Rotated Slice Sampling for Efficient and Robust Estimation of DSGE Models

Marco Ratto (with Ludovic Calès and Filippo Pericoli)

European Commission, Joint Research Centre

DYNARE GSA Course *November 2019, Ispra*

Introduction

- We investigate the performance of the rotated slice sampler first introduced by Planas Ratto Rossi (2015) (PRR2015) for the Bayesian estimation of medium/large scale DSGE models.
- We measure the relative performance of samplers based on inefficiency factors and we also test the rejection rates of the samplers against known distributions.
- We benchmark the rotated slice sampler together with the standard slice and the Metropolis-Hastings in the estimation of the Smets and Wouters (2003, 2007) model.
- the method is regularly used for the estimation of the European Commission's Global Multicountry model (Albonico et al. , 2019).

European Commission

Main outcomes

- Unlike Metropolis, the Slice algorithm does not require any ad-hoc tuning nor any proposal distribution.
- The Rotated Slice extension boosts the efficiency of the sampler in case of highly correlated posterior distributions.
- A further extension of the rotated algorithm allows to dramatically increase the mixing properties of the slice sampler in successfully exploring the shape of complex multi-modal distributions.
- A parallel implementation of the algorithm provides accurate posterior draws with a computational cost comparable to the preliminary posterior maximization required by Metropolis $\Rightarrow$ the slice algorithm strongly advisable in medium/large scale DSGE models.

Bayesian inference of DSGE

Mostly using MCMC methods: obtaining draws from the the posterior distribution $p(\theta, z|y)$ using the factorization:

$$p(\theta, z|y) = p(\theta|y)p(z|\theta, y)$$

- $\theta \sim p(\theta|y) \quad \Leftarrow$ **our main concern**
- $z \sim p(z|\theta, y)$ off-line by a simulation smoother (e.g. Durbin and Koopman, 2002) using the Kalman filter.

European Commission

Sampling model parameters

Draws $\theta \sim p(\theta|y)$ can be obtained in several ways:

- Random walk proposal Metropolis-Hastings (**Dynare**): easy to implement, faster than many other implementations, but sometimes inefficient.
- Slice sampler (Neal, 2003, also in **Dynare**): **offers an avenue that we wish to explore carefully**.
- ...

Slice sampler's main features

Advantages:

- Only require $f(\theta|y) \propto p(\theta|y)$ (like Metropolis);
- No proposal: does not require mode-Hessian or other info for defining proposal;
- Tuning parameters less important than Metropolis;
- Provides a framework for
  - adaptation
  - suppressing random walk behaviour *(over-relaxation)*.

Disadvantages:

- Not easy to generalize in the multivariate framework;
- Requires a large number of likelihood evaluations.

European Commission

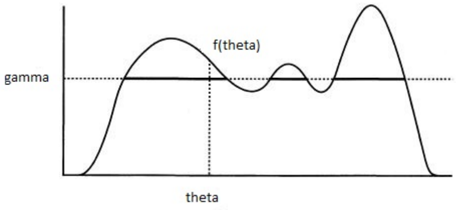
In this talk

- The univariate slice sampler
- Some univariate test cases - **to mimic** $\theta_i \sim f(\theta_i|\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_d, y)$, $i = 1, 2, \cdots, d$
- Rotated univariate sampler - **correlated and multimodal distributions**
- Test cases and DSGE applications
- Conclusions

The idea of slice sampling

Introduce an *auxiliary variable* γ and construct $p(\theta, \gamma)$ taking the marginal $p(\theta)$ unchanged. Sampling from the joint $p(\theta, \gamma)$ is not possible but ...
draws from $p(\theta)$ can be obtained iterating Gibbs updates on $\gamma|\theta$ and $\theta|\gamma$:

- sample γ given θ from a uniform pdf over the set $(0, f(\theta))$
- sample θ given γ from a uniform pdf over $S = \{\theta : \gamma < f(\theta)\}$



