

Nonparametric Statistics

What does nonparametric mean?

- Nonparametric statistics try to minimise the number of initial assumptions (everything isn't normally distributed), so that our analysis doesn't reach incorrect conclusions (as much).
- Nonparametric statistics are more basic / intuitive, and require very basic math (primary school level).
- Statistics are done using only signs and ranks (we'll get to that).

Pros & Cons

1. Very few assumptions, fewer incorrect conclusions.
2. Basic math.
3. Signs and ranks

1. We don't get a distribution, i.e. little output insight.
2. Outputs are also difficult to interpret.
3. Large loss of information
4. Computation becomes complicated for very large samples.

Let's look at an example...

Nahm F. S. (2016). Nonparametric statistical tests for the continuous data: the basic concept and the practical use. *Korean journal of anesthesiology*, 69(1), 8–14.

<https://doi.org/10.4097/kjae.2016.69.1.8>

Nonparametric statistical tests for the continuous data: the basic concept and the practical use

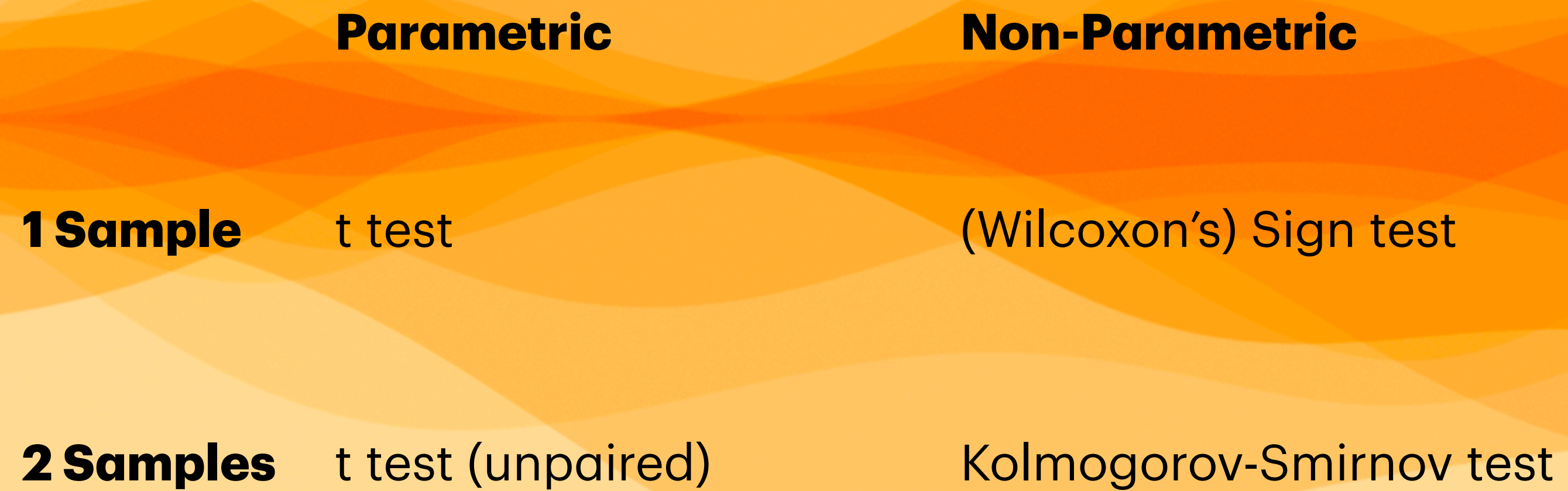
Francis Sahngun Nahm

Department of Anesthesiology and Pain Medicine, Seoul National University Bundang Hospital, Seongnam, Korea

Conventional statistical tests are usually called parametric tests. Parametric tests are used more frequently than non-parametric tests in many medical articles, because most of the medical researchers are familiar with and the statistical software packages strongly support parametric tests. Parametric tests require important assumption; assumption of normality which means that distribution of sample means is normally distributed. However, parametric test can be misleading when this assumption is not satisfied. In this circumstance, nonparametric tests are the alternative methods available, because they do not required the normality assumption. Nonparametric tests are the statistical methods based on signs and ranks. In this article, we will discuss about the basic concepts and practical use of nonparametric tests for the guide to the proper use.

Key Words: Data interpretation, Investigative technique, Nonparametric statistics, Statistical data analysis.

