The Annealed Leap-Point MCMC Sampler (ALPS) for multi-modal posterior distributions

Matt Moores

Lecturer in Statistical Science
Centre for Environmental Informatics, NIASRA
University of Wollongong, Australia

Docent in Computational Statistics
School of Engineering Science
LUT University, Finland

Acknowledgements

The following is joint work with

- Dr Nicholas Tawn
- Prof. Gareth Roberts

Department of Statistics, University of Warwick, UK.

Tawn, N.G.; Moores, M.T. & Roberts, G.O. (2021) *arXiv preprint* arXiv:2112.12908

Given an (unnormalized) target distribution, $\pi(\boldsymbol{\theta})$, with parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$,

1. Choose an initial value, $\boldsymbol{\theta}^{(0)}$ (e.g. from the *prior*)
2. For iterations $t = 1, \ldots, T$ do
   1. Propose new parameters, $\boldsymbol{\theta}' \sim q\left(\cdot \mid \boldsymbol{\theta}^{(t-1)}\right)$, e.g. $\mathcal{N}\left(\boldsymbol{\theta}^{(t-1)}, \Sigma\right)$
   2. Evaluate the Metropolis ratio,

$$\rho_t = \frac{\pi\left(\boldsymbol{\theta}'\right) \quad q\left(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}'\right)}{\pi\left(\boldsymbol{\theta}^{(t-1)}\right) \; q\left(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{(t-1)}\right)}$$

   3. Accept $\boldsymbol{\theta}'$ with probability $\min(1, \rho_t)$, so that $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}'$, or else reject so that $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$

Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953) *J. Chem. Phys.* **21** (6): 1087–1092.

When does $\pi(\boldsymbol{\theta})$ have multiple local optima?

- Mixture models (e.g. mixture of Gaussians)
- Curved exponential family (e.g. SUR regression)
- Can sometimes indicate lack of identifiability of $\boldsymbol{\theta}$ or prior-posterior mismatch
- Can arise in ill-posed inverse problems

Seemingly-Unrelated Regression

- The SUR model was introduced by Zellner (1962) for panel data.
- Until recently, the SUR likelihood was assumed to be log-concave.

SUR models involve a system of M linear regression equations, one for each response, $m = 1, \ldots, M$:

$$\vec{y}_m = \mathrm{X}_m \vec{\theta}_m + \vec{\epsilon}_m, \tag{1}$$

These vectors can be stacked on top of each other, resulting in:

$$\begin{bmatrix} \vec{y}_1 \\ \vec{y}_2 \\ \vdots \\ \vec{y}_M \end{bmatrix} = \begin{bmatrix} \mathrm{X}_1 & & \cdots & 0 \\ & \mathrm{X}_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \mathrm{X}_M \end{bmatrix} \begin{bmatrix} \vec{\theta}_1 \\ \vec{\theta}_2 \\ \vdots \\ \vec{\theta}_M \end{bmatrix} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \\ \vdots \\ \vec{\epsilon}_M \end{bmatrix}. \tag{2}$$

Zellner, A. (1962) *J. Am. Stat. Assoc.* **57**: 348–368.

The errors $\vec{\epsilon}_m$ are assumed to be jointly multivariate normal,

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\vec{0}, \boldsymbol{\Sigma}_\epsilon \otimes \mathsf{I}\right),$$

where I is a $N \times N$ identity matrix and $\boldsymbol{\Sigma}_\epsilon$ is a $M \times M$ variance-covariance matrix with entries $\sigma^2_{\ell,m}$.

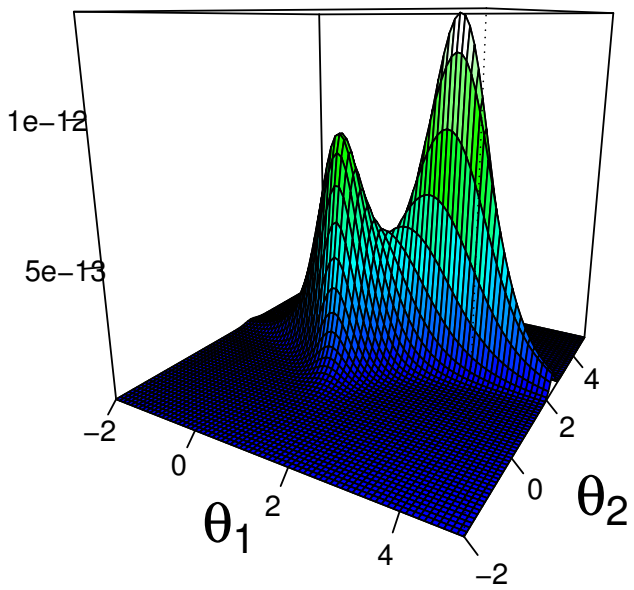
In the bivariate case when $M = 2$, the model can be written as:

$$\begin{bmatrix} \vec{y}_1 \\ \vec{y}_2 \end{bmatrix} = \begin{bmatrix} \vec{x}_1 & \vec{0} \\ \vec{0} & \vec{x}_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix}.$$

