Estimation of Logistic Smooth Transition Autoregressive Models Using Bayesian Shrinkage Methods

Mario Giacomazzo
Arizona State University, Tempe, USA - mgiacoma@asu.edu

Yiannis Kamarianakis
Arizona State University, Tempe, USA - ikamaria@asu.edu

Abstract

The logistic smooth transition autoregressive model (LSTAR) is a regime-switching nonlinear time series model that has been adopted in a wide variety of applications. LSTAR is represented by a weighted average of two or more linear autoregressive (AR) processes. Bayesian LASSO and horseshoe priors on the autoregressive coefficients provide a more flexible model selection procedure than currently used alternatives. A simulation study is used to demonstrate the efficacy of these methods. Application to a classic nonlinear time series further illustrates the ability of these methods to achieve superior forecasting performance.
Keywords: Gibbs Sampler Algorithm; Regime-Switching Model; Bayesian LASSO.

1. Introduction

Consider the univariate time series of interest y_t and let $\boldsymbol{x}_t' = [1, y_{t-1}, y_{t-2}, \cdots, y_{t-p}]$. Let also $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \cdots, \alpha_p]$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_p]$ denote two vectors of unknown coefficients. A special class of parametric nonlinear time series models follows the form in Equation 1.

$$y_t = (\boldsymbol{x}_t'\boldsymbol{\alpha})(1 - G(z_t, \gamma, \delta)) + (\boldsymbol{x}_t'\boldsymbol{\beta})G(z_t, \gamma, \delta) + \epsilon_t \text{ where } \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2) \tag{1}$$

If $0 \leq G(z_t, \gamma, \delta) \leq 1$, this model is a weighted average of two autoregressive processes of order p (AR(p)) where the weight depends on the value of the transition variable z_t. When $z_t = y_{t-d}$ the term "self-exciting" is often applied and a delay parameter d is introduced (Petruccelli & Woolford, 1984). This model becomes a logistic smooth transition autoregressive model of order p (LSTAR(p)) when $G(y_{t-d}, \gamma, \delta) = \{1 + \exp[-(\gamma^*/s_y)(y_{t-d} - \delta)]\}^{-1}$. The addition of the unknown slope parameter $\gamma^* > 0$ and threshold parameter δ makes this process nonlinear. When $y_{t-d} < \delta$, $G(y_{t-d}, \gamma, \delta) < 1/2$, the AR(p) model $\boldsymbol{x}_t'\boldsymbol{\alpha}$ in the "low regime" is favored, and when $y_{t-d} > \delta$, $G(y_{t-d}, \gamma, \delta) > 1/2$, the AR(p) model $\boldsymbol{x}_t'\boldsymbol{\beta}$ in the "high regime" is favored. The transition slope γ^* determines the speed of transition between low and high regimes around δ. Scaling γ^* by the sample standard deviation of the transition variable s_y allows for scale-free comparisons across competing STAR models with differing transition variables (Deschamps, 2008). As $\gamma^* \to \infty$, $G(y_{t-d}, \gamma, \delta) \to \mathbb{1}_{\{y_{t-d} > \delta\}}(y_{t-d})$ that evaluates to 1 if $y_{t-d} > \delta$ and 0 otherwise. In the limiting case when regime changes are abrupt, the model is called a threshold autogressive model of order p (TAR(p)). Although the focus is on the homoskedastic case, it is not hard to fathom the variance of y_t exhibiting regime switching dynamics along with the mean of y_t. The majority of research regarding STAR models revolve around the two regime case; however, extensions have been made to account for multiple (>2) regimes (MR-STAR).