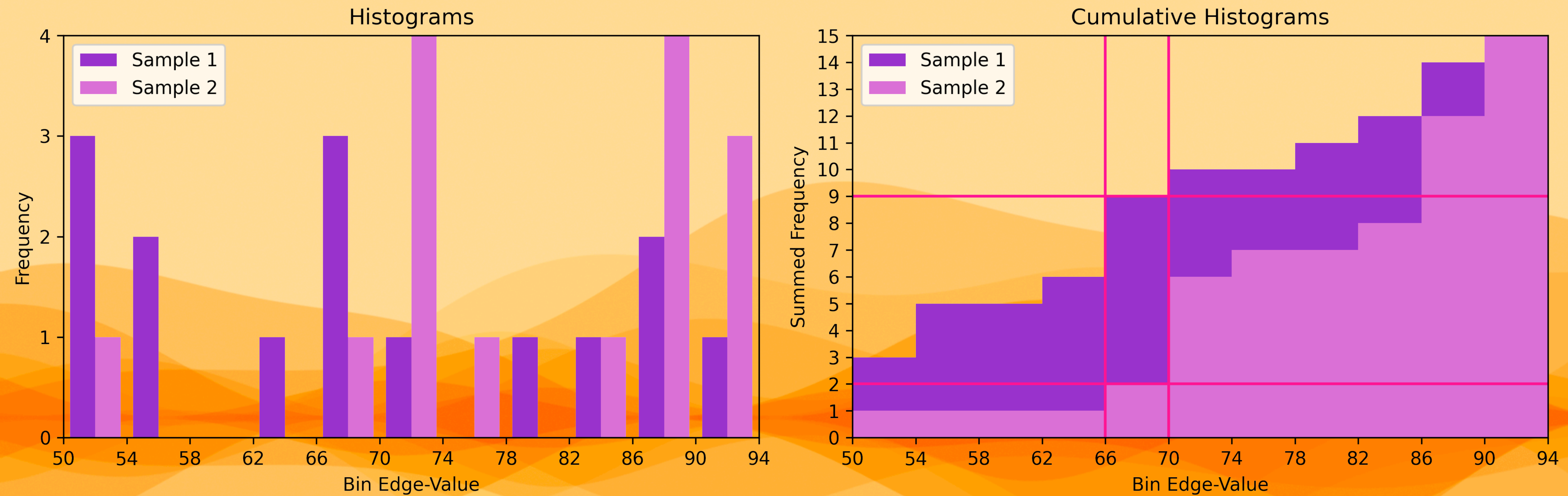
Statistical Tests



Statistical Tests for Two Samples

Null hypothesis: samples are from the same population

| | | | | | | | | | | | | | | | |
|-----------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Sample 1 | 53 | 87 | 71 | 64 | 78 | 66 | 52 | 54 | 50 | 91 | 55 | 86 | 69 | 82 | 68 |
| Sample 2 | 88 | 84 | 72 | 91 | 89 | 68 | 73 | 52 | 71 | 93 | 87 | 92 | 76 | 72 | 86 |




Kolmogorov-Smirnov test: $\max(|CDF_1 - CDF_2|) = \frac{7}{15} = 0.467 > 0.4042 (\alpha = 0.01)$

(Unpaired) t test: $\frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 2.396 > 2.048 (\alpha = 0.05, df = 28)$


Uses for Nonparametric Statistics?

Google “nonparametric statistics”

 WallStreetMojo
<https://www.wallstreetmojo.com> › Statistics Guides


Nonparametric Statistics - What Is It, Examples, Vs Parametric

Key Takeaways · **Nonparametric statistics** do not require assumptions about the data distribution or population parameters, making them more flexible than ...

 Investopedia
<https://www.investopedia.com> › ... › Math and Statistics


What Is Nonparametric Method? Analysis Vs. Parametric ...

A histogram is an example of a **nonparametric** estimate of a probability distribution. In contrast, well-known **statistical** methods such as ANOVA, Pearson's ...

 National Institutes of Health (NIH) (.gov)
<https://www.ncbi.nlm.nih.gov> › articles › PMC7643794


Nonparametric Statistical Methods in Medical Research

by P Schober · 2020 · Cited by 47 — **Nonparametric** methods are commonly used when **data** distribution assumptions of parametric **tests** are not met. In practice, researchers often ...

 Health Knowledge
<https://www.healthknowledge.org.uk> › research-methods


Parametric and Non-parametric tests for comparing two or ...

Parametric tests are those that make assumptions about the parameters of the population distribution from which the sample is drawn. This is often the ...

 Corporate Finance Institute
<https://corporatefinanceinstitute.com> › Resources


Nonparametric Tests - Overview, Reasons to Use, Types

In statistics, nonparametric tests are **methods of statistical analysis that do not require a distribution to meet the required** assumptions to be analyzed.

 Corporate Finance Institute
<https://corporatefinanceinstitute.com> › Resources


Nonparametric Statistics - Definition, Types, Examples

Nonparametric statistics **makes statistical inferences without regard to any underlying distribution**. It fits a normal distribution under no assumptions.

