Convergence Diagnostics using CODA for Optimizing the Antithetic Gibbs Sampler

Johannes Dreesman

Niedersächsisches Landesgesundheitsamt
Roesebeckstr. 4-6, 30449 Hannover

Johannes.Dreesman@nlga.niedersachsen.de

Leipzig, Dezember 2002

Gaussian Markov random fields

- Consider the regular lattice
  $L := \{(i,j) : i = 1, \ldots, I, j = 1, \ldots, J\}$.

- The set of local conditional distributions

$$X_{i,j} | \{x_{r,s} : (r,s) \neq (i,j)\} \sim N\left(\eta_{i,j}, \tau^2\right)$$
$$\text{for all } (i,j) \in L,$$
$$\text{where } \eta_{i,j} := \mu + (x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1} - 4\mu) \cdot \beta.$$

  specifies a Gaussian Markov random field (GMRF).

- Restriction: $|\beta| < 0.25$.

- $\mathbf{x} := (X_{1,1}, \ldots, X_{I,J}) \sim N(\mu, \Sigma)$,
  Σ: nondiagonal covariance matrix.

- Direct simulation by means of Cholesky factorization of the covariance matrix requires about $(I \cdot J)^3/6$ floating point operations.

- Alternative approach: Markov chain Monte Carlo (MCMC).

MCMC for GMRFs

- Natural algorithm: Gibbs sampler (GS).
- GS resamples $\mathbf{x}$ componentwise from the local conditional distributions.
- MCMC enables estimation of $E(f(\mathbf{x}))$ by $\bar{f}_M(\mathbf{x}) := \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{x}^{(m)})$.
- Current realization is starting point for generating a new one.
- In general the successive values of $f(\mathbf{x})$ are not independent.
- Usually they show positive autocorrelation.
- Barone & Frigessi (1990) suggested the introduction of an antithetic parameter which counteracts the autocorrelation.

The antithetic GS for GMRFs

- $X_{i,j}^{(m)}$ is resampled from

  $$N\left((1+\psi)\cdot\eta_{i,j}^{(m)} - \psi\cdot x_{i,j}^{(m-1)}, (1-\psi^2)\cdot\tau^2\right)$$

  instead of $N\left(\eta_{i,j}^{(m)}, \tau^2\right)$.

- Antithetic parameter $\psi \in (-1, 1)$.

- $\eta_{i,j}^{(m)}$ from the current realization of $\mathbf{x}$.

- Subtraction of $\psi\cdot x_{i,j}^{(m-1)}$ induces negative autocorrelation.

- Antithetic GS includes ordinary GS as special case for $\psi = 0$, but is no GS for $\psi \neq 0$.

Analysis of equilibrium properties by Metropolis Hastings (MH) method

- Let
  - $\pi(\mathbf{x})$ be the distribution of interest and
  - $q(\mathbf{x}'|\mathbf{x})$ the proposal density from which the MH step generates $\mathbf{x}'$, given $\mathbf{x}$. Then

$$a(\mathbf{x}'|\mathbf{x}) = \min\left(\frac{\pi(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{\pi(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}, 1\right)$$

  is the acceptance probability of $\mathbf{x}'$ and

$$1 - a(\mathbf{x}'|\mathbf{x})$$

  is the probability of taking $\mathbf{x}$ again.

- Gibbs update is single component MH update with acceptance probability 1 since

$$\frac{q(\mathbf{x}'|\mathbf{x})}{q(\mathbf{x}|\mathbf{x}')} = \frac{\pi(\mathbf{x}')}{\pi(\mathbf{x})}.$$

- For the antithetic GS

$$\frac{q(\mathbf{x}'|\mathbf{x})}{q(\mathbf{x}|\mathbf{x}')} \quad \text{is independent from} \quad \psi$$

  $\Rightarrow$ Antithetic GS is a MH sampler with stationary distribution $\pi(\mathbf{x})$ if proposals are accepted with probability 1 (Green & Han, 1992).