The Bayesian approach for estimating two-regime LSTAR(p) models was developed by Lubrano (2000). Lopes and Salazar (2006) expanded the aforementioned sampling algorithm to include the model order p, using the reversible jump markov chain monte carlo (RJMCMC) algorithm presented in Green (1995). These changes were inspired by Troughton and Godsill (1997) who applied RJMCMC to AR(p) models. Further work by Gerlach and Chen (2008) accounted for heteroskedasticity. Current Bayesian estimation methods of the LSTAR(p) typically assume that the autoregressive order p is the same in both regimes and include all autoregressive terms y_{t-k} for $k \in \{1, 2, ..., p\}$. If the true nonlinear data generating process (DGP) has

regime-specific model orders and some autoregressive terms are not significant, the above-mentioned estimation method is expected to be suboptimal in terms of out-of-sample predictive accuracy.

Section 2 explains how two Bayesian estimation methods for sparse signals can be incorporated in the sampling algorithm for LSTAR models. Section 3 provides simulation results showing the efficacy of the two methods. Section 4 presents a forecasting exercise which is based on benchmark data that have been analyzed extensively in previous studies. Section 5 gives a positive outlook on how these methods may further advance Bayesian estimation of complicated nonlinear processes.

2. Methodology

Recall the 2-regime LSTAR(p) model in Equation 1 and define the full vector of unknown parameters $\boldsymbol{\theta} = [\alpha_0, \alpha_1, \cdots, \alpha_p, \beta_0, \cdots, \beta_p, \gamma, \delta, \sigma, d, p]'$ where $\gamma = \gamma^*/s_y$. In regards to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and σ^2 the prior specifications are generally $\alpha_k \sim N(\mu_\alpha, \sigma_\alpha^2)$, $\beta_k \sim N(\mu_\beta, \sigma_\beta^2)$, $1/\sigma^2 \sim IG(a_{\sigma^2}, b_{\sigma^2})$. To ensure that sufficient representation exists in both regimes, the prior for δ is defined as follows: $\delta \sim U[q_Y(0.15), q_Y(0.85)]$ where $q_Y(.)$ is the empirical quantile function of the observed transition variable. Using the 15^{th} and 85^{th} percentiles requires at least 15% of the data to be part of both regimes. Due to the difficulty in the estimation for $\gamma = \gamma^*/s_y$ discussed across the literature, a variety of priors with positive supports have been used: Gamma (Lopes & Salazar, 2006), Truncated Normal (Livingston & Nur, 2016), and Log Normal (Gerlach & Chen, 2008). For reasons justified by Gerlach and Chen (2008), the prior favored herein is $\gamma^* \sim LN(\mu_\gamma, \sigma_\gamma^2)$. The parameter d is typically given a discrete uniform prior $P(d = \tilde{d}) = 1/d_{max}$ for $\tilde{d} \in \{1, 2, \cdots, d_{max}\}$, where d_{max} is chosen *a priori*. From this point on, d is assumed to be known and is removed from $\boldsymbol{\theta}$.

Sampling algorithms of the joint posterior $f(\boldsymbol{\theta}|\boldsymbol{y})$ exploit the fact that the LSTAR(p) model is conditionally linear given γ^* and δ. Specifically, Gibbs sampling is applied for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and σ^2 (Gelfand & Smith, 1990) and Metropolis-Hastings (Metropolis et.al., 1953; Hastings, 1970) for γ^* and δ. Since the length of $\boldsymbol{\theta}$ increases with the model order p, Lopes and Salazar (2006) extend the sampling algorithm outlined by Lubrano (2000) to incorporate a reversible jump step to include the model order p in $\boldsymbol{\theta}$. RJMCMC allows the dimension of the sampled vector $\boldsymbol{\theta}$ to change dimension from $2(p+1)+3$ to $2(p'+1)+3$ whenever proposed changes from p to p' are accepted. Posterior analysis of p relies on comparing posterior model probabilities. The most likely model order $\hat{p}$ is defined as $\hat{p} = \max_p \#\{p_s = p | s \in 1, \cdots, S\}/S$ where S represents the number of samples from the joint posterior distribution after burn-in and p_s represents the sampled value at iteration s.