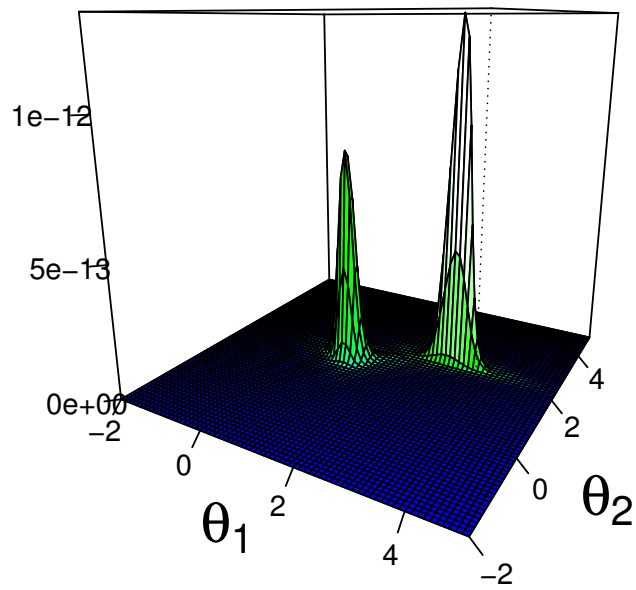
Drton & Richardson (2004) demonstrated that the likelihood $\pi(\boldsymbol{\theta})$ can be multimodal.

Drton, M. & Richardson, T.S. (2004) *Biometrika* **91**: 383–392.

Profile Likelihood for SUR



(a) $\pi(\boldsymbol{\theta})$



(b) $\pi(\boldsymbol{\theta})^{10}$

Parallel Tempering

The parallel tempering algorithm, also known as MCMCMC or MC^3, was previously the state of the art for statistical inference from multi-modal distributions.

The idea is to run multiple MCMC chains, each at different *temperatures* $\beta_j \in (0, 1]$. For example, $\boldsymbol{\beta} = (1, 0.5, 0.333, 0.25, 0.2, 0.167)^T$

Each chain is independently initialized at a different starting value $\boldsymbol{\theta}_j^{(0)}$

At each iteration, 2 different types of proposals can be considered:

- Within-temperature move targeting $\pi(\boldsymbol{\theta})^{\beta_j}$
- Swap move between neighbouring temperatures (either $j, j+1$ or $j, j-1$)

Geyer, C.J. (1991) *Comput. Sci. & Stat.* **23**: 156–163.
Neal, R.M. (1996) *Stat. & Comp.* **6**: 353–366.

Torpid Mixing

Woodard et al. (2009a,b) proved that both parallel tempering and simulated tempering can converge exponentially slowly in dimension d (in terms of the "spectral gap") for some multi-modal distributions:

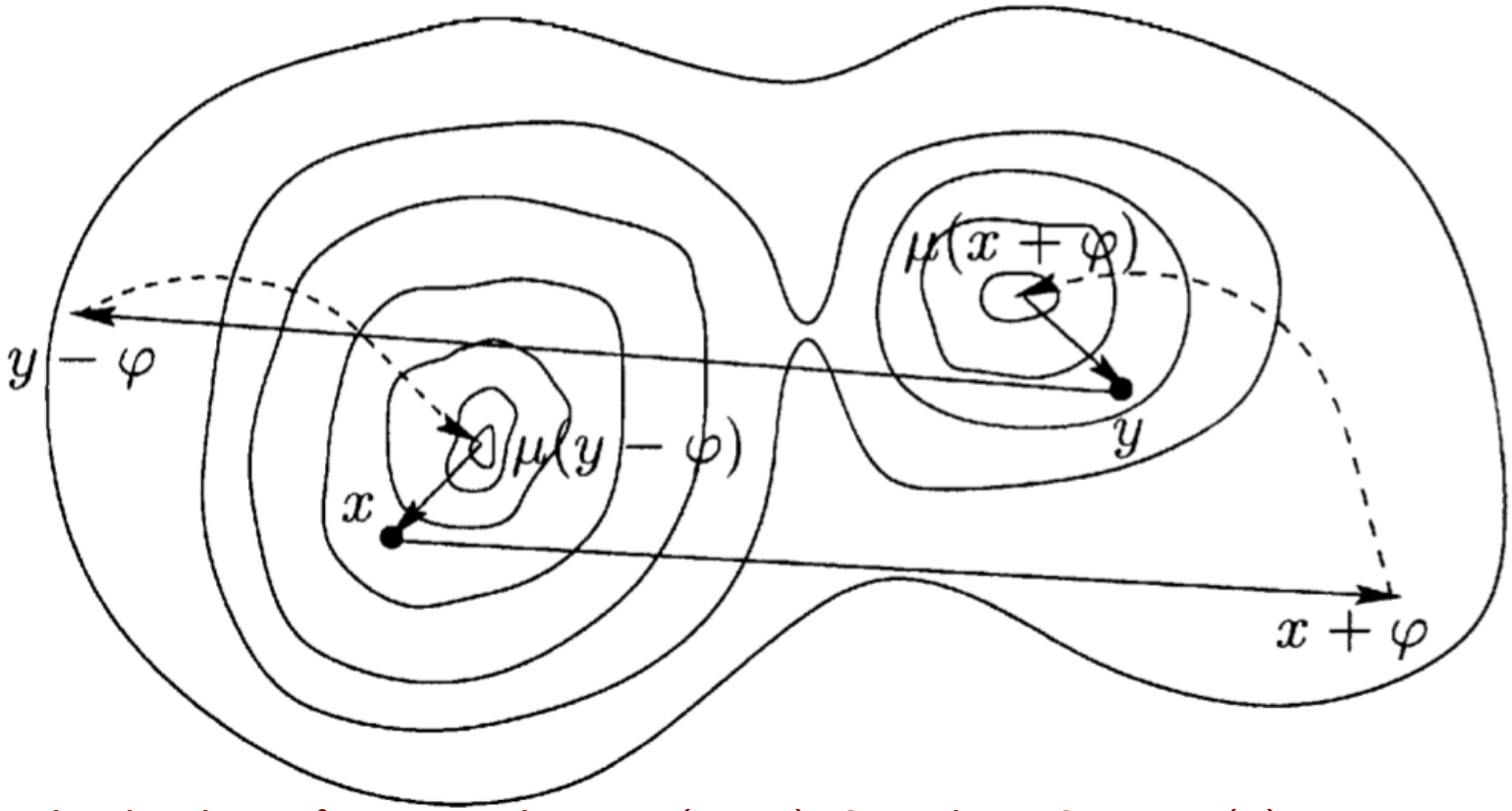
- If the **conductance** is exponentially decreasing with d,
- and the **persistence** is exponentially decreasing with d,
- and the **overlap** is exponentially decreasing with d,