 Investopedia
<https://www.investopedia.com> › ... › Financial Analysis


Nonparametric Statistics: Overview, Types, and Examples

Nonparametric statistics refer to a **statistical method in which the data is not required to fit a normal distribution**. Rankings should not change.

 Machine Learning Mastery
<https://machinelearningmastery.com> › Blog

A Gentle Introduction to Nonparametric Statistics

10 Nov 2019 — In the case of ordinal or interval data, **nonparametric statistics are the only type of statistics that can be used**. For real-valued data, ...

 Mayo Foundation for Medical Education and Research
<https://www.mayo.edu> › research › doc-20408960 PDF


Parametric and Nonparametric: Demystifying the Terms

Nonparametric statistical procedures rely **on no or few assumptions about the shape or parameters of the population distribution from which the sample was drawn**.

 BMC Medical Research Methodology
<https://bmcmmedresmethodol.biomedcentral.com> › articles

Parametric versus non-parametric statistics in the analysis of ...

by AJ Vickers · 2005 · Cited by 465 — It has generally been argued that **parametric statistics** should not be applied to **data** with **non-normal** distributions.

 National Institutes of Health (NIH) (.gov)
<https://www.ncbi.nlm.nih.gov> › articles › PMC4754273

Nonparametric statistical tests for the continuous data

by FS Nahm · 2016 · Cited by 350 — Abstract. Conventional statistical tests are usually called parametric tests. Parametric tests are used more frequently than **nonparametric tests** in...

Uses for Nonparametric Statistics?

Google “nonparametric statistics”

- Finance
- Medicine
- Machine learning / statistics
- Discrete data / populations in general

So how are nonparametric statistics relevant for us?

Bayesian Nonparametric MCMC

Bayesian Nonparametric MCMC

See: <https://www.stats.ox.ac.uk/~teh/teaching/npbayes/mlss2011F.pdf>

A nonparametric model is:

- A really large parametric model.
- A parametric model where the number of parameters increases with the amount of data.
- A family of distributions \rightarrow sampling over function space, rather than parameter space.

By having the ability to increase the number of parameters / functions, we combat under-fitting.

... But we still use a Normal distribution in our prior.

An ambiguous number of parameters / functions sounds scary. How can we ensure that our algorithm converges, and that we sample efficiently?

MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster

S. L. Cotter, G. O. Roberts, A. M. Stuart and D. White

Abstract. Many problems arising in applications result in the need to probe a probability distribution for functions. Examples include Bayesian nonparametric statistics and conditioned diffusion processes. Standard MCMC algorithms typically become arbitrarily slow under the mesh refinement dictated by nonparametric description of the unknown function. We describe an approach to modifying a whole range of MCMC methods, applicable whenever the target measure has density with respect to a Gaussian process or Gaussian random field reference measure, which ensures that their speed of convergence is robust under mesh refinement.

“Since in nonparametric Bayesian problems the unknown of interest (a function) naturally lies in an infinite-dimensional space, numerical schemes for evaluating posterior distributions almost always rely on some kind of finite-dimensional approximation or truncation to a parameter space of dimension d_u , say... The larger d_u is, the better the approximation to the infinite-dimensional true model becomes. However, off-the-shelf MCMC methodology usually suffers from a curse of dimensionality so that the numbers of iterations required for these methods to converge diverges with d_u ”

MCMC Methods for Functions

We can modify existing and well-understood algorithms such that they still perform well as we increase the number of parameters.

- Metropolis-Hastings (MH) \rightarrow preconditioned Crank-Nickolson (pCN)

$$(1.2) \quad I(u) = \Phi(u) + \frac{1}{2} \|\mathcal{C}^{-1/2} u\|^2$$

and consider the following version of the standard random walk method:

- Set $k = 0$ and pick $u^{(0)}$.
- Propose $v^{(k)} = u^{(k)} + \beta \xi^{(k)}, \xi^{(k)} \sim N(0, \mathcal{C})$.
- Set $u^{(k+1)} = v^{(k)}$ with probability $a(u^{(k)}, v^{(k)})$.
- Set $u^{(k+1)} = u^{(k)}$ otherwise.
- $k \rightarrow k + 1$.