Markov chain Monte Carlo maximum likelihood (MCMCML)

- Goal: Estimation of spatial dependence parameter β of GMRFs.

- Problem: ML estimation usually is intractable because of normalizing term.

- Alternatives: Coding, Pseudolikelihood: simple, but less efficient (Besag, 1974, 1975, 1977).

- MCMCML: ML estimation of β even if likelihood contains unknown normalizing term depending on β (Geyer, 1991).

- Consider a family of densities $\{f_\beta\}$:

$$f_\beta(\mathbf{x}) = \frac{1}{z(\beta)} h_\beta(\mathbf{x}),$$

where h_β is known and
$z(\beta) = \int h_\beta(\mathbf{x}) d\mathbf{x}$ may be intractable.

- Maximize log-likelihood-ratio $lr(\beta, \beta^* | \mathbf{x}) :=$

$$\log \left(\frac{L(\beta|\mathbf{x})}{L(\beta^*|\mathbf{x})} \right) = \log \left(\frac{h_\beta(\mathbf{x})}{h_{\beta^*}(\mathbf{x})} \right) - \log \left(\frac{z(\beta)}{z(\beta^*)} \right)$$

for observed $\mathbf{x}$ and arbitrary β^*, $f_{\beta^*} \in \{f_\beta\}$, instead of likelihood.

- Estimate $\frac{z(\beta)}{z(\beta^*)} = E_{\beta^*} \left(\frac{h_\beta(\mathbf{x})}{h_{\beta^*}(\mathbf{x})} \right)$ by
$\frac{1}{M} \sum_{m=1}^{M} \frac{h_\beta(\mathbf{x}^{(m)})}{h_{\beta^*}(\mathbf{x}^{(m)})}$, $\mathbf{x}^{(m)} \sim f_{\beta^*}$.

- Generate sample $\left(\mathbf{x}^{(m)}\right)_{m=1,\ldots,M}, \mathbf{x}^{(m)} \sim f_{\beta^*}$, by MCMC, which is possible even if the normalizing term of f_{β^*} is unknown.

$\Rightarrow$ MCMCML estimator is obtained by maximizing $\widehat{lr}_M(\beta, \beta^*|\mathbf{x}) :=$

$$\log\left(\frac{h_\beta(\mathbf{x})}{h_{\beta^*}(\mathbf{x})}\right) - \log\left(\frac{1}{M}\sum_{m=1}^{M}\frac{h_\beta(\mathbf{x}^{(m)})}{h_{\beta^*}(\mathbf{x}^{(m)})}\right),\ \mathbf{x}^{(m)} \sim f_{\beta^*},$$

with respect to β.

- For the GMRF $\widehat{lr}_M(\beta, \beta^*|\mathbf{x})$ reduces to

$$(\beta - \beta^*)\cdot g(\mathbf{x}) - \log\left(\frac{1}{M}\sum_{m=1}\exp\Big((\beta - \beta^*)\cdot g(\mathbf{x}^{(m)})\Big)\right),\ \mathbf{x}$$

where

$$\begin{aligned} g(\mathbf{x}) \ = \ & \sum_{(i,j)\in L}(X_{i,j} - \mu)\cdot(X_{i+1,j} - \mu) + \\ & \sum_{(i,j)\in L}(X_{i,j} - \mu)\cdot(X_{i,j+1} - \mu). \end{aligned}$$

- If antithetic GS is used for simulation, goodness properties depend on ψ, which interacts with β.

- Theoretical results of Green & Han (1992) do not apply since $g(\mathbf{x})$ is no linear function of $\mathbf{x}$.

