Markov-Chain Monte Carlo (MCMC)
1. Theory

Karri Muinonen
Department of Physics, University of Helsinki, Finland

Data Analysis and Inverse Methods in Astronomy, Advanced Course,
Department of Physics, University of Helsinki, February 3, 2012

Tutorial

- Markov-Chain Monte-Carlo methods (MCMC) allow sampling of multivariate probability density functions (p.d.f.) that are difficult to sample directly
- Metropolis-Hastings algorithm is a practical rejection sampling algorithm for MCMC (Metropolis et al. 1953, Hastings 1970)
  - Random-Walk Metropolis-Hastings
  - Independence-Chain Metropolis-Hastings
- Single requirement: a value proportional to p.d.f. can be computed for a given set of parameters (avoiding p.d.f. normalization)

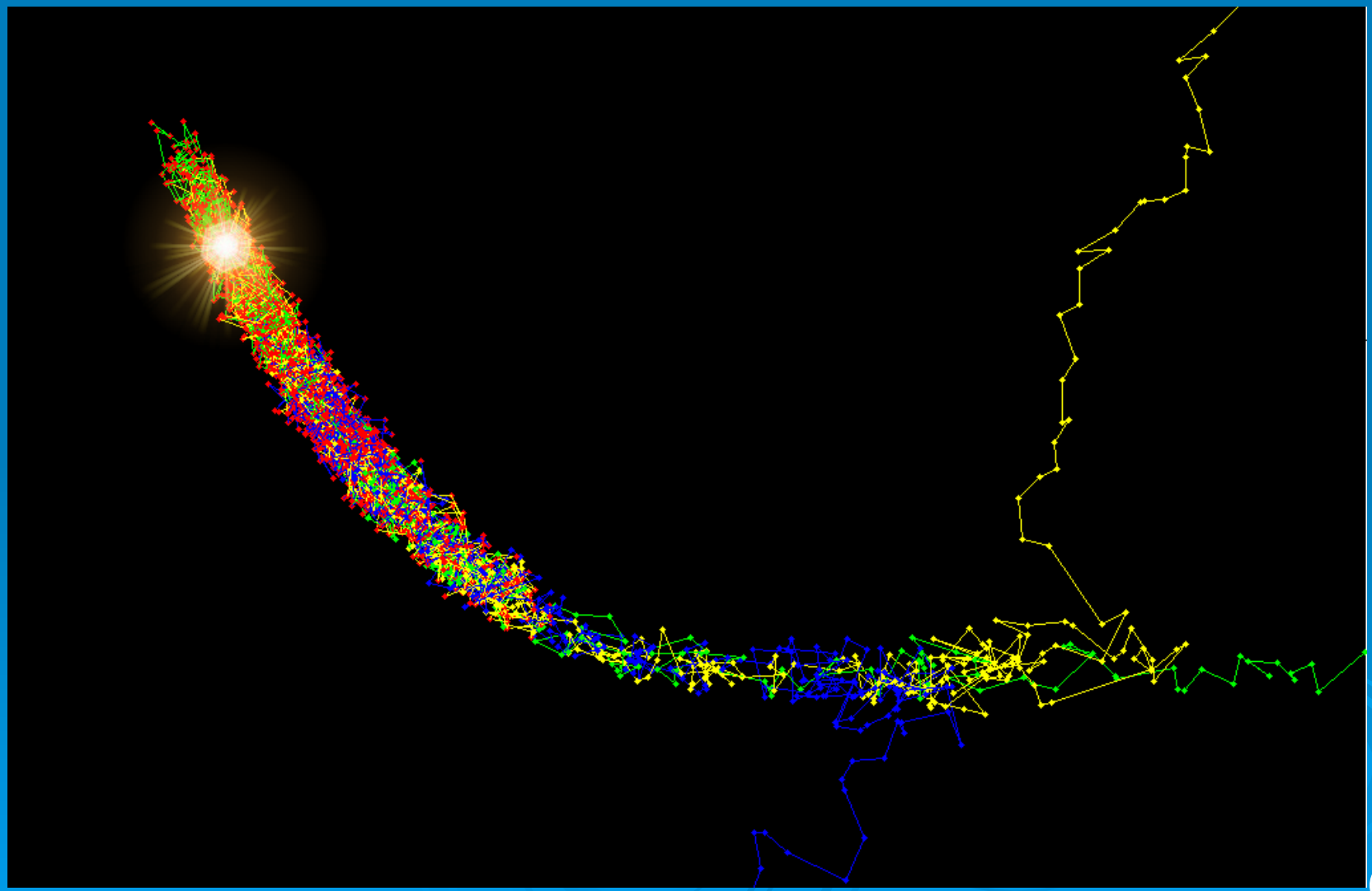
Tutorial

- For a proposal p.d.f. $q(x,x_j)$ centered at x_j with the proposed new value x, a new set of parameters is accepted as $x_{j+1}=x$ if for a random deviate y within [0,1] $y < [p(x)q(x,x_j)] / [p(x_j)q(x_j,x)]$, where p(x) is the p.d.f. to be sampled; if the proposed set is rejected, $x_{j+1}=x_j$
- If q is symmetric, the acceptance criterion reduces to $y < p(x)/p(x_j)$ for a uniform random deviate y within [0,1]
- Typical proposal p.d.f.: Gaussian p.d.f. with standard deviations to be tuned

Tutorial

- Example: 3D Rosenbrock Function
- Non-convex function widely used in testing optimization algorithms (with extensions to higher dimensions)
  - $f(x,y) = (1-x^2)+100(y-x^2)^2$
  - global minimum at x=y=1, in a long flat parabolic valley
  - converted to a p.d.f.
- Source for this example: Wikipedia

Tutorial



Tutorial

- Optimum proposal p.d.f. coincides with the one to be sampled (but typically not available)
- Challenges in tuning proposal p.d.f. parameters: how to guarantee global coverage in a finite amount of computing time?
- “Optimum acceptance rate” around 10-40%

Introduction

- For arbitrary multivariate probability distributions, efficient direct sampling methods do not exist
- MCMC provides the facility to draw dependent samples from a posteriori distributions