Slice sampling in practice

Sampling θ from a uniform over $S = \{\theta : \gamma < f(\theta)\}$ is difficult to achieve exactly (perfect slice sampling). In practice:

- Position $I = (L, R)$ around θ^0 at random that contains S as much as possible;
- Draw θ from the set $A = \{\theta : \theta \in S \cap I \;\; and \;\; \Pr(I|\theta) = \Pr(I|\theta^0)\}$

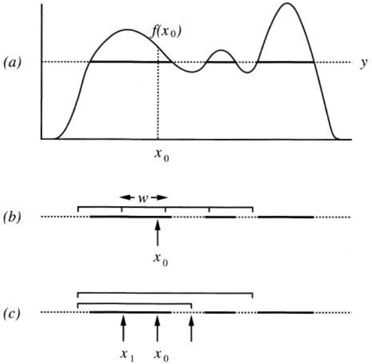
Neal (2003) proposes some strategies:

(i) stepping out;

(ii) doubling;

(iii) random positioning, etc.

Stepping out



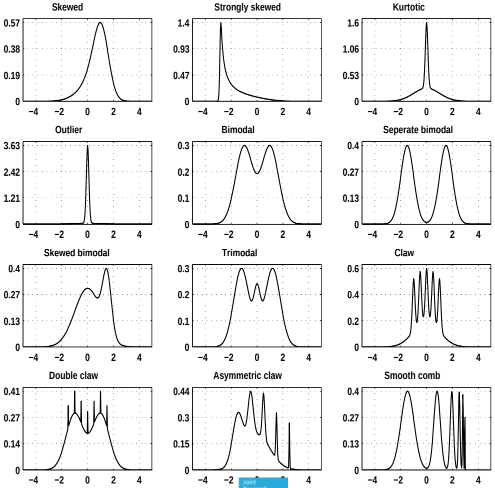
European Commission

The performance of the slice sampler

We start using a battery of univariate pdfs. By 1000 replications of sample size $G = 5000$, we measure

- NSE: $Var(\frac{1}{G}\sum_{i=1}^{G}\theta^{i})^{1/2}$, $i = 1, 2, \cdots, G$ (small)
- ρ_1: the 1st order autocorrelation of the chain $\theta^1, \cdots, \theta^G$ (close to 0)
- $IF = 1 + 2\sum_{j=1}^{p}\omega_j\rho_j$, ω_j the Parzen-weights (close to 1)
- the (average) number of calls to $f(\theta)$ (small)
- the number of rejections of the Cramer-Von Mises (CVM) test (5%)

Univariate test cases - Marron and Wand (1992)



Univariate test cases - summary results

	RW-MH	Slice							
		Stepping out				Doubling			
W		$\frac{1}{2}\sigma$	3σ	10σ	100σ	$\frac{1}{2}\sigma$	3σ	10σ	100σ
NSE	3.77	2.75	1.92	1.83	1.79	2.30	2.02	1.85	1.80
ρ_1	0.71	0.23	0.15	0.12	0.11	0.24	0.18	0.14	0.12
IF	6.38	3.97	1.51	1.38	1.34	2.24	1.68	1.42	1.35
N eval	1	9.42	6.09	6.61	9.72	23.31	14.65	9.47	10.00
CVM	0.59	0.22	0.13	0.11	0.11	0.20	0.15	0.12	0.12
RE	1	4.1	1.42	1.42	2.04	7.44	3.72	2.12	2.12

Slice sampling for multivariate distributions

- Generalizing the stepping out procedure for a d-dimensional θ (PRR2015):
  - need to evaluate $f(\theta|y)$ 2^d-times to compute the vertices of the hypercube that approximates the Slice
    $\Rightarrow$ parallelization as in Tibbits et al. (2011)?
  - not easy to approximate the Slice in many dimensions by axis-aligned hypercubes
    $\Rightarrow$ gradient of $f(\theta)$
    $\Rightarrow$ directional hypercubes for high-correlated parameters
  - still difficult to solve the curse of dimensionality issue
- **Here: Rotated univariate slice for high-correlated parameters** as first proposed by PRR2015 and further developed here.