then parallel and simulated tempering are torpidly mixing (Corollary 3.2).

Woodard, D.B.; Schmidler, S.C. & Huber, M. (2009a) *Ann. Appl. Prob.* 617–640.
Woodard, D.B.; Schmidler, S.C. & Huber, M. (2009b) *Electr. J. Prob.* **14**: 780–804.

Mode Jumping

Tjelmeland & Hegstad (2001) introduced mode-jumping proposals for MCMC:



Tjelmeland, H. & Hegstad, B.K. (2001) *Scand. J. Stat.* **28**(1): 205–223.
Tjelmeland, H. & Eidsvik, J. (2004) *J. R. Stat. Soc.* (Ser. B) **66**: 411–427.

Drawbacks of Mode Jumping

Although very promising, the algorithm of Tjelmeland & Hegstad (2001) has a few drawbacks:

- First, you need to *find* the modes
- To preserve reversability, the algorithm *forgets* the modes it has found
- Due to Laplace approximation, the modes need to be very close to Gaussian in shape (i.e. not skewed or heavy-tailed)

Our ALPS algorithm addresses these deficiencies in the previous methods.

ALPS combines the following features:

- The **exploration component** or "mode hunting" uses a random walk at a "hot" temperature (e.g. $\beta \propto d^{-1}$) combined with optimization (BFGS) to locate the modes
- The **annealing component** uses *parallel* chains at a sequence of "cold" temperatures, e.g. $\boldsymbol{\beta} = (1, 1.04, 1.10, 1.25, 2, 4, 8, 16, 32)^T$, with *quantile* tempering (QuanTA) & *hessian-adjusted* tempering (HAT) & c.

At the coldest temperature, the modes are much closer to Gaussian:

- this is where **mode-jumping** proposals are used.

Temperature-swapping proposals are used to propagate information between the chains, down to the target temperature ($\beta = 1$).

Example 1: 2D SUR

Drton & Richardson (2004) had an example dataset with $N = 8$ observations and $M = 2$ regression equations, each with only $J = 1$ coefficients.

Two modes:

$$M_1\text{: } (0.78,\ 1.54)^T \quad — \quad \log\{\pi(M_1)\} = -27.722$$
$$M_2\text{: } (2.76,\ 2.50)^T \quad — \quad \log\{\pi(M_2)\} = -27.349$$

and a saddle point at $\boldsymbol{\theta} = (1.62, 2.03)^T$

Example 1: systemfit

The iterated, generalised least squares (GLS) algorithm of Zellner (1962) is implemented in the R package "`systemfit`" (Henningsen et al., 2007):

```
library(systemfit)
Drton ← data.frame(cbind(Y,X))
names(Drton) ← c("Y1","Y2","X1","X2")
eq1 ← Y1 ~ -1 + X1
eq2 ← Y2 ~ -1 + X2
eqSys ← list(first = eq1, second = eq2)
fitsur ← systemfit(eqSys, method="SUR", data=Drton)
```

after 38 iterations, the algorithm converges to $\hat{\boldsymbol{\theta}} = (0.777, 1.547)^T$, which is the sub-optimal local mode, M_1.

This algorithm has failed to find the global optimum, M_2.

Henningsen, A. & Hamann, J.D. (2007) *J. Stat. Soft.* **23**: 1–40.

Example 1: ALPS

For this 2D example, we ran ALPS with 3 temperatures $\boldsymbol{\beta} = (1, \sqrt{10}, 10)^T$ plus the hot state $\beta = 0.5$.

It took 2.6 seconds to perform 10,000 iterations in total.

Currently, the within-temperature Gaussian random walk proposals and the position-dependent proposals need to be tuned manually.

There are many other parameters in the R package for fine-tuning ALPS, but most of these have sensible defaults.