The acceptance probability is defined as

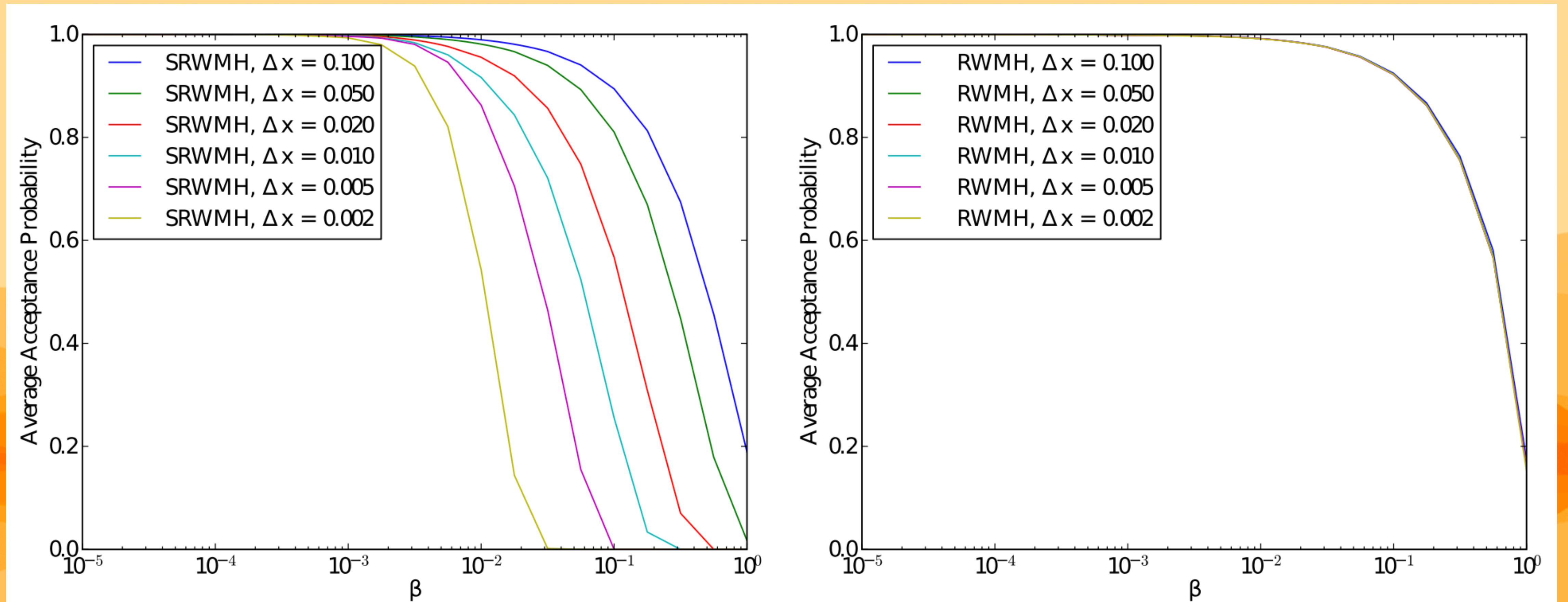
$$a(u, v) = \min\{1, \exp(I(u) - I(v))\}.$$

The pCN method is the following modification of the standard random walk method:

- Set $k = 0$ and pick $u^{(0)}$.
- Propose $v^{(k)} = \sqrt{(1 - \beta^2)} u^{(k)} + \beta \xi^{(k)}, \xi^{(k)} \sim N(0, \mathcal{C})$.
- Set $u^{(k+1)} = v^{(k)}$ with probability $a(u^{(k)}, v^{(k)})$.
- Set $u^{(k+1)} = u^{(k)}$ otherwise.
- $k \rightarrow k + 1$.

