Is my inference any good? What does “good” mean?

May 7, 2025

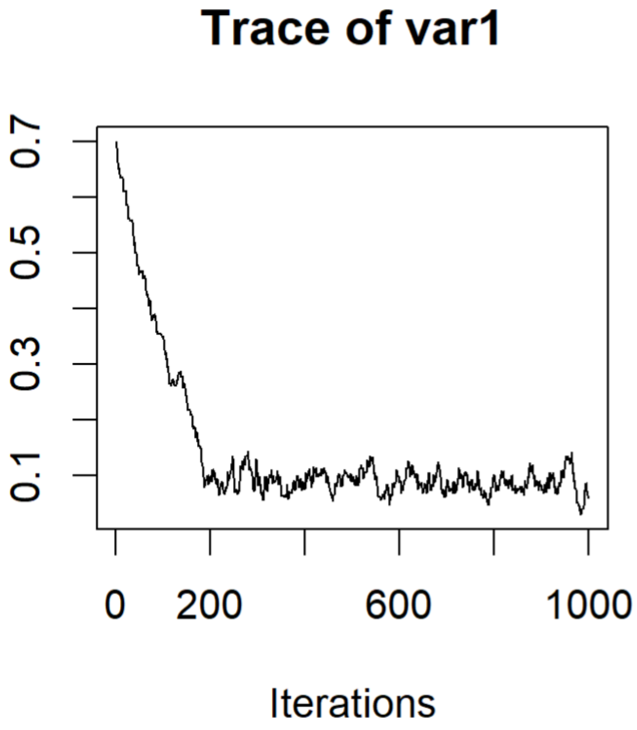
Prof. Gwendolyn Eadie

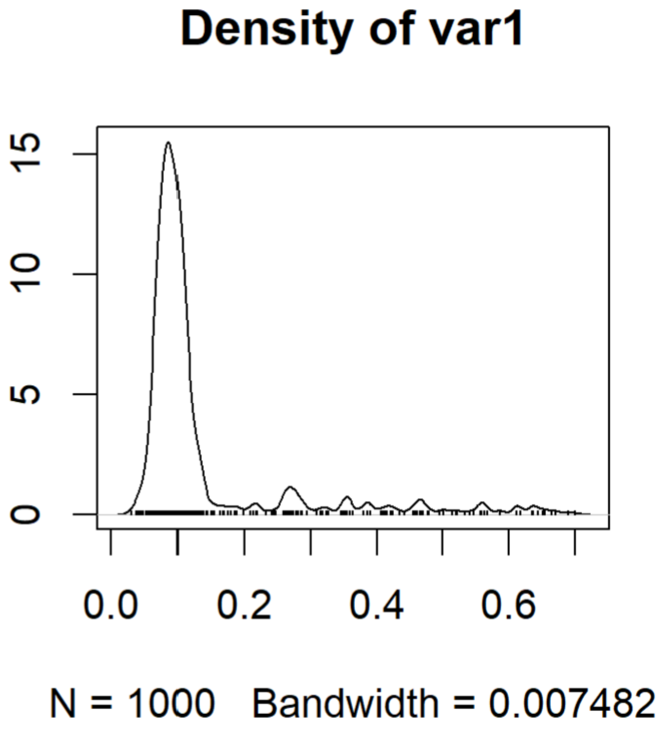
Inference and Convergence

- Inference:
  - Plotting and summarizing posterior samples
  - Computing quantiles, moments, other summary statistics, etc.
  - Posterior predictive simulations
- **But, the above depends heavily on whether your samples are a sample representative of the target distribution!**
- Convergence to the target distribution:
  - Design simulations that allow monitoring of convergence
  - Monitor the convergence with multiple diagnostics
    - many tests, diagnostics, etc have and continue to be developed to assess and monitor convergence
  - Techniques that avoid getting into bad places in parameter space to begin with