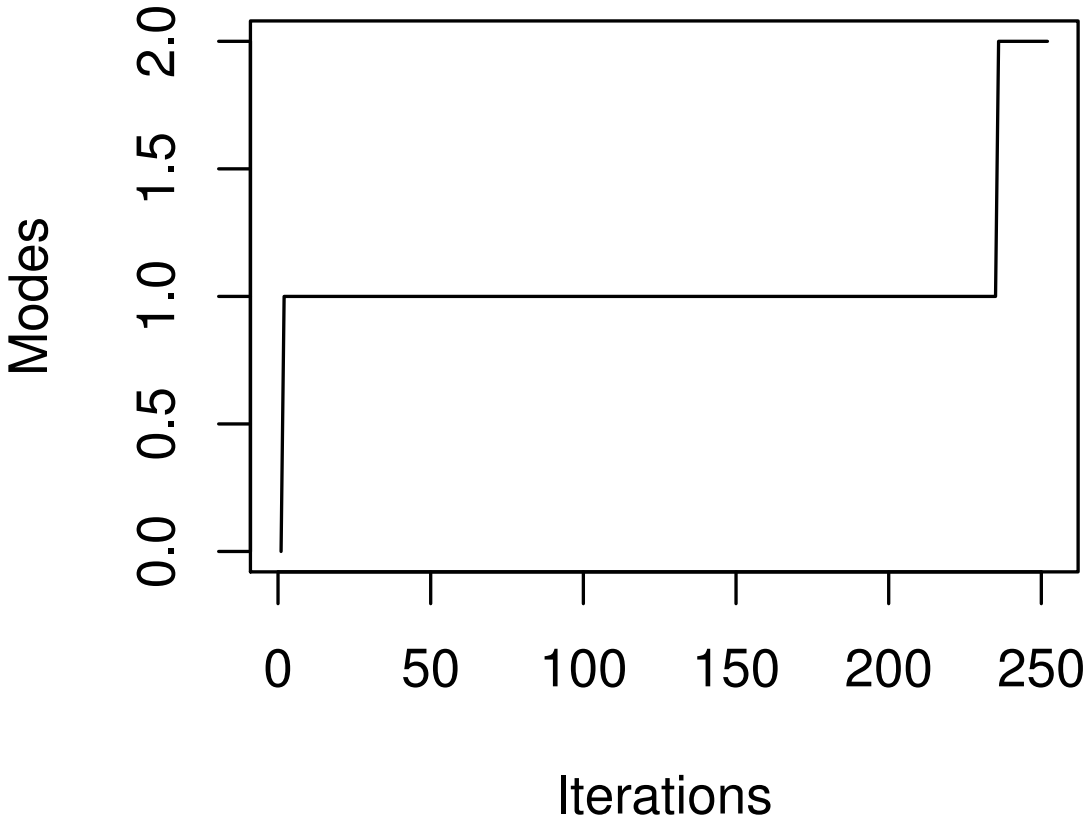
Example 1: R package

```
library(ALPS)
dd ← 2
scaler ← 0.7
TmpSch ← c(1,sqrt(10),10)
sdprops ← 2.38*scaler/(sqrt(dd)*sqrt(TmpSch))
ShallJump ← c(rep(0,(length(TmpSch)−2)),0.7,0.7)
HotTem = 1/dd
HotTuners ← list(prob=0.1,HotTem=HotTem,Hotsig=2.7/sqr
   Type_opt="BFGS", Toler=1e−8, max_lts_opt=2000, p_res

sur2D  ← ALPS_seq(Targ=profLikeSUR,centres=0,n=2500,w=
   Pjump=ShallJump,sigs=sdprops,sigsPD=6/TmpSch,
   PDRWM=0.8,dimension=dd,inits=rep(0,dd),maxmodes=2,
   QuanTLocs = (levsm1),Weight_Preservation=T,MahalUsed=
   newmode_attempt=HotTuners)
```

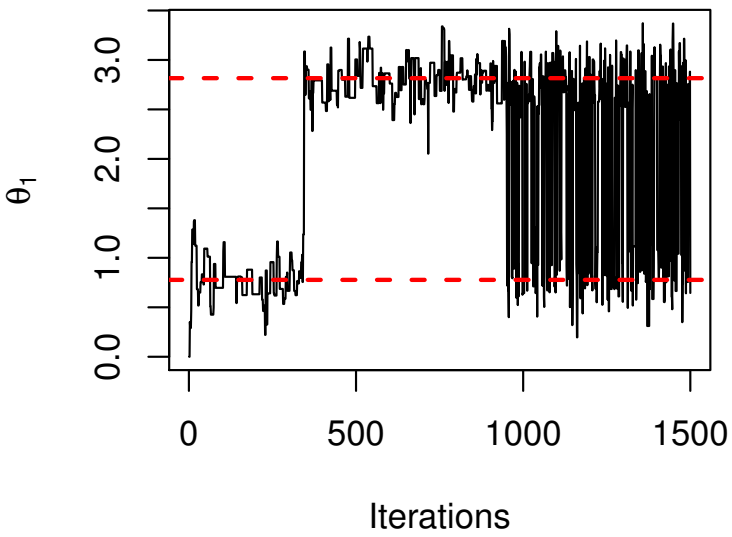
Example 1: exploration component

The hot-temperature "mode hunting" takes 236 iterations

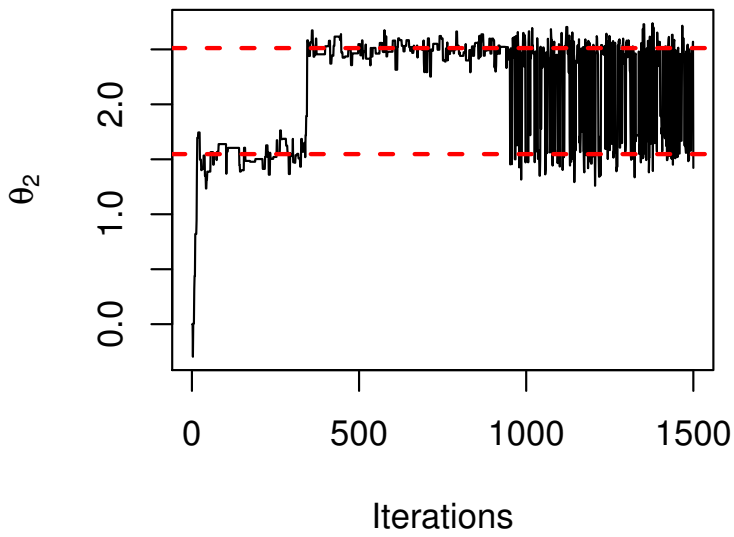


Example 1: MCMC chains

236 iterations of mode hunting corresponds to 944 iterations of annealing:



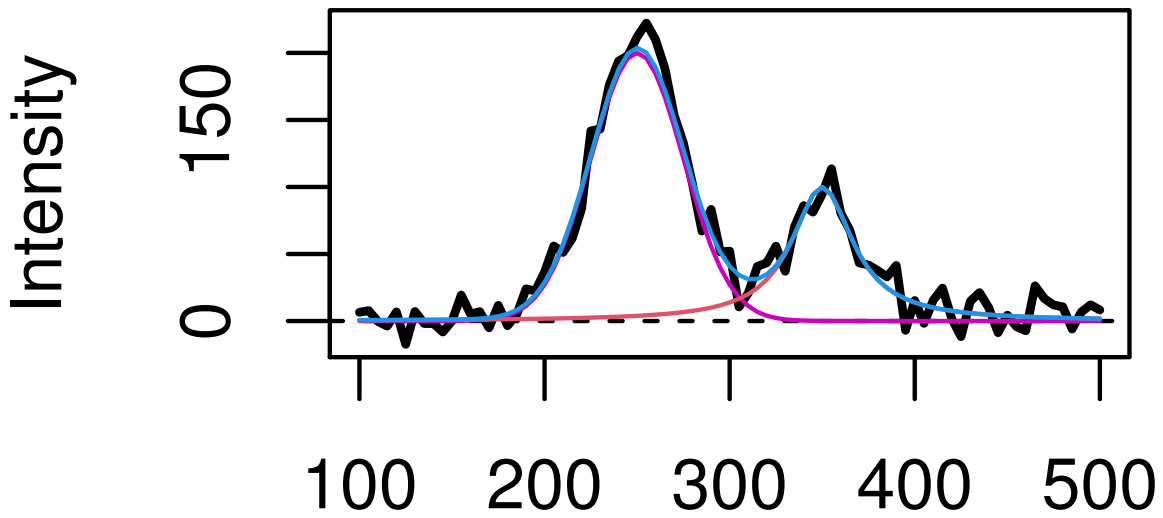
(a) θ_1



(b) θ_2

Example 2: spectroscopy

Consider a spectrum that has one Gaussian peak at $\nu_1 = 250$ cm^{-1} and one Lorentzian peak at $\nu_2 = 350$ cm^{-1}. This spectrum is observed with zero-mean, additive Gaussian noise ($\sigma = 10$):



Example 2: multi-modality

The likelihood $\pi(\boldsymbol{\theta})$ has two modes, with $\boldsymbol{\theta} \in \mathbb{R}_+^6$:

$$M_1\text{: } (352.70,\ 16.97,\ 82.67,\ 251.63,\ 24.51,\ 230.35)^T$$
$$M_2\text{: } (250.47,\ 24.16,\ 210.93,\ 350.66,\ 21.09,\ 100.06)^T$$

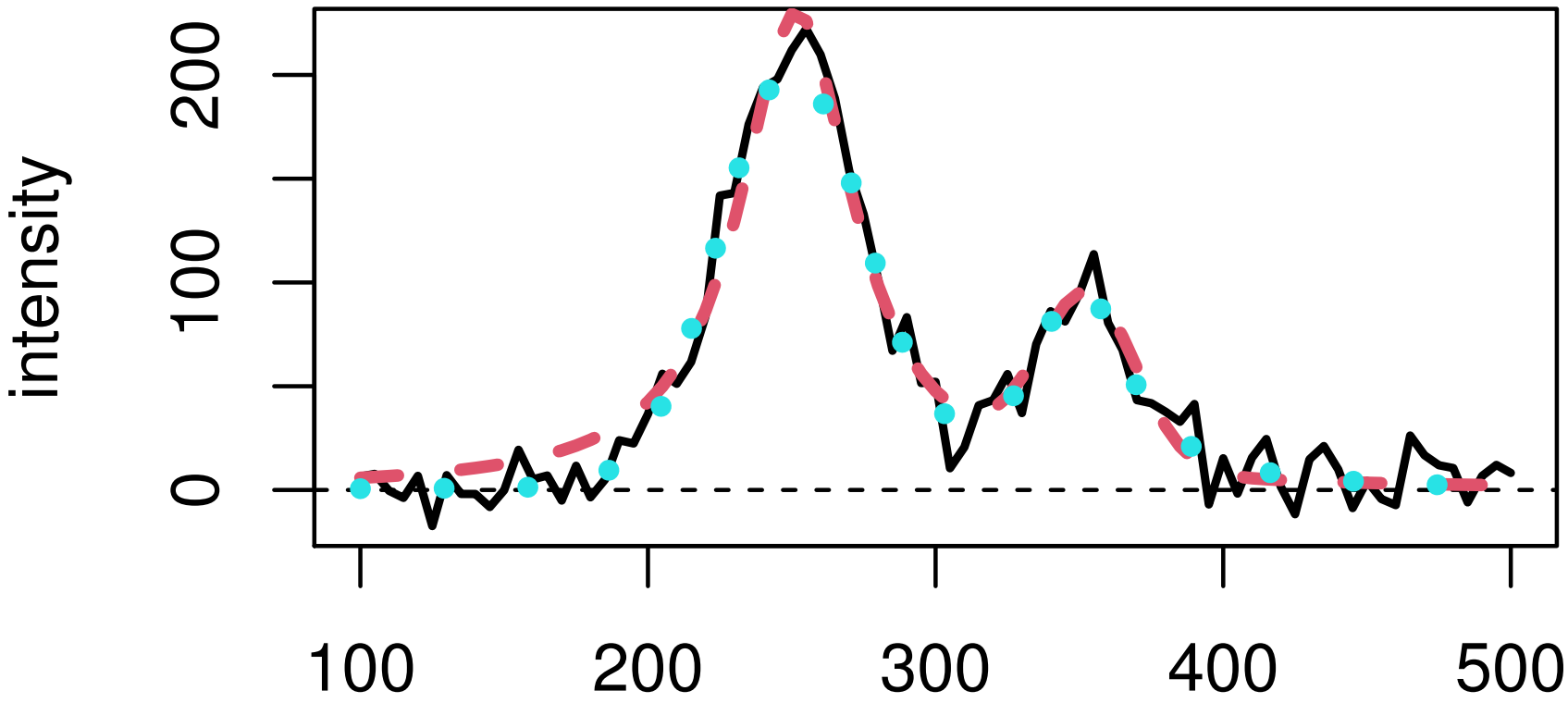
M_2 is very close to the true parameter values for the peak locations, broadening, and amplitudes.