Now, let p_1 be the true linear AR model order in the low regime and p_2 be its equivalent in the high regime. Furthermore, let $p = \max\{p_1, p_2\}$. The two cases where the previously explained method may never sample from the correct parameter space are when $p_1 \neq p_2$ or when $\exists\, j < p$ such that $\alpha_j = 0 \cup \beta_j = 0$. Minor adjustments may be made to allow for a more flexible posterior inference of the true DGP by specifying p *a priori* and utilizing alternative Bayesian model selection procedures. The conditional linear nature of the LSTAR(p) model invites a plethora of Bayesian techniques for model selection through the editing of the priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For insight into the varieties, see O'Hara and Sillanpaa (2009); the focus here is on two Bayesian analogues to the regularization techniques developed by the statistics and the machine learning communities, known as LASSO-type estimators.

Bayesian methods for linear models have been developed to mimic penalized least squares designed to combat overfitting by shrinking insignificant parameters to 0. Replacing the normal priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with double-exponential (Laplace) priors is the cornerstone of Bayesian LASSO (BLASSO; Park & Casella, 2008). The prior hierarchy is inspired from Andrews and Mallows (1974) who demonstrated the double-exponential distribution can be expressed as a scaled-mixture of normal distributions. A global shrinkage parameter λ is introduced and conditional priors are expressed as shown in Equation 2:

$$\alpha_j|\sigma^2, \tau_{\alpha_j}^2 \sim N(0, \sigma^2\tau_{\alpha_j}^2),\ \beta_j|\sigma^2, \tau_{\beta_j}^2 \sim N(0, \sigma^2\tau_{\beta_j}^2),\ \tau_{\alpha_j}^2| \sim EXP(\lambda^2/2),\ \tau_{\beta_j}^2| \sim EXP(\lambda^2/2). \tag{2}$$

The parameter λ is usually chosen via cross-validation in frequentist analyses; in Bayesian spirit the gamma hyperprior $\lambda^2 \sim G(a_\lambda, b_\lambda)$ is applied here. The full Gibbs sampler outlined by Park and Casella (2008) may

be applied to the LSTAR(p) model with Metropolis-Hastings still used for parameters $\{\gamma^*, \delta\}$.

The horseshoe prior of Carvalho et. al. (2009) is similar to BLASSO since it can be expressed as a scale-mixture of normals. Along with a global shrinkage parameter λ, Bayesian horseshoe adds local shrinkage parameters λ_{α_j} and λ_{β_j}. This change allows finer discrimination between relevant and non-significant autoregressive parameters by preventing the simultaneous over-shrinking that may occur to the parameter space in BLASSO. The Bayesian horseshoe (BHS) prior hierarchy in the LSTAR(p) context is presented in Equation 3 with C^+ denoting the half-Cauchy distribution:

$$\alpha_j|\lambda_{\alpha_j} \sim N(0, \lambda_{\alpha_j}),\ \beta_j|\lambda_{\beta_j} \sim N(0, \lambda_{\beta_j}),\ \lambda_{\alpha_j} \sim C^+(0, \lambda),\ \lambda_{\beta_j} \sim C^+(0, \lambda),\ \lambda|\sigma^2 \sim C^+(0, \sigma). \tag{3}$$

Unfortunately, posterior sampling here does not compare to the ease of the Gibbs sampler for BLASSO since full conditional distributions cannot be found analytically. Nevertheless, Carvalho et. al. (2009) provide theoretical and empirical justifications for the horseshoe prior over the double-exponential prior and slice sampling methods have been developed (Makalic & Schmidt, 2016).

Once a maximum considered model order p is specified *a priori*, BLASSO and BHS provide flexible model building alternatives: unlike RJMCMC, nonlinear LSTAR(p) model estimation can be easily conducted using popular Bayesian software such as JAGS (Plummer, 2003). Also, these methods may be applied to general STAR and TAR nonlinear models. Because the regressors are lags of the endogenous time series y_t, scaling is unnecessary in this context. Also, intercepts in time series models carry a different interpretation than the usual linear regression context; therefore, these parameters are also considered for shrinkage.