- MCMC constitutes the major reason for the increasing application of Bayesian methods

Introduction

- MCMC methods simulate a discrete-time homogeneous Markov chain
- Markov chain has only a one-step memory
- Asymptotically, under certain requirements, the Markov chain represents a sample from the a posteriori distribution
- MCMC constitutes a (pseudo-)random-number generator for an arbitrary probability density

Introduction

- How can we ensure that the invariant distribution of a Markov chain is the distribution that we wish to sample from?
- How can we ensure that the Markov chain satisfies the conditions for the distribution of parameters to converge to the invariant distribution?

Introduction

- How long do we need to run the chain before we can suppose that the distribution of parameters is (sufficiently close to) the invariant distribution?
- How long do we need to run the chain in order to get a sufficiently large sample to compute the required summaries of the distribution?

Introduction

- Markov chain theory and methods for constructing Markov chains
- Metropolis-Hastings method and Gibbs sampler
- Optimization of MCMC methods
- Convergence diagnostics
- Presence of model uncertainty

Markov Chains

- A discrete-time homogeneous Markov chain is defined by a transition kernel that specifies the conditional distribution for x_{i+1} given x_i and is independent of i
- Target distribution = a posteriori distribution
- Detailed balance equations ensure that the Markov chain has the a posteriori distribution as its invariant (or stationary) distribution
- A Markov chain that satisfies the detailed balance equations is said to be reversible

Markov Chains

- If P is the transition kernel of a Markov chain which has invariant distribution f then provided that it is f-irreducible, aperiodic, and Harris recurrent it is said to be ergodic.
- An ergodic Markov chain tends to produce a sample of the a posteriori distribution as the number of transitions of the chain tends to infinity.
- The time it takes for the chain to approximately achieve its equilibrium distribution is often called the burn-in.