In M_1, the peak locations are swapped around. However, the broadening parameters and the amplitudes are also incorrect.

Depending on where the Maximum Likelihood algorithm is initialized, it will either converge to M_1 or M_2.

The more peaks there are, the more modes there will be in the likelihood!

Example 2: fitted model



Example 2: ALPS

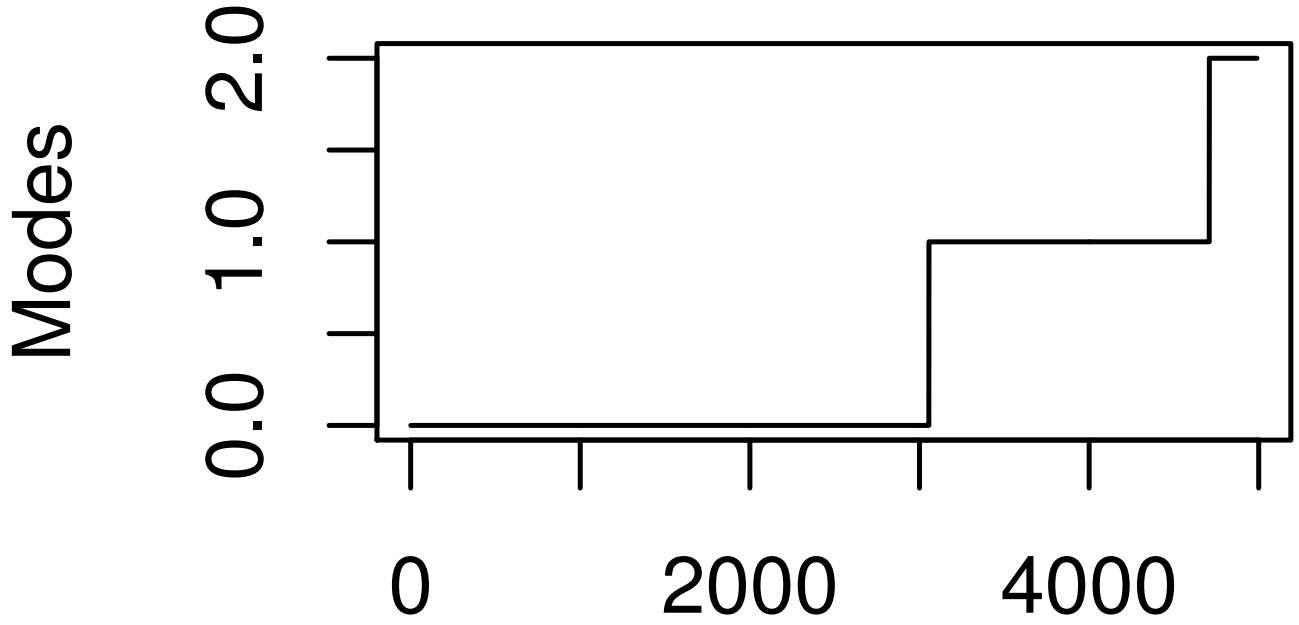
For this 6D example, we ran ALPS with 6 temperatures $\boldsymbol{\beta} = (1.000, 1.216, 1.341, 1.800, 3.240, 5.832)^T$ plus the hot state $\beta = 0.01$.

Mode jumping was performed at the coldest 3 temperatures.

It took 135 seconds to perform 200,000 iterations in total.
The first 20,000 were discarded as burn-in.

