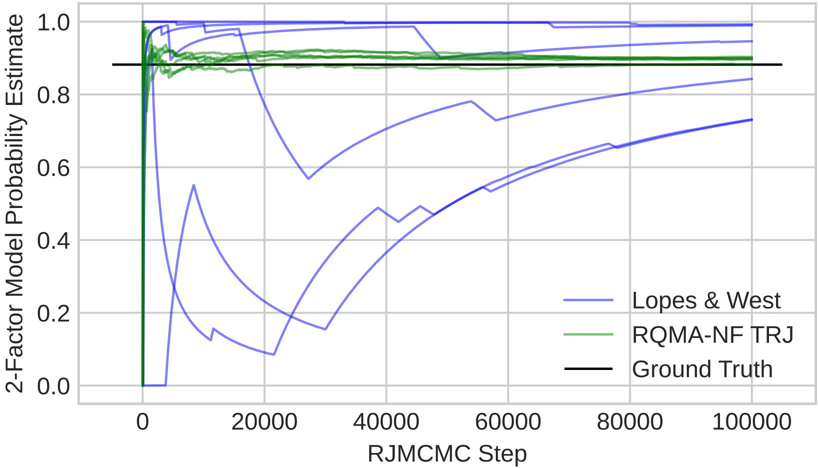
where for $k \in \mathcal{K}$, $q_k(\boldsymbol{\beta}_k) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}_k}, 2\boldsymbol{B}_k)$, and $q_k(\Lambda_{ii}) = \mathcal{IG}(18, 18 v_{k,i}^2)$ where $v_{k,i}^2$ is the approximate conditional posterior mode of Λ_{ii} given k.

We compare the [Lopes and West, 2004] proposal to Affine and RQMA-NF TRJ trained on finite draws $\boldsymbol{s} \sim \pi(\boldsymbol{\theta}_k | k)$ obtained via HMC-NUTS (for $k = 3$) and SMC (for $k = 2$).

Bayesian Factor Analysis: Proposal Comparison



Bayesian Factor Analysis: Running Estimates from RJMCMC Chain



Bayesian Factor Analysis: MBE Study

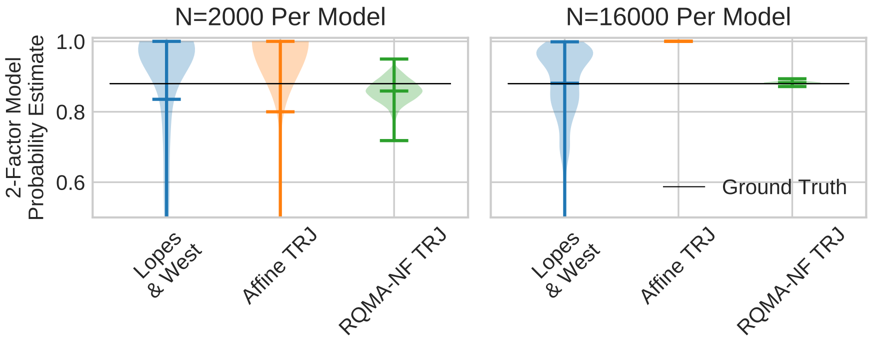


Figure: Violin plot showing the variability of the 2-factor model probability estimates in the case where only the 2-factor and 3-factor models are compared. Model probability estimates are obtained via the MBE. Ground truth is estimated via extended individual SMC runs ($N = 5 \cdot 10^4$).

Table of Contents

Using One Transport (to rule them all)

Problem: Currently, we need to train an approximate TM for each model $k \in K$.

Using One Transport (to rule them all)

Problem: Currently, we need to train an approximate TM for each model $k \in K$.

Solution: Re-frame target so that a *conditional* approximate TM can be used.

Conditional Transport Reversible Jump Proposals

The dimension-saturation approach, originally formalized in [Brooks et al., 2003], is an equivalent formulation of RJMCMC that involves an augmented target. Writing the maximum model dimension as $n_{\max}$ (assumed to be finite) and recalling that $\boldsymbol{x} = (\boldsymbol{\theta}_k, k)$, we define our augmented target to be

$$\tilde{\pi}(\tilde{\boldsymbol{x}}) = \pi(\boldsymbol{x})(\otimes_{n_{\max}-n_k}\nu)(\boldsymbol{u}_{\frown k}), \tag{16}$$

where $\tilde{\boldsymbol{x}} = (k, \boldsymbol{\theta}, \boldsymbol{u}_{\frown k})$, and "$\frown k$" identifies that the auxiliary variable is of dimension $n_{\max} - n_k$. In this setting, one can obtain approximate TMs by training a single conditional NF with the conditioning vector being the model index $k \in \mathcal{K}$. The associated proposals analogous to those in (4) are constructed as

$$(\boldsymbol{\theta}'_{k'}, \boldsymbol{u}_{\frown k'}) = c_{k'}^{-1} \circ \tilde{T}^{-1}(\cdot|k') \circ \tilde{T}(\cdot|k) \circ c_k(\boldsymbol{\theta}_k, \boldsymbol{u}_{\frown k}), \tag{17}$$

where $k' \sim j_k$, and c_k is simply concatenation.