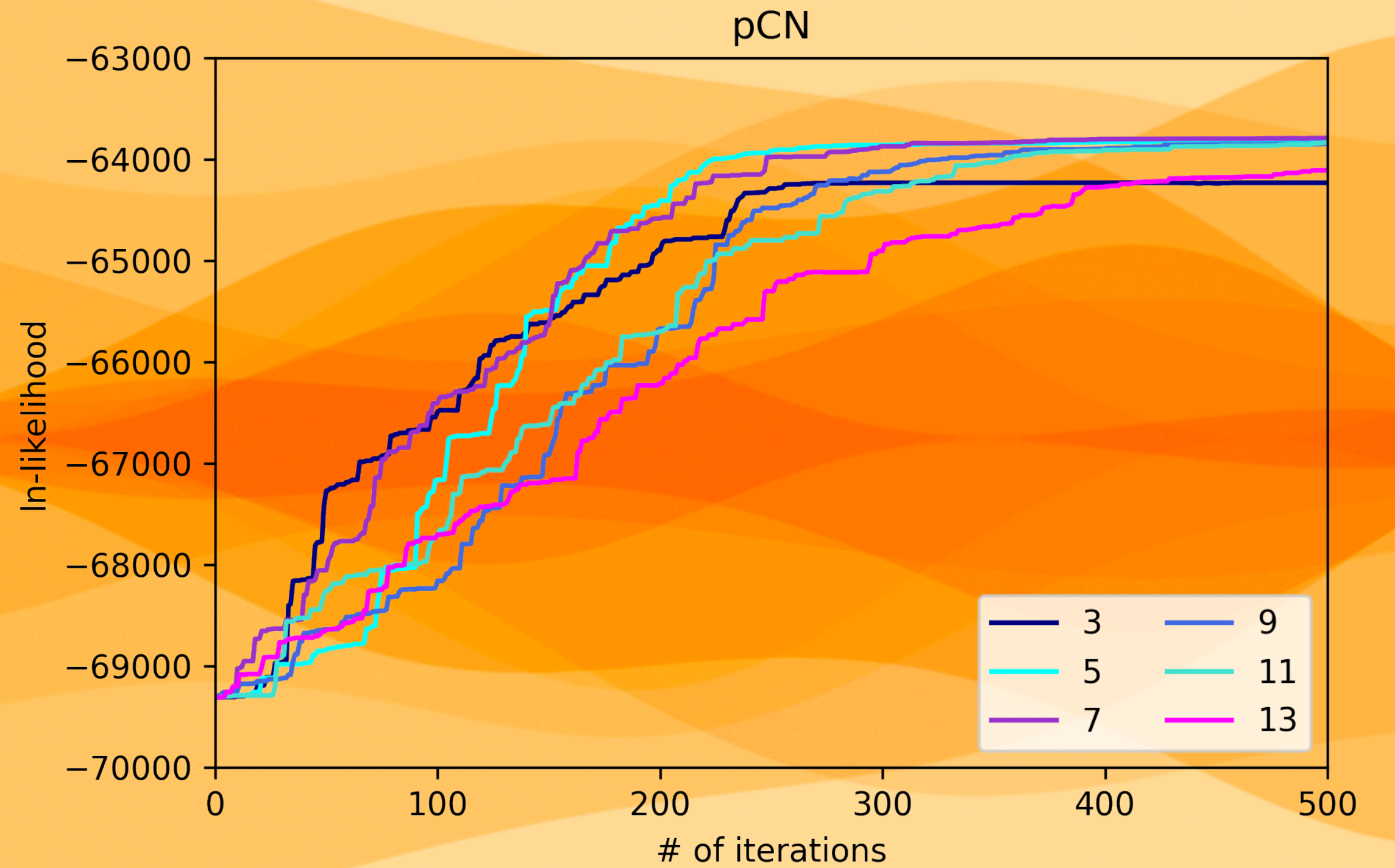
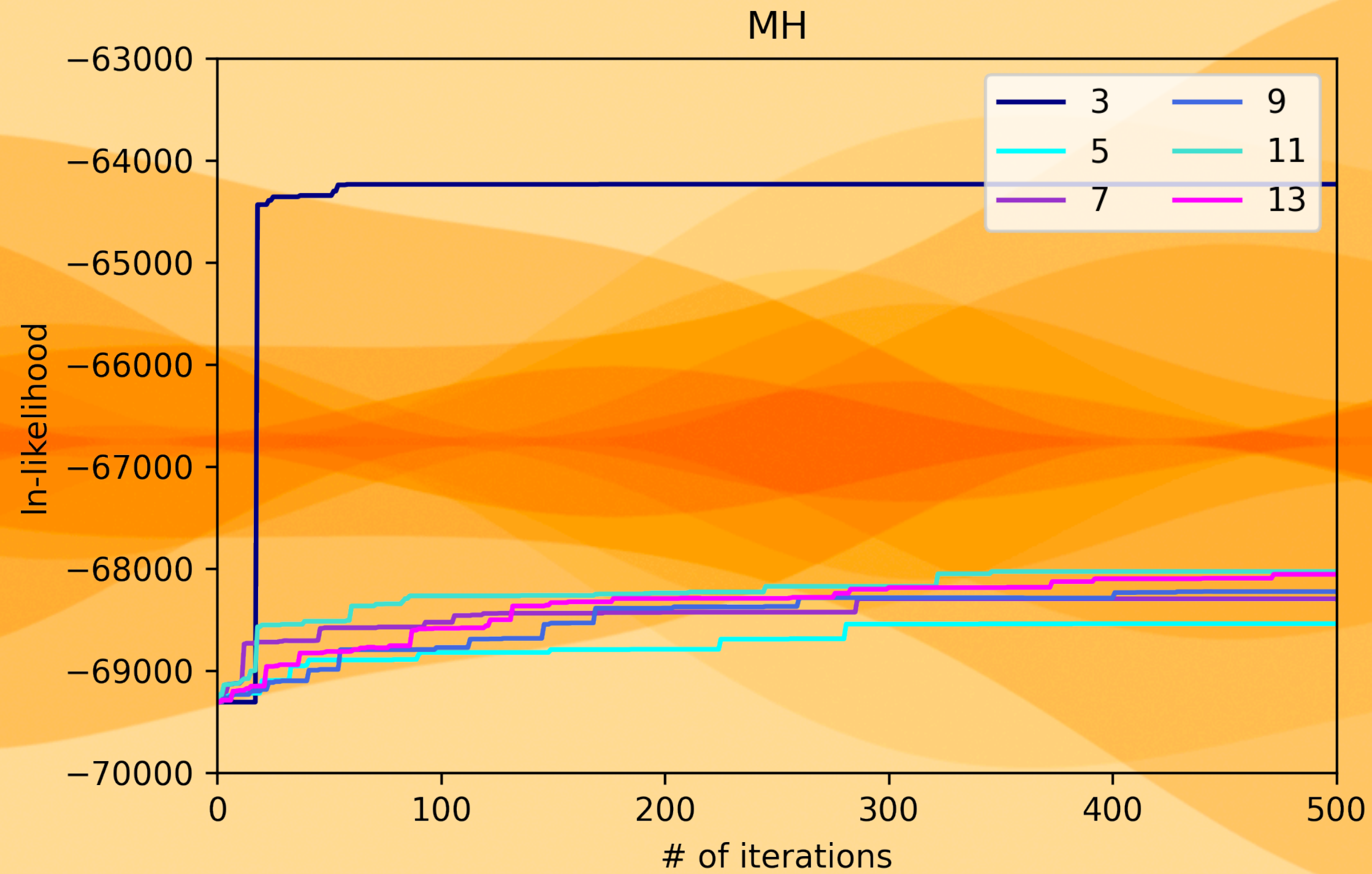
Now we set