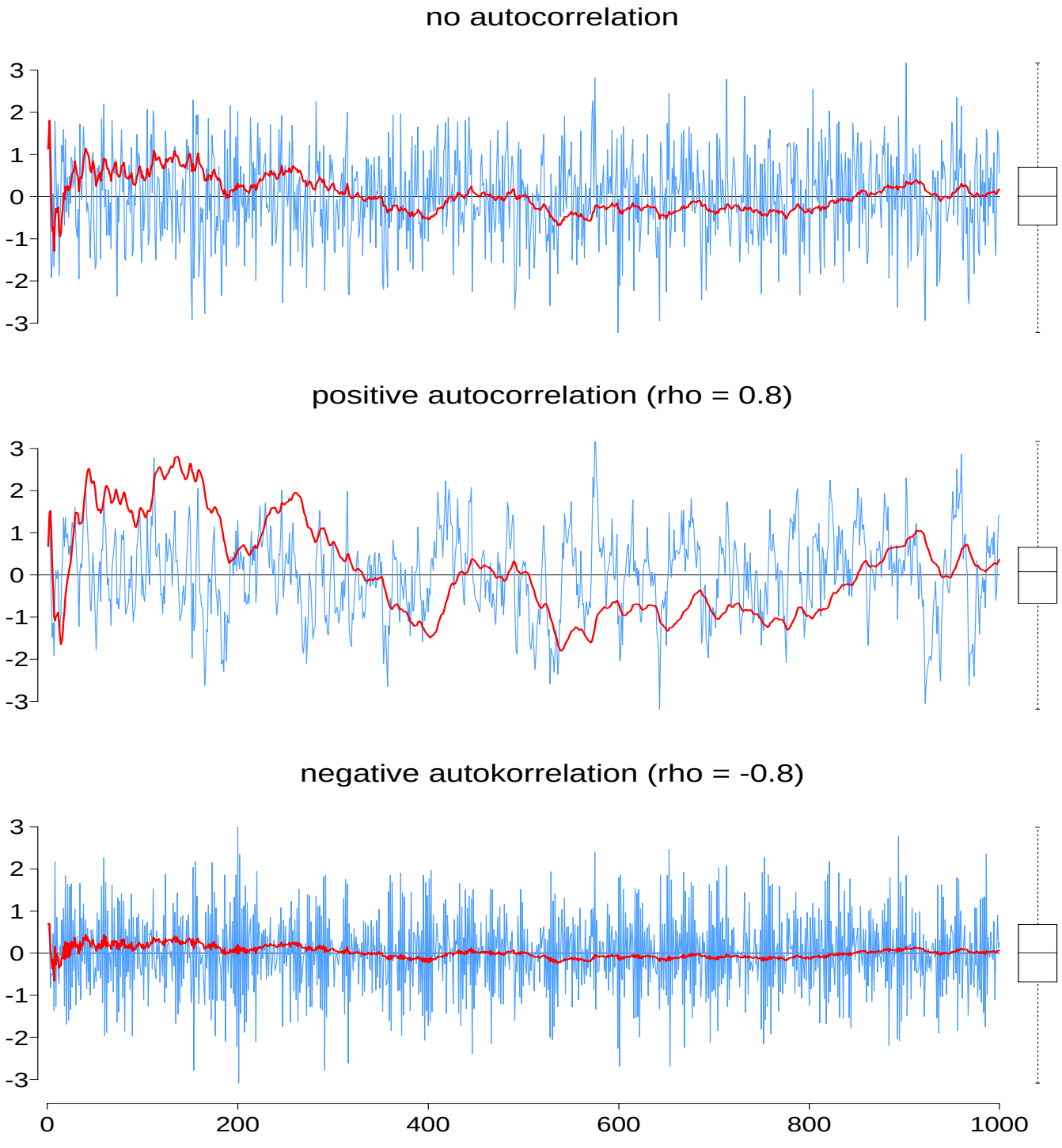
Simulation study

- For $\beta = 0.01, 0.1, 0.2, 0.24$.
- For $\psi = -0.95, -0.9, -0.8, \ldots, 0.8, 0.9, 0.95$.
- 2 simulations of GMRF $\mathbf{x}$ on 16×16 torus lattice.
- Run length: 10000,
  burn-in period length: 100.
- Visiting schedule: Coding sets.
- Analysis of output $\left(g(\mathbf{x}^{(m)})\right)_{m=1,\ldots,10000}$.
  Diagnostic statistics calculated with CODA (Best et al., 1995).