Markov Chains

- We require the central limit theorem for Markov chains, that is, that the difference between the ergodic sample average of a function of the parameters and the ensemble average tends to normality as the number of samples tends to infinity
- This is true when the chain is reversible and a certain variance (sum of covariances along the chain) remains finite
- The variance can, in principle, be estimated from the output of the MCMC sampler

Markov Chains

- "Irreducible" means that, for any starting value of the chain, the chain can eventually reach every region of the parameter space with positive probability
- The chain is periodic if it cycles between disjoint subsets of the parameter space. Any chain that is not periodic is aperiodic.
- The chain is Harris recurrent if any subset of the parameter space with positive probability is visited infinitely often by the chain.

Metropolis-Hastings Algorithms

- A proposal parameter value is generated from a proposal density and accepted/ rejected on the basis of a straightforward criterion (see earlier).
- Unnormalized probability densities can be readily utilized.
- Extremely general method for sampling from a general posterior distribution

Random Walk Algoritms

- In a random walk algorithm, a zero-mean random increment with, e.g., multivariate normal probability density function is added to current value of the chain independently of the value
- If the distribution is symmetric, the proposal density drops off completely
- In a random-walk algorithm, the chain will spend more time in regions of high probability

Random Walk Algorithms

- If the proposal variance is set to a small value, proposals have a high probability of being accepted, but successive realizations of the chain will be very close to one another, leading to a slow transition between distant areas of the parameter space.
- If the variance is large, larger transitions are proposed. If the variance is too large, proposed values may be in areas of the parameter space with low posterior probability and hence likely to be rejected.

Random Walk Algorithms

- Correlation structure and/or the relative variances in the proposal covariance can be chosen based on an approximation to the posterior distribution.
- The overall scale is nevertheless likely to be crucial in determining the convergence of the chain.
- Acceptance rates between 0.1 and 0.4 ought to perform close to optimal.
- Advisable to spend some time tuning the scale of the proposal distribution.

The Independence Sampler

- In the independence sampler, the proposal distribution is fixed and independent of the value of the chain.
- Dependence in the transition distribution arises through the dependence of the acceptance probability on the current value of the chain.
- Independence sampler can fail to converge. It should be designed to have strong tails.
- The independence sampler should mimic the target posterior density as closely as possible.

Combining MCMC Samplers

- The Markov chain transition kernel can be a mixture kernel that is a weighted sum of a number of transition kernels (with individual probabilities).
- A mixture proposal will exhibit generally superior convergence properties to the mixture kernel.
- A mixture proposal is, however, less efficient to generate.

Single Variable and Block Updates

- In the variable-at-a-time Metropolis-Hastings algorithm, single parameter is proposed transitions for.
- By updating blocks or individual components, it is often possible to construct mobile samplers
- It remains as a challenge to obtain an ergodic sampler.

Gibbs Sampling

- For the Gibbs sampler, the proposal probability density is the conditional a posteriori density of the block parameters given the current values of the other parameters
- Proposal is always accepted
- Gibbs sampler among the most popular MCMC methods
- Paper on the Gibbs sampler by Gelfand and Smith (1990) marks the real beginning of MCMC in Bayesian computation

Sampling from the Conditional Distributions

- The form of the conditional distributions is immediately available by examining the form of the unnormalized joint density