Example: Block Variable Selection in Robust Regression

We are interested in realizations of a random response variable Y through a linear combination of predictor variables X_1, X_2, X_3 and $\boldsymbol{\beta} = (\beta_0, ..., \beta_3)$ parameters in a regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

We model the residual error term as a mixture between standard normal variable and a normal variable with a large variance. Use the notation for the model space $k = (1, k_1, k_2, k_2)$ where $k_i \in \{0, 1\}$ for $i = 1, 2$. The prior distributions are specified as

$$\begin{aligned} &k_i \sim \text{Bernoulli}(1/2),\ i \in \{1, 2\}, \quad \text{and} \\ &\beta_i \sim \mathcal{N}(0, 10^2),\ i \in \{0, 1, 2, 3\}. \end{aligned} \tag{18}$$

The target π is then the posterior distribution over the set of models and regression coefficients defined using Bayes' Theorem.

Example: Block Variable Selection in Robust Regression

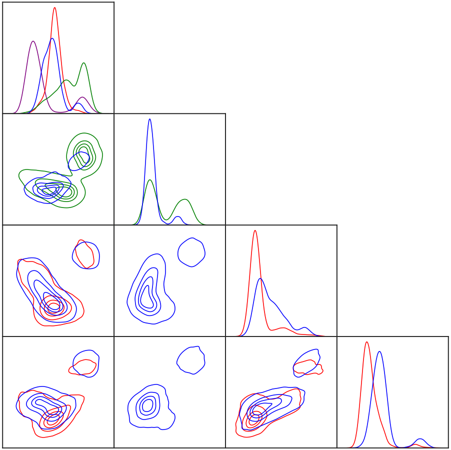


Figure: Pairwise plot of the conditional bivariate posterior densities in the Bayesian variable selection example. All four models feature: $k = (1, 0, 0, 0)$ (*Purple*) , $k = (1, 1, 0, 0)$ (*Green*), $k = (1, 0, 1, 1)$ (*Red*), and $k = (1, 1, 1, 1)$ (*Blue*).

Example: Block Variable Selection in Robust Regression

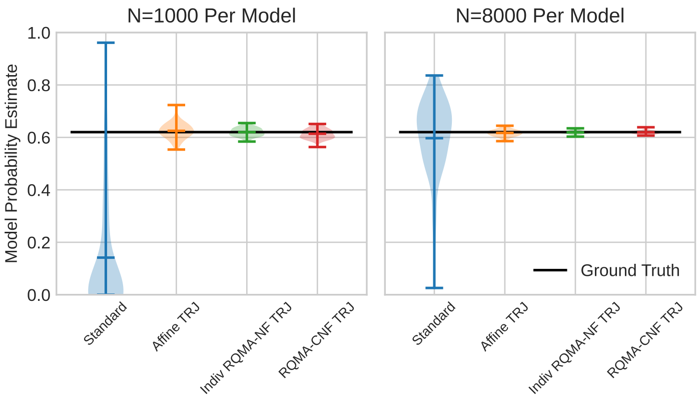


Figure: Violin plot showing the variability of the $k = (1, 1, 1, 1)$ model probability estimate for each proposal type using the MBE vs ground truth individual SMC ($N = 5 \cdot 10^4$). Individual SMC with $N = 1000, 8000$ particles sampled conditional targets split into training/test samples for a total of 80 passes.

Table of Contents

Conclusions and Future Work

- We have introduced a theoretically rejection-free approach for RJMCMC proposal design utilizing transport maps.

Conclusions and Future Work

- We have introduced a theoretically rejection-free approach for RJMCMC proposal design utilizing transport maps.
- Proposals using approximate transport maps yield good results when compared to baseline approaches.

Conclusions and Future Work

- We have introduced a theoretically rejection-free approach for RJMCMC proposal design utilizing transport maps.
- Proposals using approximate transport maps yield good results when compared to baseline approaches.
- There is the caveat of requiring samples to train the approximate TMs (but pilot runs are also required in many other approaches).
- We have introduced the idea of using a conditional normalizing flow to reduce training time. This would be useful for large model spaces!

Conclusions and Future Work

- We have introduced a theoretically rejection-free approach for RJMCMC proposal design utilizing transport maps.
- Proposals using approximate transport maps yield good results when compared to baseline approaches.
- There is the caveat of requiring samples to train the approximate TMs (but pilot runs are also required in many other approaches).
- We have introduced the idea of using a conditional normalizing flow to reduce training time. This would be useful for large model spaces!
- Efforts are justified in expensive-likelihood scenarios.

Conclusions and Future Work

- We have introduced a theoretically rejection-free approach for RJMCMC proposal design utilizing transport maps.
- Proposals using approximate transport maps yield good results when compared to baseline approaches.
- There is the caveat of requiring samples to train the approximate TMs (but pilot runs are also required in many other approaches).
- We have introduced the idea of using a conditional normalizing flow to reduce training time. This would be useful for large model spaces!
- Efforts are justified in expensive-likelihood scenarios.
- Finally, whilst the MBE benchmark was used to assess cross-model proposal quality, the results seem promising and justify further investigation in lieu of standard RJMCMC.

References

Bartolucci, F., Scaccia, L., and Mira, A. (2006).
Efficient Bayes factor estimation from the reversible jump output.
Biometrika, 93(1):41–52.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003).
Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1):3–39.

Jones, M. C. and Pewsey, A. (2009).
Sinh-arcsinh distributions.
Biometrika, 96(4):761–780.

Lopes, H. F. and West, M. (2004).
Bayesian model assessment in factor analysis.
Statistica Sinica, 14(1):41–67.

West, M. and Harrison, J. (1997).
Bayesian forecasting and dynamic models.
Springer series in statistics. Springer, New York, 2nd ed edition.