Rotated univariate sampler

We use a rotated orthonormnal basis:

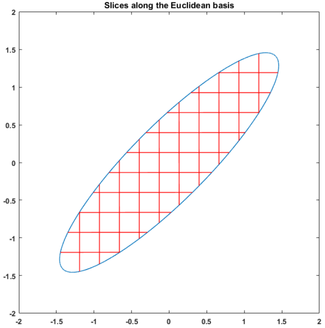
Let $\Sigma \equiv Var(\theta|y)$. Any rough estimate (e.g. burn-in period) will suffice.

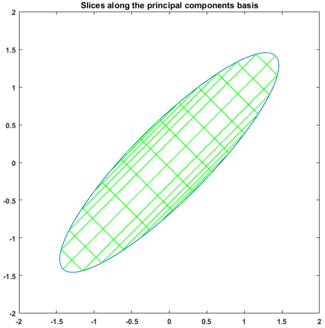
By spectral decomposition $\Sigma = A\Lambda A'$.

- A, the eigenvectors of Σ, suggests in which direction to rotate the axes;
- the eigenvalues in Λ suggest the scale of the slices along each direction. We use $W_i = 3\Lambda_{ii}^{1/2}$.

This simple algorithm allows to increase the efficiency of the univariate sample for correlated multivariate distributions [same cost as plain OAT for orthogonal ones].

Illustrations of the classical stepping out method by Neal (2003) (left) and the rotated stepping out method (right).





DYNARE Implementation

The rotated slice sampling is based on procedure with at least 2 steps:

1. burnin OAT, calling `estimation` with options
   - `posterior_sampling_method='slice'`
   - the parameter `mh_replic=50, mh_blocks= ...`
2. Rotated
   - `posterior_sampling_method='slice'`
   - `posterior_sampler_options= ('rotated',1,'use_mh_covariance_matrix',1)`
   - `load_mh_file`, which is loading the variance-covariance matrix obtained in the previous step.
   - the parameter `mh_replic=50-1000, mh_blocks= ...`

Multivariate test cases

We test the accuracy of some multivariate slice algorithms in four different examples:

1. $\theta \sim N_d(0, I_d)$, $d = 2, 5, 10$

   $\Rightarrow$ highlights issues even for well behaving distributions

2. Bivariate Mixture of 2 Normals (Chib - Ramamurthy, 2010)

   $\Rightarrow$ slow mixing of RW-MH even for small dimension; effectiveness of rotated slice

3. $\theta \sim N_d(0, \rho 11' + (1 - \rho) I_d)$, $d = 10$

   $\Rightarrow$ shows the effectiveness of rotated slice

4. Mixture of 4 Normals (Chib - Ramamurthy, 2010) $d = 12$

   $\Rightarrow$ shows the effectiveness of rotated slice

Case 1. Gaussian - uncorrelated variables

$\theta \sim N_d(0, \quad I_d),\ d = 2, 5, 10$

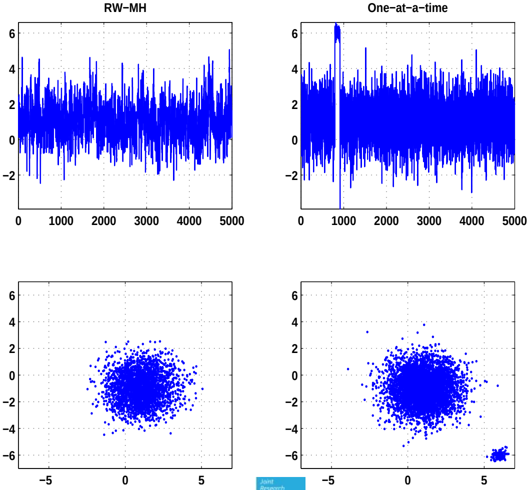
	RW-MH			Slice		
				One-at-a-time		
d	**2**	**5**	**10**	**2**	**5**	**10**
NSE	5.0	5.8	8.2	1.4	1.4	1.4
ρ_1	.78	.90	0.95	.00	.00	.00
IF	7.84	17.7	40.7	.95	.96	.95
N eval	1	1	1	12	30	60
CVM	.83	.95	.99	.06	.07	.07
RE	1	1	1	1.46	1.62	1.40