$$a(u, v) = \min\{1, \exp(\Phi(u) - \Phi(v))\}.$$



2D fluid dynamics experiment,
therefore: $d_u \propto \Delta x^{-2}$

(pCN in Parameter Space)



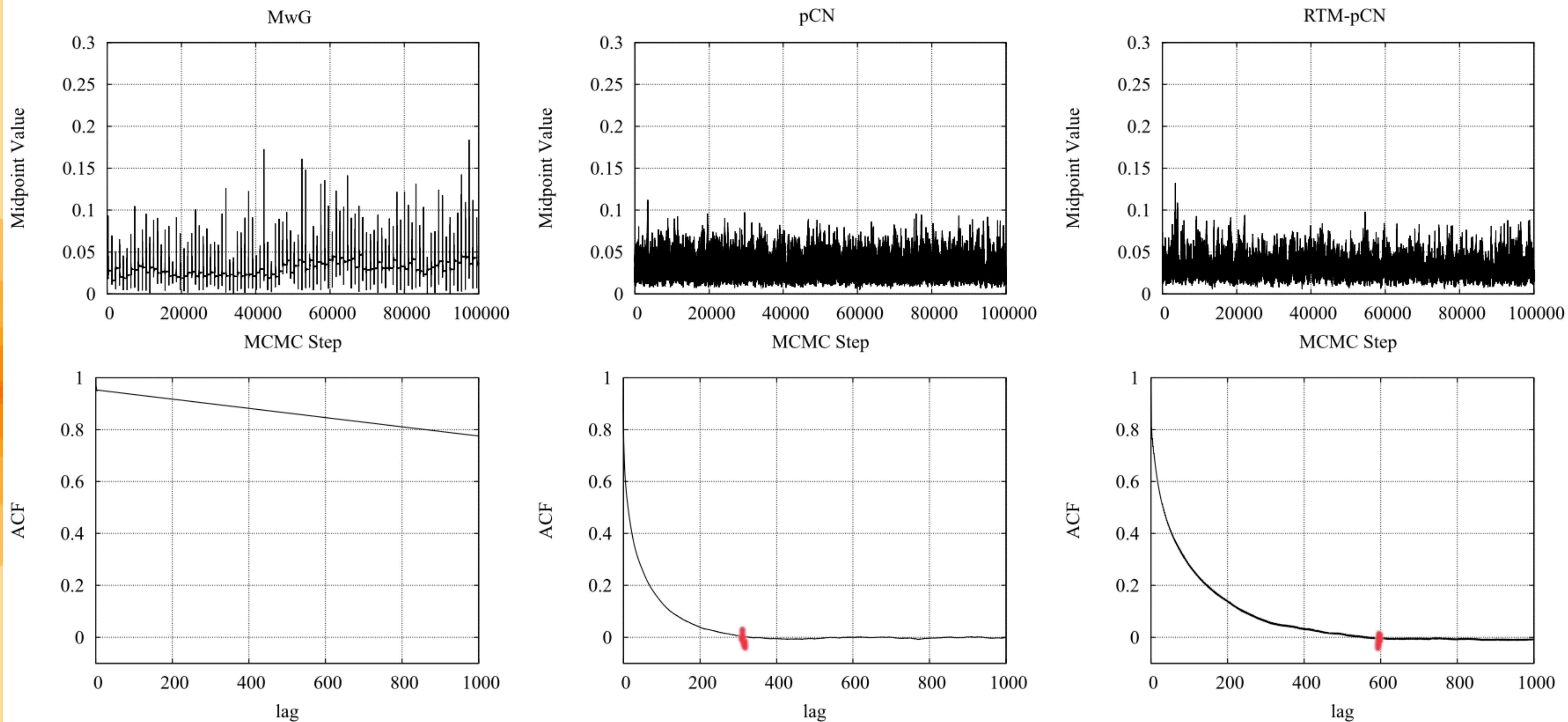


FIG. 2. Trace and autocorrelation plots for sampling posterior measure with true density ρ using MwG, pCN and RTM-pCN methods.