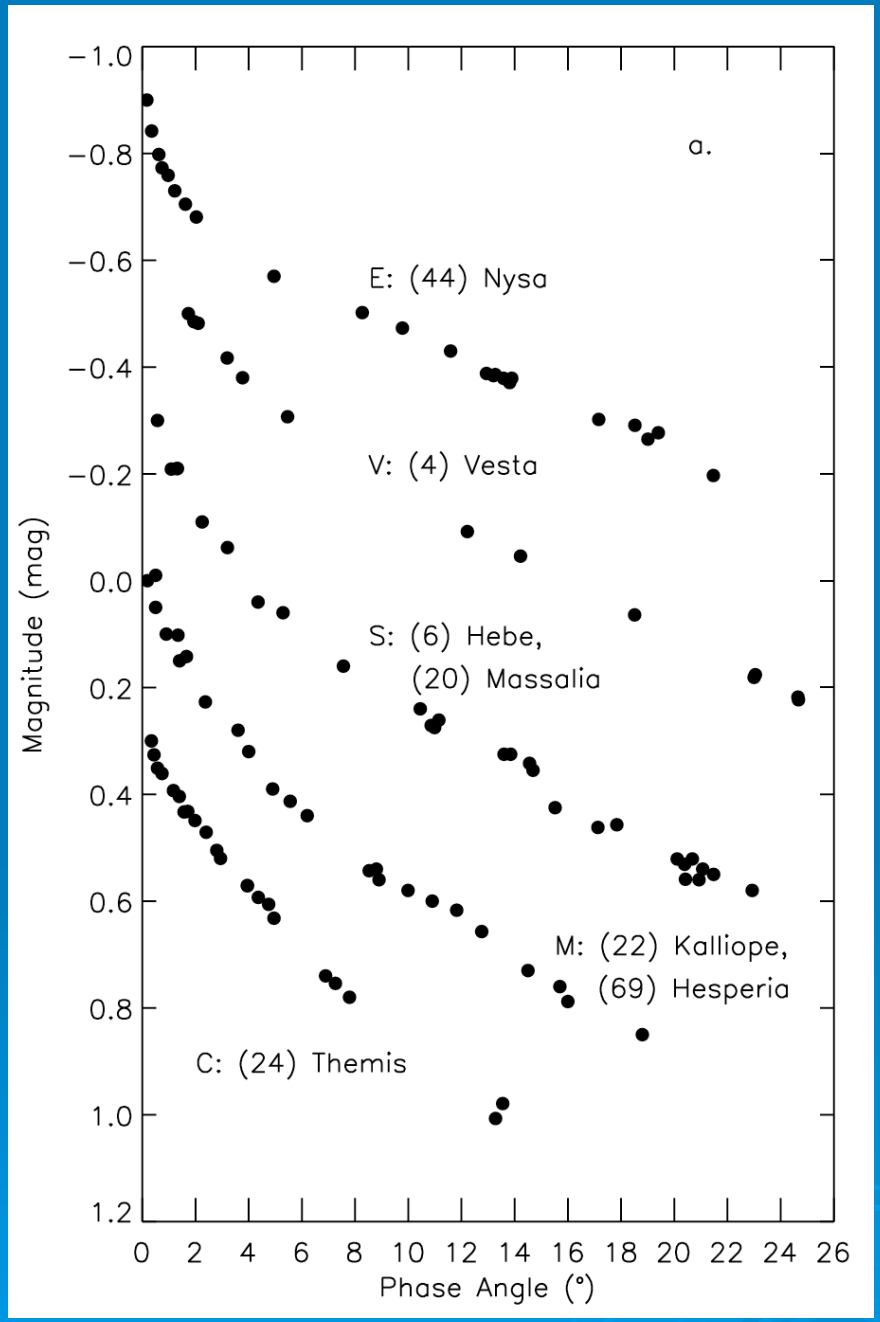
Rejection Methods

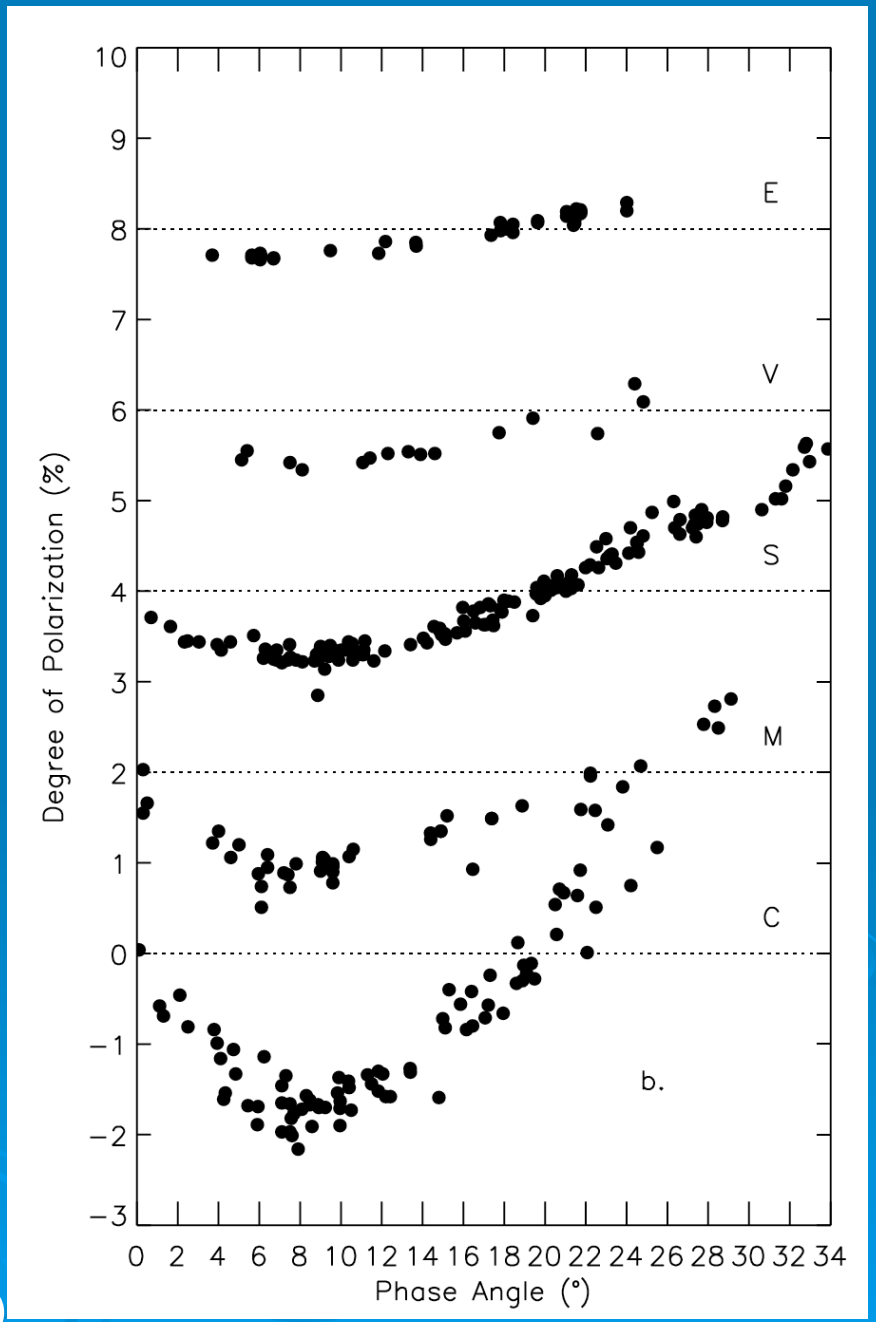
- Let s(x) be a density from which we can conveniently sample x and let the ratio g(x)/s(x) be bounded, $A \geq g(x)/s(x)$, for all x.
- Sampling from g can proceed in the following way: take two independent random draws, x from s(x) and y from the uniform distribution on [0,1]. If $Ay \leq g(x)/s(x)$ retain the draw x, otherwise reject and draw fresh x and y, and continue doing so until succeeding.
- Note that rejection sampling could be applied to the a posteriori distributions directly. But, as stated earlier, the implementation is hard without obtaining enormous numbers of rejections.