Case 2. Chib Ramamurthy (2001)

$\theta \sim .99 \times N_2(\mu_1, \Sigma_1) + .01 \times N_2(\mu_2, \Sigma_2)$
where $\mu_1 = (1, -1)$, $\mu_2 = 6\mu_1$, $\Sigma_1 = 1.3 I_2$, and $\Sigma_2 = 0.05 I_2$.

	RW-MH	Slice One-at-a-time
Max NSE	5.28	2.87
Max ρ_1	0.81	0.04
Max IF	12.94	11.9
N eval	1	11.99
CVM	0.82	0.2
RE	1	11.02

Case 2. Chib Ramamurthy (2010)



European Commission

Case 3. Gaussian - high correlated variables

$$\theta \sim N_d(0, \quad .95 \times 11' + .05 \times I_d),\ d = 10$$

	RW-MH	Slice	
		OAT	Rotated OAT
Max NSE	10.4	11.3	1.8
Max ρ_1	.98	.94	0.2
Max IF	91	122	3.3
N eval	1	71.2	61
CVM	1	.96	0.1
RE	1	95.6	2.2

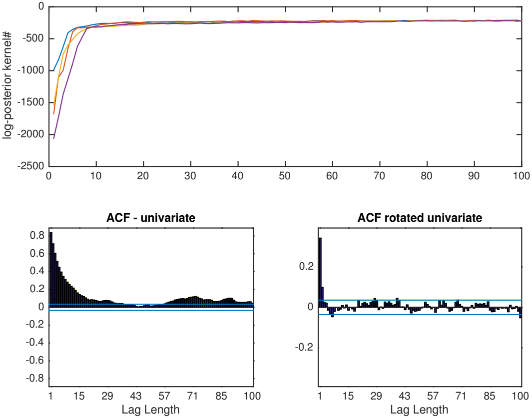
Slice sampler

- Rule of thumb for the tuning parameter $W \simeq 3\sigma$;
- Stepping out procedure generally outperforms the doubling one (and other not reported here);
- Univariate slice sampling more appealing than multivariate one when variables are loosely correlated;
- Rotated univariate sampler works well with highly-correlated variables and multi-modal problems.
- How bout medium-large scale problems?

DSGE examples

- Smets Wouters (2003, 2007)
- EC Global Multicountry model (Albonico et al. , 2019);
- An Schorfheide (2007);

Smets Wouters (2003)



Smets Wouters (2003, 2007)

SW2003 32 parameters

	RW-MH	Slice Rotated OAT
IF	110	3.2
N eval	1	183
RE	1	5.37

SW2007 36 parameters

	RW-MH	Slice Rotated OAT
IF	110	3.3
N eval	1	210
RE	1	6.3

Modified Implementation: go parallel

Once chains properly start, not clear trade-off between slice and MH. BUT: can slice be run at the same cost of optimizer??

1. burnin OAT, calling `estimation` with options
   - `posterior_sampling_method='slice'`
   - the parameter `mh_replic=50` → 30-35, `mh_blocks=250`
2. Rotated
   - `posterior_sampling_method='slice'`
   - `posterior_sampler_options= ('rotated',1,'use_mh_covariance_matrix',1)`
   - `load_mh_file`, which is loading the variance-covariance matrix obtained in the previous step.
   - the parameter `mh_replic=100-1000` → 15-20, `mh_blocks=250`

Q: How short can be slice/MH chains?