Example 2: exploration component

The hot-temperature "mode hunting" took 4,708 iterations



Example 2: results

```
Jump Rate:
[1]       NaN       NaN       NaN 0.9269 0.9469 0.9609
Within Level Simple RWMM Acceptance Rates:
[1] 0.2518 0.2333 0.2308 0.2325 0.2600 0.2464
Within Level PDRWMM Acceptance Rates:
[1] 0.8038 0.7961 0.7898 0.7867 0.7931 0.8010
Swap Level Acceptance Rates:
[1] 0.6881 0.9604 0.9535 0.9488 0.9571
Hot Level Acceptance Rates:
[1] 0.3175
Elapsed time (seconds):
[1] 134.76
```

Example 3: investment demand

Panel data from U.S. manufacturing firms (Grunfeld, 1958):

$$\vec{y}_m = \vec{x}_{1,m}\theta_{1,m} + \vec{x}_{2,m}\theta_{2,m} + \vec{x}_{3,m}\theta_{3,m} + \vec{\epsilon}_m,$$

where $y_{i,m}$ is the gross investment by firm m during the ith year

- $N = 15$ years of data (1935 to 1949)
- $J = 3$ covariates:
  intercept $\vec{x}_{1,m} = 1$; market value $\vec{x}_{2,m}$; & capital stock $\vec{x}_{3,m}$
- $M = 5$ firms:
  General Motors, Chrysler, General Electric, Westinghouse, & US Steel

Grunfeld, Y. (1958) PhD thesis, University of Chicago.

Example 3: systemfit

The GLS algorithm (Zellner, 1962) takes 52 iterations to converge:

$\hat{\boldsymbol{\theta}} =$

$(41.2, 89.3, 188.1, 12.8, 64.0, 140.7, -46.1, 56.3, 92.3, 7.9, 51.4, -34.1, 107.3, 126.2, 19.1)^T$

```
data("GrunfeldGreene")
GGPanel <- pdata.frame(GrunfeldGreene, c("firm","year"))
formulaGrunfeld <- invest ~ value + capital
eqSys <- list(gm = eq1, chry = eq2, ge = eq3,
    west = eq4, steel = eq5)
fitsur <- systemfit(eqSys, method="SUR", data=GGPanel,
    control = systemfit.control(maxiter = 100))
```

Example 3: ALPS

For $\boldsymbol{\theta} \in \mathbb{R}^{15}$, we use 9 temperature levels
$\boldsymbol{\beta} = (1.00, 1.04, 1.10, 1.25, 2, 4, 8, 16, 32)^T$ with hot state $\beta = 1/15$.

We needed to truncate the annealed HAT targets to avoid problems with machine precision at the coldest temperatures.

It took 31 minutes to perform 200,000 iterations in total.

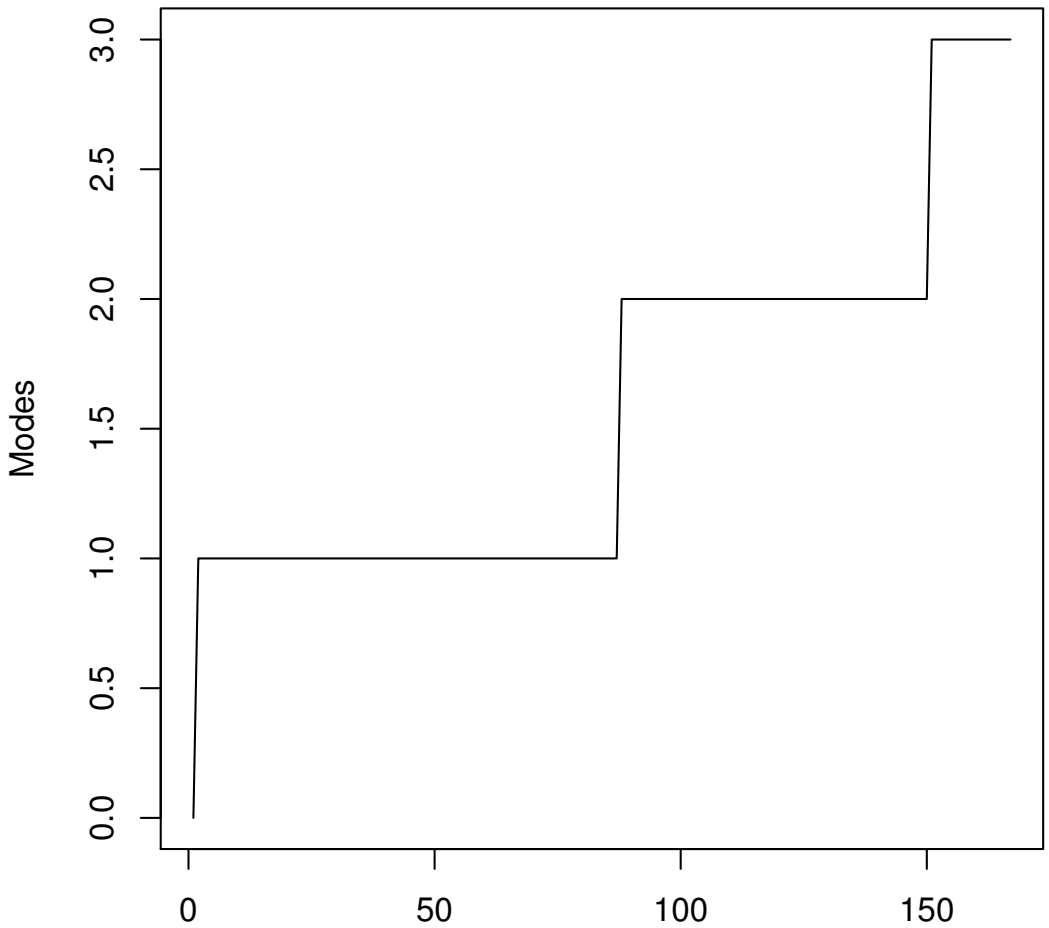
There were 3 modes found, with
$\pi(M_1) = -426.79$, $\pi(M_2) = -515.52$, & $\pi(M_3) = -577.23$.