CODA-tools for analysis of efficiency

Autocorrelation

Autocorrelation affects the efficiency (mean square error) of the estimator for the ergodic mean.



Estimating the MCMC-Variance

Time Series Methods

MCMC-Variance: $\tilde{\varsigma}_g^2 := \lim_{M \to \infty} M \cdot var(\bar{g}_M)$

$(g^{(m)})_{m=1,\ldots,M} \sim iid$

$$\Rightarrow var(\bar{g}_M) = \frac{\sigma_g^2}{M} \Leftrightarrow M \cdot var(\bar{g}_M) = \sigma_g^2$$

In general

$$\tilde{\varsigma}_g^2 = \sum_{k=-\infty}^{\infty} \tilde{\sigma}_g(k) = \sigma_g^2 \cdot \left(\sum_{k=-\infty}^{\infty} \tilde{\rho}_g(k) \right),$$

where $\tilde{\sigma}_g(k)$ denotes the autocovariance and $\tilde{\rho}_g(k)$ the autocorrelation at lag k.

Spectral density is identical to fourier transformation of series of autocovariances:

$$\tilde{f}_g(\lambda) = \sum_{k=-\infty}^{\infty} \tilde{\sigma}_g(k) \cdot e^{i2\pi\lambda k}.$$

Estimating the MCMC-Variance

Time Series Methods II

At $\lambda = 0$ the sum of autocovariances is obtained:

$$\widehat{\widetilde{\varsigma}_g^2} = \widehat{\sum_{k=-\infty}^{\infty} \tilde{\sigma}_g(k)} = \widehat{\tilde{f}_g(0)}.$$

For estimation Geweke (1992) used the periodogram

$$I(\lambda) = M \cdot \left| \frac{1}{M} \sum_{m=1}^{M} (g^{(m)} - \bar{g}_M) \cdot e^{i2\pi\lambda m} \right|^2,$$

smoothed with bandwidth: $(\sqrt{M}/0.3 + 1)^{-1}$.

$\Rightarrow$ numerical standard error (NSE) (Geweke 1992).

Estimating the MCMC-Variance

Batch Means Estimator

Trajectory is divided into B batches of length M_B

Batch means

$$\frac{1}{M_B} \cdot \sum_{m=(b-1)\cdot M_B+1}^{b\cdot M_B} g^{(m)}, \quad b = 1, \ldots, B,$$

are calculated, which are asympt. (for $M \to \infty$)
- independent and
- distributed according to $N(\mu_g, \frac{\tilde{\varsigma}_g^2}{M_B})$.

$\Rightarrow \tilde{\varsigma}_g^2$ can be calculated from the distribution of the batch means.

If the number of batches is small, t-distribution should be used.

(Geyer, 1992)

Rafterty and Lewis

Theory for binary Markov chains exists.

$\Rightarrow$ A binary Markov chain ($u(g^{(m)})$) is generated from the output using the quantile v:

$$u(g^{(m)}) = \begin{cases} 1, & \text{if } g^{(m)} \leq v, \\ 0, & \text{otherwise.} \end{cases}$$

$u(g^{(m)})$ might be a higher order Markov chain.
$\Rightarrow$ $u(g^{(k)}), u(g^{(2k)}), \ldots$ is used for further investigations where k is choosen to give a 1st order MC according to BIC.