- We test the performance of the slice sampler and MH for estimating the parameters of a SW2003-like posterior (approximated with a truncated normal).
- We apply the multivariate version of the Cramer test proposed by Baringhaus and Franz (2004)
- We explore the posterior of the model through samples of 250 points, obtained as the last element of 250 parallel chains (each point has 32 parameters/dimension)
- We repeat the experiment 30 times to get the distribution of the p-value of the test (each point has 32 parameters/dimension)

Slice sampler performance after rotation - non truncated normal

We give the p-values of the Cramer test at different poinst of the chain, during burnin and after rotation. Null of the test: the sample comes from the known distribution of the posterior.

SLICE SAMPLER CRAMER TEST - NON TRUNCATED NORMAL

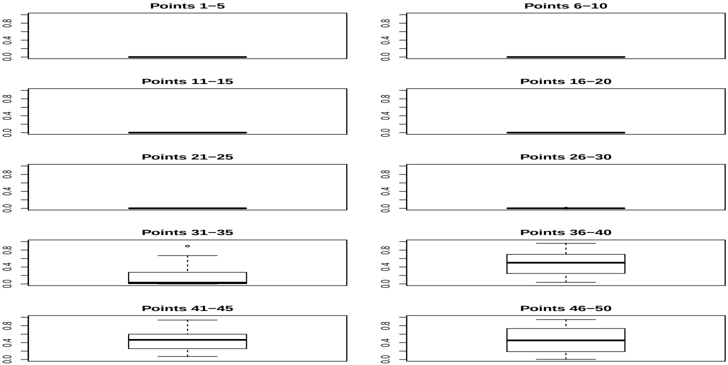
	BURNIN						SAMPLER AFTER ROTATION									
	1	10	20	30	40	50	150	250	350	450	550	650	750	850	950	1050
µ	0.00	0.00	0.00	0.02	0.34	0.45	0.57	0.48	0.51	0.49	0.56	0.54	0.43	0.56	0.53	0.56
σ	0.00	0.00	0.00	0.02	0.21	0.29	0.26	0.34	0.27	0.27	0.28	0.29	0.24	0.28	0.30	0.27
# <0.05	30	30	30	27	1	3	1	3	0	0	0	1	1	0	1	1
% <0.05	100.0%	100.0%	100.0%	90.0%	3.3%	10.0%	3.3%	10.0%	0.0%	0.0%	0.0%	3.3%	3.3%	0.0%	3.3%	3.3%

SLICE SAMPLER CRAMER TEST - TRUNCATED NORMAL

	BURNIN						SAMPLER AFTER ROTATION									
	1	10	20	30	40	50	150	250	350	450	550	650	750	850	950	1050
µ	0.00	0.00	0.00	0.01	0.46	0.45	0.50	0.46	0.53	0.48	0.46	0.56	0.47	0.49	0.39	0.51
σ	0.00	0.00	0.00	0.01	0.23	0.29	0.26	0.30	0.27	0.28	0.29	0.22	0.28	0.34	0.30	0.30
# <0.05	30	30	30	30	2	2	0	3	0	2	3	0	2	5	2	0
% <0.05	100.0%	100.0%	100.0%	100.0%	6.7%	6.7%	0.0%	10.0%	0.0%	6.7%	10.0%	0.0%	6.7%	16.7%	6.7%	0.0%

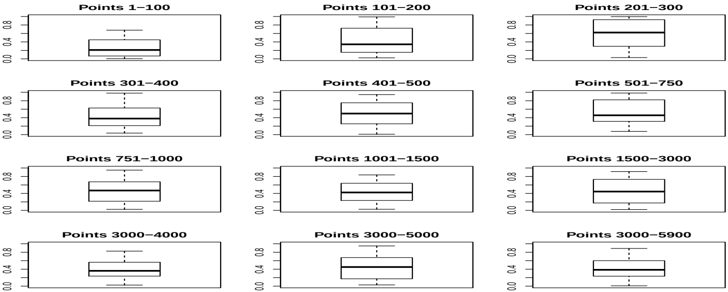
Adaptive Slice sampler performance before/after rotation - truncated normal

OAT length: 30-35 iterations; rotated length: 15-20 iterations.