MCMC Optimization

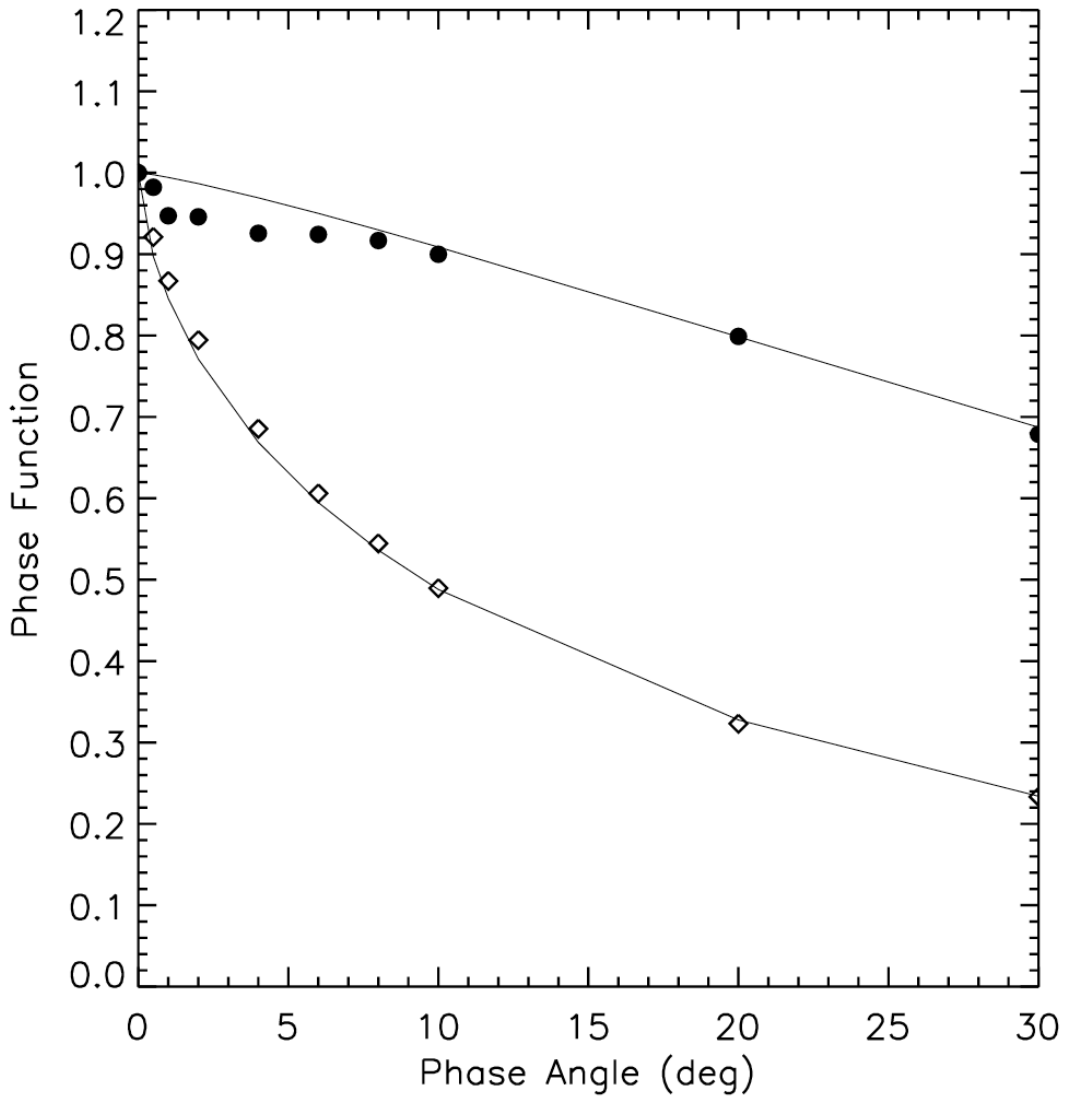
- Empirical systems for asteroid brightness and polarization phase curves
  - Revision of the H, G magnitude system of the International Astronomical Union
  - Revision of the trigonometric polarization system
- Instead of pre-determined basis functions, utilize MCMC in a search of optimum functions via large numbers of parameters
- MCMC can yield an arbitrary number of systems to choose from