3. Simulation Study

Consider the nonlinear time series in Equation 4 for which 100 replications are made each of length 1000 after a burn-in period of 500. This simulation study is identical to the one found in Lopes and Salazar (2006). The maximum model order across regimes is specified $p = 4$ *a priori* and $d = 2$ is assumed to be known. Under this specification, $\boldsymbol{\theta} = [\alpha_0, \alpha_1, \cdots, \alpha_4, \beta_0, \cdots, \beta_4, \gamma, \delta, \sigma]' = [0, 1.8, -1.06, 0, 0, 0.02, 0.9, -0.265, 0, 0, 100, 0.02, 0.02]'$.

$$\begin{aligned} y_t &= (1.8y_{t-1} - 1.06y_{t-2})[1 - G_1(y_{t-2})] + (0.02 + 0.9y_{t-1} - 0.265y_{t-2})[G_1(y_{t-2})] + \epsilon_t \\ &\text{where: } G_1(y_{t-2}) = \left\{1 + \exp\big[-100(y_{t-2} - 0.02)\big]\right\}^{-1} \\ &\text{and } \epsilon_t \sim \text{i.i.d. } N(0, 0.02^2). \end{aligned} \tag{4}$$

Bayesian estimation of the underlying LSTAR(2) model is conducted using BLASSO and BHS priors. To match the regularization paths of the common LASSO, posterior medians are used parameter estimates in BLASSO (Park & Casella, 2008). For BHS, Carvalho et. al. (2009) recommend using posterior means. After a burn-in period of 15,000 with a thinning of 10, to reduce autocorrelation and control computer memory usage, the sampler is updated a maximum of 20 times until both convergence across three chains is met and the effective sample size for each parameter exceeds 150. All prior parameters are chosen to be non-informative and starting values are randomly chosen. A non-informative log normal prior $LN(3, 1)$ is used for γ^*. For BLASSO, 91% of the replications converged; for these replications, the mean and median number of samples required were 11,615 and 2,000 respectively. All replications converged for BHS with a mean and median number of samples required of 1,600 and 1,000, respectively.

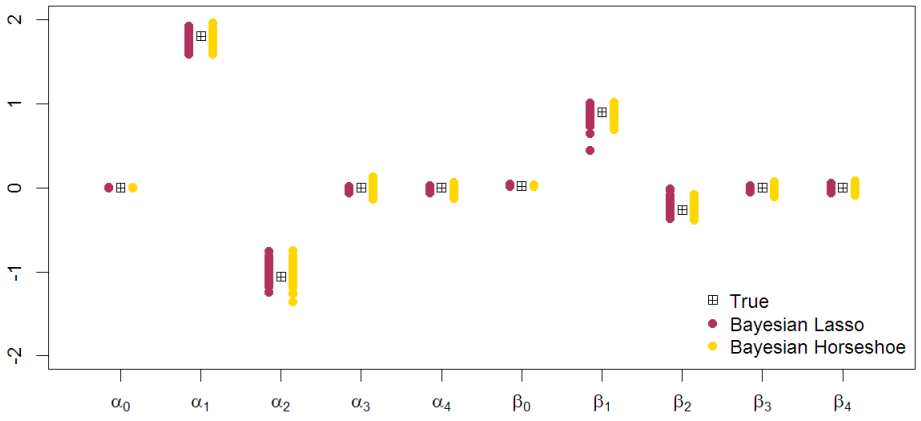
Table 1 provides summary statistics of the posterior estimates using BLASSO and BHS. Medians and means of all posterior estimates are given for replications that converged. Rather than reporting the standard deviation of the estimates, $RMSE(\theta) = \sqrt{\sum(\hat{\theta} - \theta)^2/n}$ is reported for each parameter to measure estimation error. Consistent overestimation and large uncertainty for γ is commonly reported in literature. Although significant model parameters are shrunk closer to 0, accurate signal detection is observed for both BLASSO and BHS. Figure 1 plots all 100 posterior estimates of the autoregressive parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. BLASSO performs better in identifying parameters that are truly zero but also tends to over shrink non-zero parameters

in the high regime. Examination of RMSE in Table 1 highlights this.