Metropolis Hastings - truncated normal

We randomize the p-values of the Cramer test at different point of the chain. Null of the test: the sample comes from the known distribution of the posterior.



A: How short can be slice/MH chains?

- A few tens of slice iterations (*without* the fixed cost of optimizer)
- A few hundreds of MH iterations (*after* the optimization step)

are sufficient to get good iid sample from a lot of parallel chains

- Rotated slice iterations can start after about 20-30 iterations, reducing computational cost.
- How about overhead cost of optimization for MH?

Computational cost of slice.

One slice iteration uses on average 6-8 function evaluation per parameter. So, for 50 iterations, the total number of function evaluations N for n parameters can estimated as:

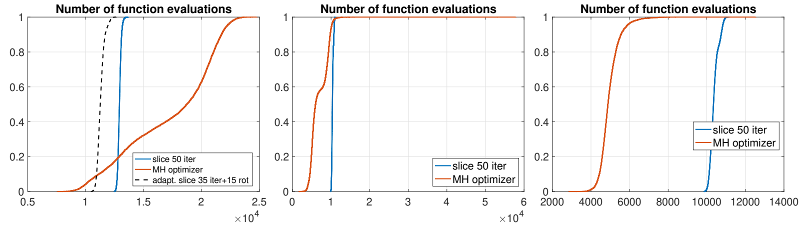
$$N = 50 \times 7 \times n$$

Note. Never set the 'usual' huge number of MCMC iterations applied for MH. Slice will last forever for such a big number!

Computational cost: slice vs. optimizer

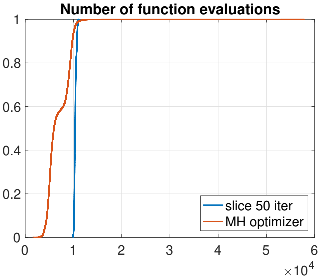
Number of function evaluations to run 50 slice iterations vs optimizer.
Across different tests, optimizer may last more than the 50 slice iterations.
Optimization becomes more difficult:

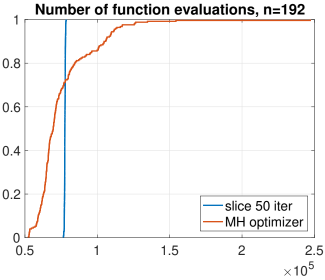
- when posterior is more concentrated w.r.t. prior
- presence of cliffs (e.g. no solution, B-K violations etc.)
- large number of parameters n.



Computational cost: slice vs. optimizer. $n = 32$ vs $n = 192$

Number of function evaluations to run 50 slice iterations vs optimizer.
Increasing the size of the problem, cost of optimizer increases w.r.t. the 50 slice iterations.





European Commission

Rotated univariate sampler: multimodal problems

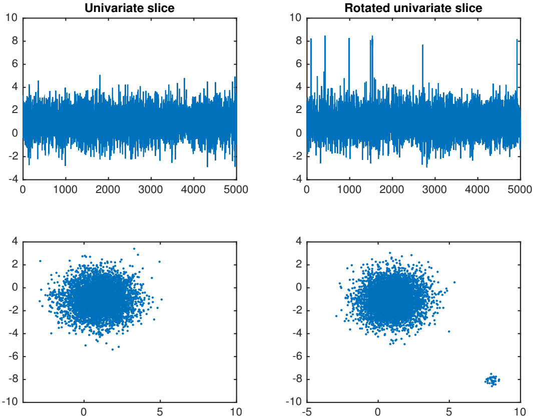
The same idea can also be used to improve mixing and efficiency of the sampler in the case of multi-modal problems. As in Chib Ramamurthy (2010) rely on assumption that we know the location of the different local optima:

1. identify (at least some of) the multiple local optima characterizing the posterior distribution;
2. perform slice sampling along rotated axes, parallel to lines joining different local optima