Asteroids' III, p.123, obs. ref. therein





Magnitude system

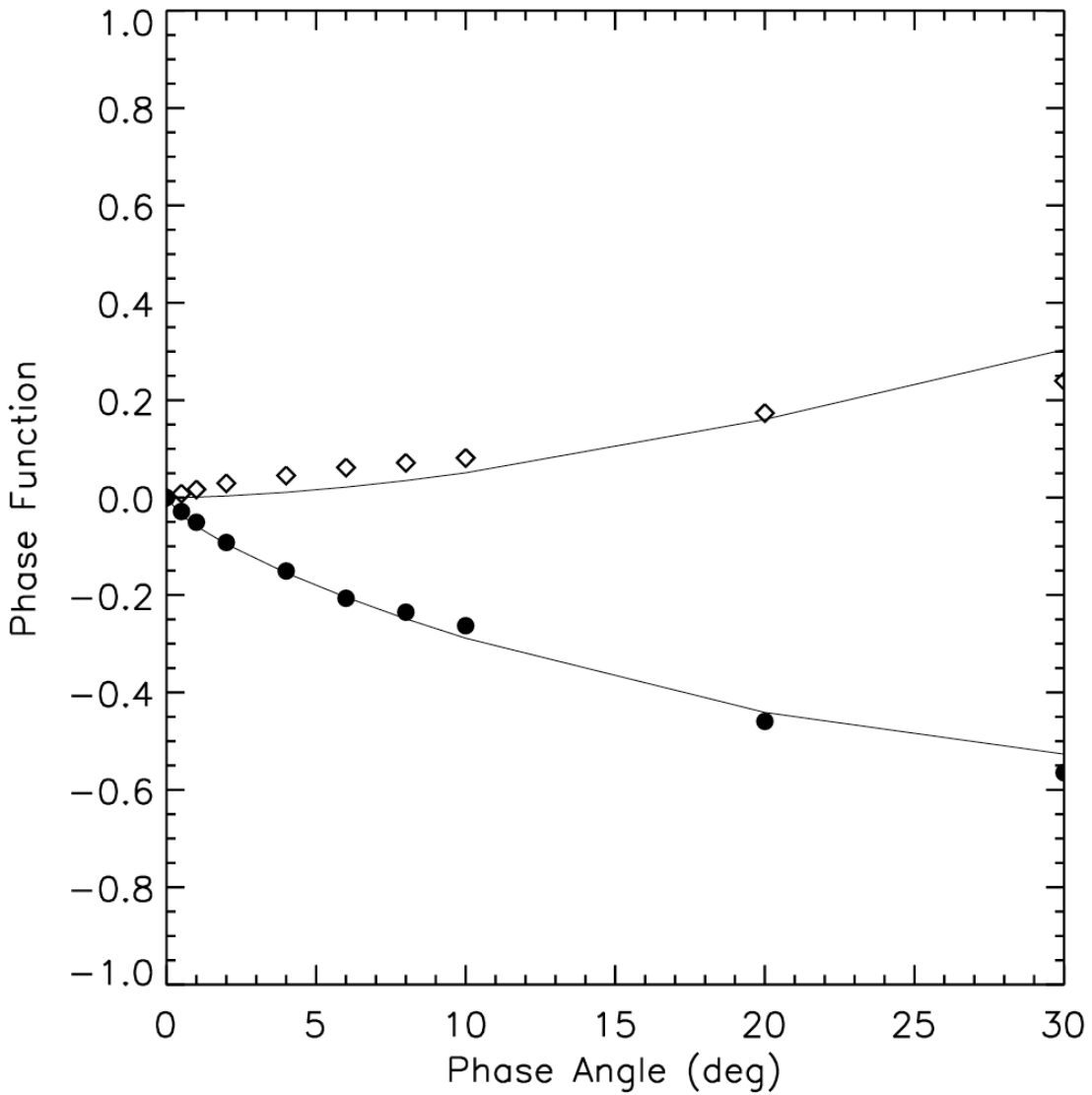


Solid lines: H, G system

Symbols: revised system

Number of parameters: 18

Polarization system



Solid lines: trigonometric system

Symbols: revised system

Number of parameters: 18

MCMC Case Study 1

- Determine the parameters of a linear-exponential model for the disk-integrated brightness of atmosphereless solar-system objects near opposition.