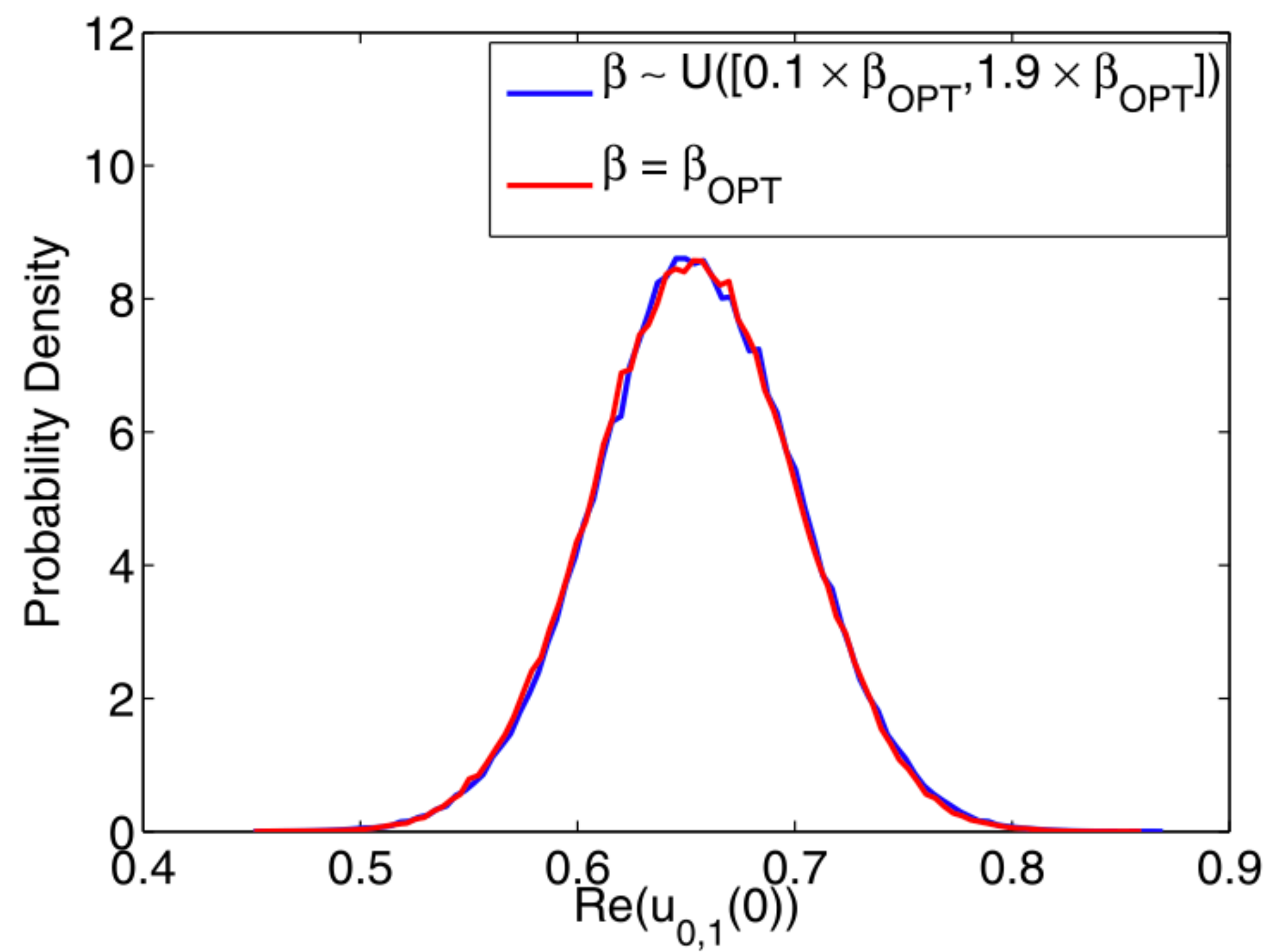
Benefits of the pCN algorithm

- Is able to maintain a high acceptance probability despite undergoing mesh refinement (increase in # of dimensions).
- Is able to produce independent samples an order of magnitude faster than non-function space samples, such as the Metropolis-within-Gibbs (MwG).
- Conclusion: very efficient probability distribution sampler.