Case 2. Chib Ramamurthy (2010) REVISITED

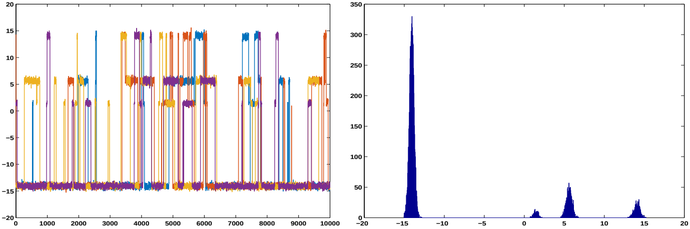
$\theta \sim .99 \times N_2(\mu_1, \Sigma_1) + .01 \times N_2(\mu_2, \Sigma_2)$

where $\mu_1 = (1, -1)$, $\mu_2 = 8\mu_1$, $\Sigma_1 = 1.3I_2$, and $\Sigma_2 = 0.05I_2$.



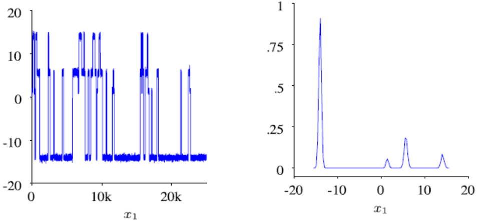
Case 4. Chib Ramamurthy (2010)

Mixture of 4 normals in 12 dimensions with probability $p_j = [0.75, 0.05, 0.15, 0.05]$



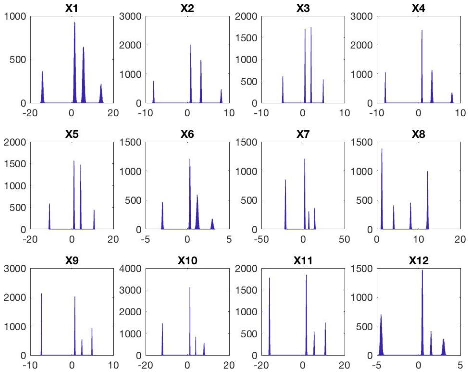
Case 4. Chib Ramamurthy (2010)

Mixture of 4 normals in 12 dimensions with probability $p_j = [0.75, 0.05, 0.15, 0.05]$.
Excerpt from Chib Ramamurthy:



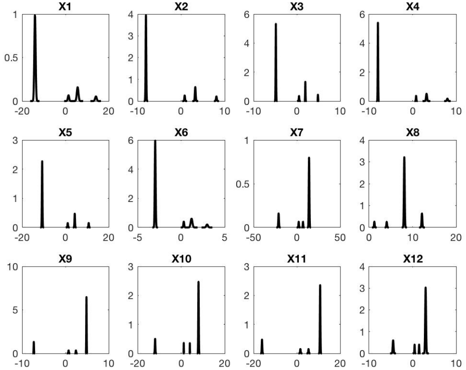
Case 4. Chib Ramamurthy (2010): go parallel

Parallel implementation of mixture of 4 normals in 12 dimensions with probability $p_j = [0.75, 0.05, 0.15, 0.05]$. It easily finds the four modes, but misses relative weight.



Case 4. Chib Ramamurthy (2010): go parallel

Compute the marginal likelihood for the four optima: $\hat{p}_j = [0.73, 0.05, 0.14, 0.05]$.
Re-weight estimated densities accordingly:

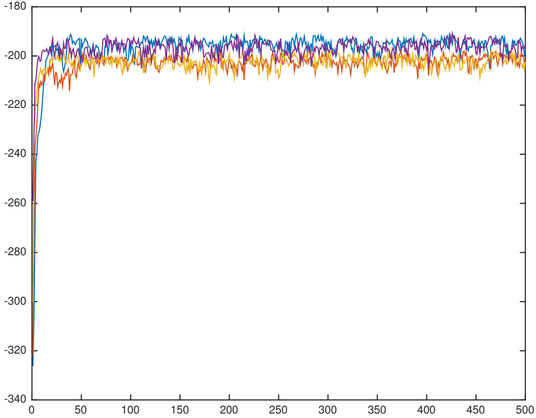


An Schorfheide (2007)