Assessing Convergence: discard the burn-in

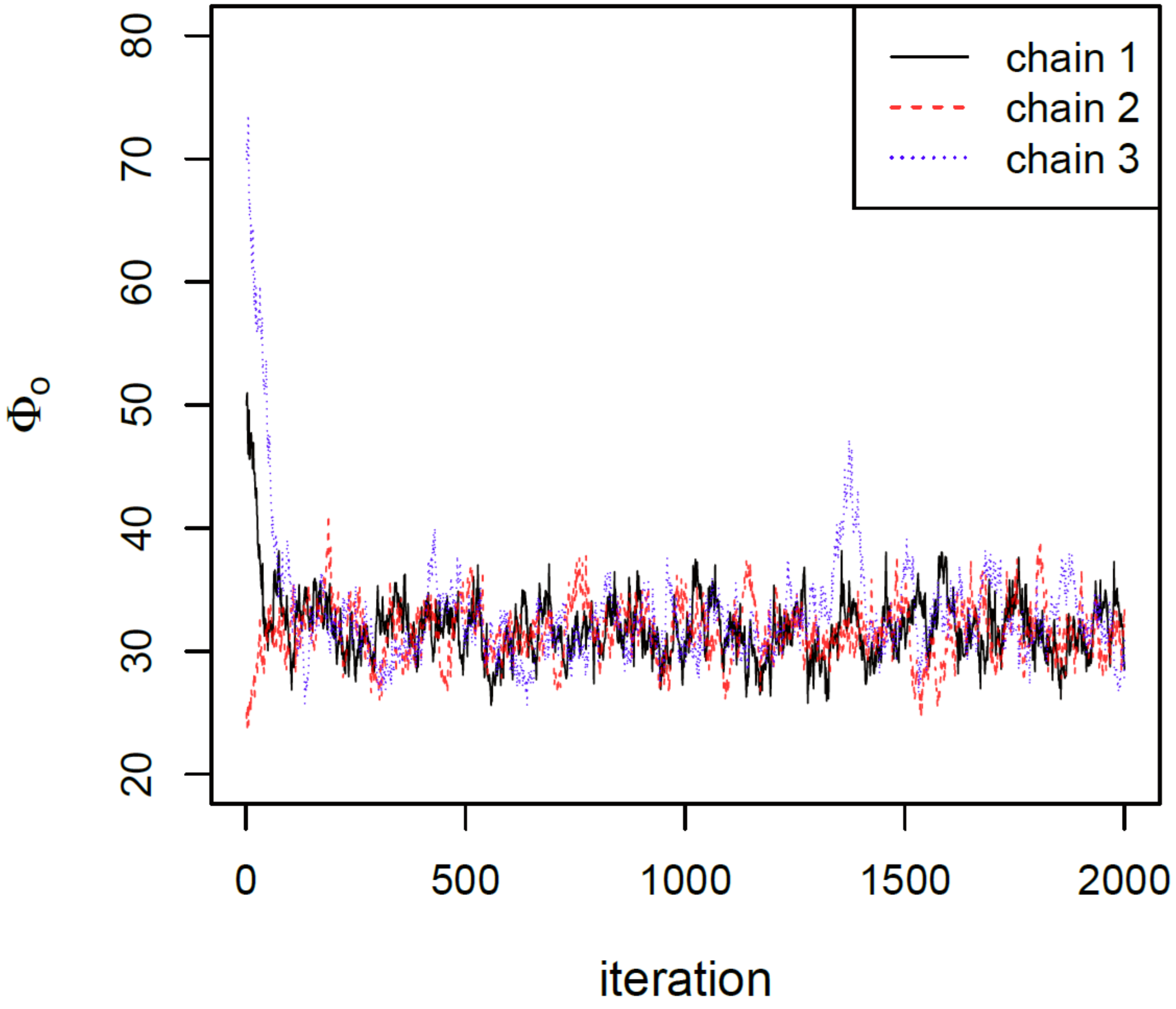
- Discard the “burn-in” from a Markov chain
  - Look at the traceplot of the samples, get rid of the first few hundred (or more!) samples





Assessing Convergence: run multiple chains

- Run independent, multiple chains
- Start them in different parts of parameter space to make sure then converge to the same place
- Evaluate convergence diagnostics such as $\hat{R}$ that assess between-chain and within-chain information



Assessing Convergence: the $\hat{R}$

In *Bayesian Data Analysis*, Gelman et al (2014) suggest a convergence diagnostic

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}(\psi|y)}{W}}$$

Which can be computed when multiple, independent chains are run. In the equation above,

$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

where W and B are the estimates of the within and between chain variances respectively.

As $n \to \infty$, $\hat{R}$ declines to 1.

A general rule is that if $\hat{R}$ is less than 1.1 (or closer to 1 if you want to be more conservative), then you can probably safely assume that you don't need to run the chain longer.

Inference and Assessing Convergence: effective sample size

- Ideally, every sample should only depend on the sample before it, but in practice there may be more autocorrelation

- If autocorrelation is high but unavoidable, then you need to *thin* the chains
  - Take every k^{th} sample

- Calculate the *effective sample size* across m chains of length n (see BDA by Gelman et al for more details):

$$\hat{n}_{eff} = \frac{mn}{1 + 2\Sigma_{t=1}^{T}\, \widehat{\rho_t}}$$

where $\widehat{\rho_t}$ are the estimated autocorrelations and T is the first odd positive integer for which $\widehat{\rho_{T+1}} + \widehat{\rho_{T+2}}$ is negative (Gelman et al 2014, BDA)

Inference and Assessing Convergence

- The $\hat{R}$ and $\hat{n}_{eff}$ do not work great for highly non-Gaussian distributions
- Sometimes sampling a transformation of the parameter is better, e.g.,
  - log-transformations of parameters
  - Logit transformation of quantities that fall in (0,1)
  - Rank transform for long-tailed distributions