M_1 corresponds to the same values of $\boldsymbol{\theta}$ found by GLS.

M_3 had negligible probability mass.

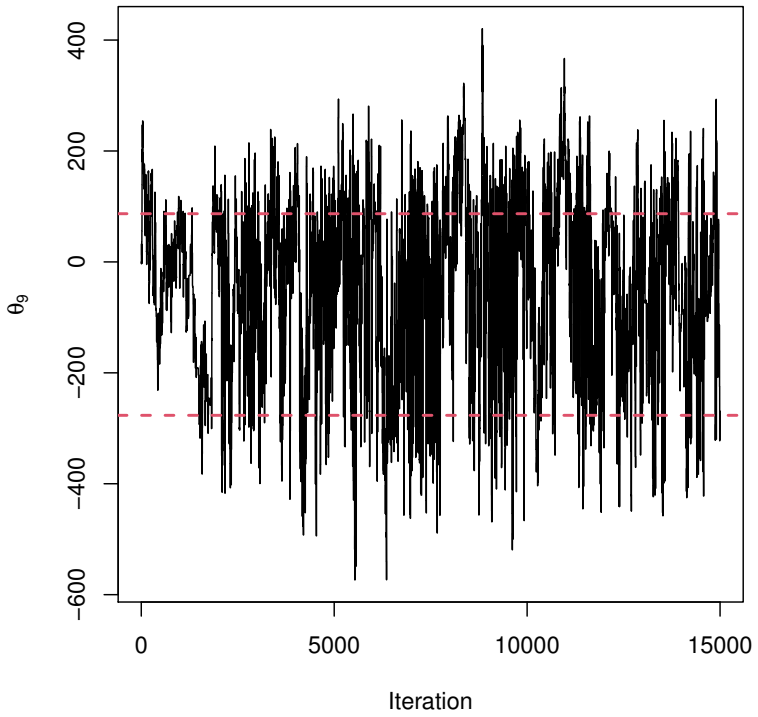
Example 3: exploration component

The hot-temperature "mode hunting" took 155 iterations to find 3 modes:

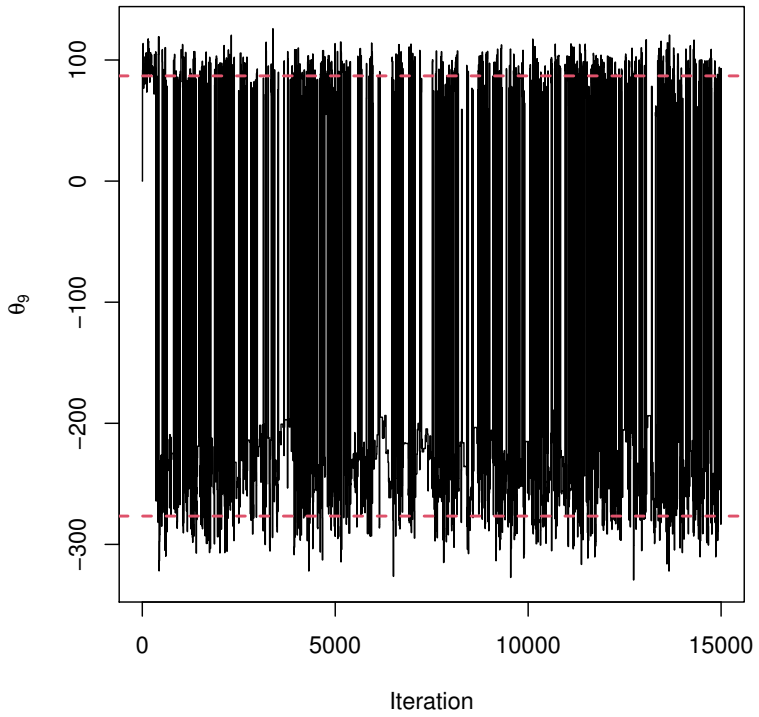


Example 3: traceplots

155 iterations of mode hunting corresponds to 620 iterations of annealing:



(a) θ_9 at $\pi(\boldsymbol{\theta})^1$



(b) θ_9 at $\pi(\boldsymbol{\theta})^{32}$

Conclusion

ALPS can be used to explore multi-modal posterior distributions:

- Curve fitting in spectroscopy
- SUR model for panel data in econometrics

R package coming soon.

Current and future work:

- ALPS for intractable likelihoods (e.g. pseudo-marginal)
- ALPS for inverse problems

For Further Reading I


Nicholas Tawn, Matt Moores & Gareth Roberts
Annealed Leap-Point Sampler for Multimodal Target Distributions.
arXiv preprint arXiv:2112.12908


Gareth Roberts, Jeff Rosenthal & Nicholas Tawn
Skew Brownian Motion and Complexity of the ALPS Algorithm.
Advances in Applied Probability, in press, 2022.


Dawn Woodard, Scott Schmidler & Mark Huber
Sufficient Conditions for Torpid Mixing of Parallel and Simulated Tempering.
Electronic Journal of Probability, **14**: 780–804, 2009.


Håkon Tjelmeland & Bjørn Kåre Hegstad
Mode Jumping Proposals in MCMC.
Scandanavian Journal of Statistics, **28**(1): 205–223, 2001.