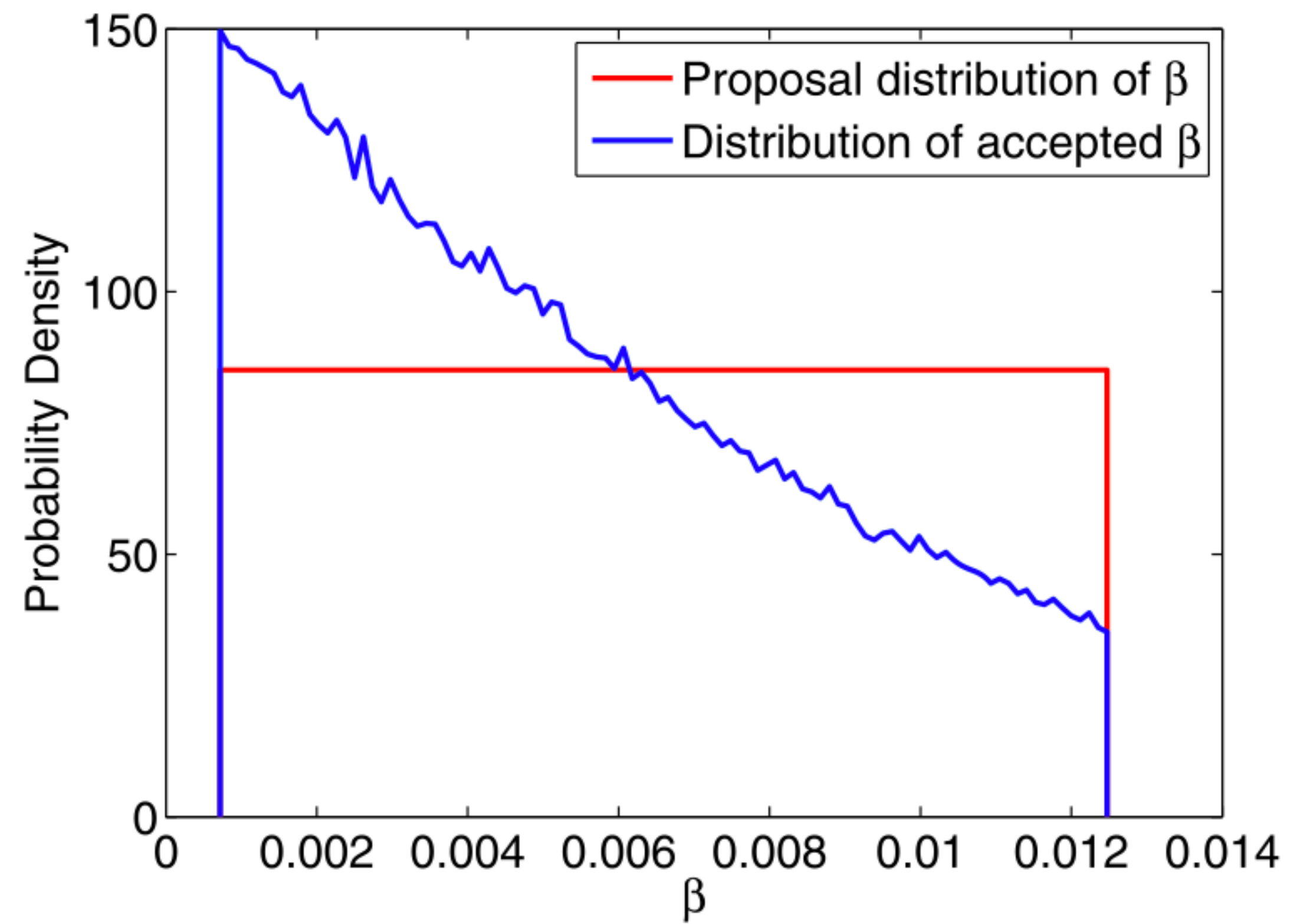
- How do we choose the step size, β ?

TABLE 2
Comparison of computational timings for target ρ

| Algorithm | Time for 10^6 steps (s) | Time to draw an indep sample (s) |
|-----------|---------------------------|----------------------------------|
| MwG | 262 | 0.234 |
| pCN | 451 | 0.0331 |
| RTM-pCN | 278 | 0.0398 |



(a)



(b)

Conclusions?

- Nonparametric MCMC is very powerful if we want to reduce the number of assumptions, however dimensionality will render traditional algorithms useless.
- The pCN algorithm maintains high acceptance probabilities through mesh refinement.
- Using a pCN algorithm over function space produces independent samples much faster than a MwG over parameter space.
- The choice of step size is somewhat ambiguous. We can safely add a random element such that the Markov chain is able to jump across areas of high PD.



Questions?