Table 1: Results Summarizing Posterior Results from BLASSO and BHS

		Bayesian Lasso			Bayesian Horseshoe		
Parameter	Actual	Mean	Median	RMSE	Mean	Median	RMSE
α_0	0	0.0012	0.0014	0.0024	0.0007	0.0006	0.002
α_1	1.8	1.7638	1.7662	0.0726	1.768	1.7689	0.0746
α_2	-1.06	-1.0037	-1.0087	0.1002	-1.0104	-1.011	0.1159
α_3	0	-0.0042	-0.0033	0.0102	-0.0125	-0.0107	0.0434
α_4	0	-0.0035	-0.0012	0.0138	-0.0028	-0.0011	0.031
β_0	0.02	0.0198	0.0192	0.0038	0.0202	0.0197	0.0035
β_1	0.9	0.8763	0.8856	0.0742	0.8899	0.8949	0.0528
β_2	-0.265	-0.2294	-0.2357	0.0706	-0.2496	-0.2538	0.053
β_3	0	-0.0067	-0.0042	0.0134	-0.0081	-0.004	0.0296
β_4	0	0	-0.0004	0.0148	0.0019	0.0011	0.0253
σ	0.02	0.0201	0.0202	0.0004	0.0201	0.0201	0.0004
γ	100	134.5089	111.5546	85.2702	174.15	128.7485	148.8093
δ	0.02	0.0216	0.0211	0.0047	0.0208	0.0201	0.0038

Figure 1: Posterior Estimates from BLASSO and BHS



4. Application to Annual Sunspot Numbers

Daily international sunspot numbers are gathered and updated by the World Data Center SILSO, Royal Observatory of Belgium, Brussels. Since 1957 (Granger, 1957), the annually aggregated time series x_t =Annual Average Sunspot Number at year t has served as classical illustration of a nonlinear process. As proposed by Ghaddar and Tong (1981), nonlinear time series analysis is applied to the square root transformed time series $y_t = 2[\sqrt{1+x_t} - 1]$. Annual average sunspot numbers from 1700 to 1979 are used to fit the models, while years 1980 to 2006 are held out to examine forecasting performance for horizons $h \in \{1, 2, \cdots, 5\}$.

A textbook example by Terasvirta (2010) compares three nonlinear time series models, namely STAR, TAR, and Artificial Neural Nets (AR-NN), to the baseline linear AR model. Nonlinear least squares and hypothesis tests were used to fit and select the best model under each of these model types. The LSTAR model in Equation 5 outperformed all other models based on $RMSFE(h)$ for forecast horizons $h \in \{1, 2, \cdots, 5\}$. This model exhibits different autoregressive orders between the low and high regimes and also exhibits gaps.

$$
\begin{aligned}
y_t = &(1.46y_{t-1} - 0.76y_{t-2} + 0.17y_{t-7} + 0.11y_{t-9})[1 - G_1(y_{t-2}, 5.5, 7.9)] \\
&+ (2.7 + 0.92y_{t-1} - 0.01y_{t-2} - 0.47y_{t-3} + 0.32y_{t-4} - 0.26y_{t-5} \\
&+ 0.17y_{t-7} - 0.24y_{t-8} + 0.11y_{t-9} + 0.17y_{t-10})G_1(y_{t-2}, 5.5, 7.9) + \hat{\epsilon}_t \\
\text{where: } &\hat{\epsilon}_t \sim N(0, 1.898^2).
\end{aligned} \tag{5}
$$

The fully saturated LSTAR(10) model estimated via nonlinear least squares is presented as a baseline. Under prior assumptions that $d = 2$ and the maximum model order $p \leq 10$, models estimated through BLASSO and BHS are compared to the aforementioned models. BLASSO required 4000 posterior samples to converge while Bayesian horseshoe required 11,000 posterior samples. This finding regarding efficiency is contrary to the results from the simulation experiment.