- multimodal posterior distribution;
- although ‘low’ mode has very low probability (0.005%), MH or standard slice samplers incapable to jump to high density region
- rotated slice sampler can deal with this (likewise Chib Ramamurthy, 2010, with Tailored randomized block MCMC)

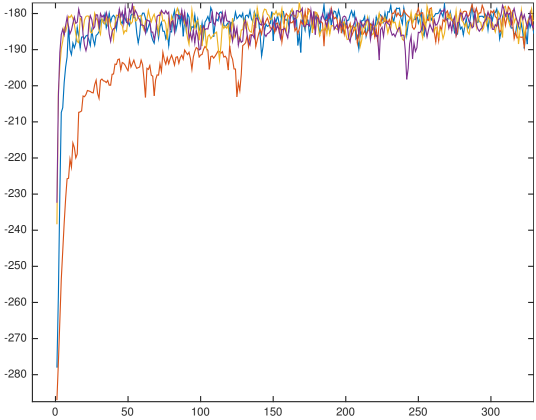
An Schorfheide (2007)

Univariate slice sampler:



An Schorfheide (2007)

Rotated univariate slice sampler:



An Schorfheide (2007): go parallel

- perform many short chains as usual: very easily get to the two modes.
- Compute the relative marginal likelihood for the two optima: $\hat{p}_{2,1} = 0.03\%$ [relative likelihood at mode: $\hat{L}_{2,1} = 0.005\%$]
- Re-weight estimated densities accordingly...

Conclusions

Slice sampler has appealing properties for usage with DSGE models:

- does not require optimization to compute the mode and the Hessian;
- small IF and short chains: good to get a few 'good' points in the posterior space, with a limited initial budget of function evaluations;
- rotated slice makes the method much more efficient, even at short number of iterations;
- available in DYNARE.

[Note. Optimizer 5 of DYNARE is based on the same principle, with Gibbs steps.]

Conclusions

Many short parallel chains are to be preferred to long chains:

- provides iid samples;
- applies to any kind of sampler MH/slice etc.
- applies to large problems;
- short slice chains (50 iterations) are often faster than the optimizer needed to start MH;
- effective also for multimodal problems.

European Commission

References

Albonico, A., Calès, L., Cardani, R., Croitorov, O., Ferroni, F., Giovannini, M., Hohberger, S., Pataracchia, B., Pericoli, F., Raciborski, R., Ratto, M., Roeger, W., Vogel, L. (2019). Comparing post-crisis dynamics across Euro Area countries with the Global Multi-country model. Economic Modelling, 2019, https://doi.org/10.1016/j.econmod.2019.04.016.

Chib, A. and Ramamurthy, S. (2010). Tailored Randomized Block MCMC Methods with Application to DSGE Models. Journal of Econometrics, vol. 155(1), pp. 19-38.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. Biometrika, vol. 57(1), pp. 97-109.

Kollmann, R., Pataracchia, B., Raciborski, R., Ratto, M., Roeger, W., and Vogel, L. (2015). The Post-Crisis Slump in the Euro Area and the US: Evidence from an Estimated Three-Region DSGE Model. European Economic Review, vol. 88, pp. 21-41.

Marron, J. S. and Wand M.P. (1992). Exact Integrated Squared Error. The Annals of Statistics, vol. 20(2), pp. 712-736.

Neal, R. M. (2003). Slice Sampling. Annals of Statistics, vol. 31(3), pp. 705-767.

Planas, C., M. Ratto, and A. Rossi (2015). Slice Sampling in Bayesian Estimation of DSGE models. DYNARE Conference.

Smets, F. and R. Wouters (2003). An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area. Journal of the European Economic Association, vol. 1(5), pp. 1123-1175.

Smets, F. and R. Wouters (2007). Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach. American Economic Review vol. 97(3), pp. 586-606.