Out-of-sample forecasts are obtained using bootstrapped samples from the empirical error distribution and a rolling window. Evaluation of $RMSFE(h)$ for horizons 1 to 5 is presented in Table 2. The LSTAR(10) model was estimated by nonlinear least squares and the poor results indicate the necessity for a more parsimonious specification. It is important to note that Bayesian estimation of the saturated LSTAR(10) model failed to converge after 200,000 iterations. The Bayesian shrinkage methods not only converged but achieved forecasting performance at least as good as the best model estimated by Terasvirta. In this example, the results from Bayesian horseshoe are impressive especially at long horizons when compared to the other three methods.

Table 2: $RMSFE(h)$ for Horizons $h \in \{1, 2, \cdots, 5\}$

h	Terasvirta	LSTAR(10)	Bayesian Lasso	Bayesian Horseshoe
1	1.51	1.86	1.66	1.42
2	2.24	3.21	2.28	1.96
3	2.65	3.72	2.64	2.28
4	2.58	3.6	2.55	2.18
5	2.67	3.25	2.41	2.08

6. Conclusions

Bayesian shrinkage methods can be applied to nonlinear regime-switching models with different transition functions to the one presented here. Future work involves applying and evaluating these methods to the estimation of multiple regime smooth transition models (MR-STAR) and nonlinear models with exogenous regressors. When exogenous regressors are lags of another time series, the autoregressive order representing the maximum significant lag of the exogenous time series becomes a new parameter of interest. In both cases, the dimension of the model matrix increases dramatically; Bayesian shrinkage estimation in these more complex situations appears as a viable alternative to multiple nested RJMCMC routines.

References

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society Series B (Methodological)*, 99-102.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research*, 5, 73-80.

Deschamps, P. J. (2008). Comparing smooth transition and Markov switching autoregressive models of US unemployment. *Journal of Applied Econometrics*, 23, 435-462.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398-409.

Gerlach, R., & Chen C.W.S. (2008). Bayesian inference and model comparison for asymmetric smooth transition heteroskedastic models. *Statistics and Computing*, 18(4), 391-408.

Ghaddar, D., & Tong, H. (1981). Data Transformation and Self-Exciting Threshold Autoregression. *Journal of the Royal Statistical Society*, 30(Series C), 238-248.

Granger, C. W. J. (1957). A Statistical Model for Sunspot Activity. *The Astrophysical Journal*, 126, 152.

Green, P. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4), 711-732.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97-109.

Livingston Jr, G., & Nur, D. (2017). Bayesian inference for Smooth Transition Autoregressive (STAR) model: A prior sensitivity analysis. *Communications in Statistics-Simulation and Computation*, 1-22.

Lopes, H. F., & Salazar, E. (2006). Bayesian model uncertainty in smooth transition autoregressions. *Journal of Time Series Analysis*, 27(1), 99-117.

Lubrano, M. (2000). Bayesian analysis of nonlinear time series models with a threshold. In *Nonlinear Econometric Modeling in Time Series: Proceedings of the Eleventh International Symposium in Economic Theory* (Vol. 11, p. 79). Cambridge University Press.

Makalic, E., & Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1), 179-182.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087-1092.

O'Hara, R. B., & Sillanpaa, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1), 85-117.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

Petruccelli, J., & Woolford, S. (1984). A threshold AR(1) model. Journal of Applied Probability, 21(2), 270-286.

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

SILSO World Data Center. International sunspot number monthly bulletin and online catalogue. (1970-2006). Available from: http://www.sidc.be/silso/

Terasvirta, T., Tjostheim, D. & Granger, C. W. J. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Troughton, P. T., & Godsill, S. J. (1997). A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. Technical Report. University of Cambridge: Department of Engineering, Cambridge, UK.