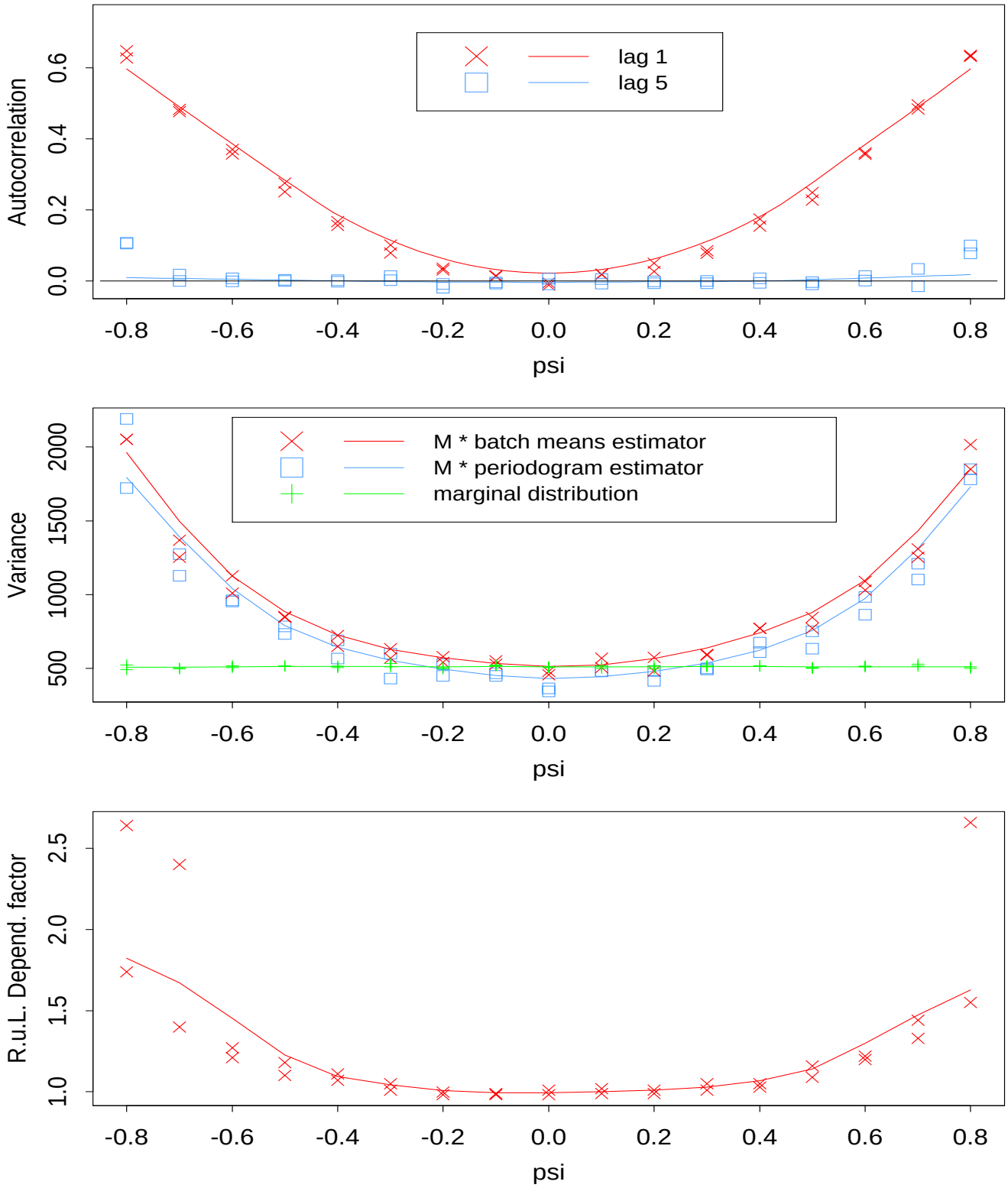
$v = 0.025$-quantile,
$95\% CI = [0.02, 0.03]$ for reestimation of 0.025
$\Rightarrow$ 3748 iterations under independence.

Dependence factor =
Necessary number of iterations / 3748.

Burn-in length: Also obtained from two state Markov chain theory.

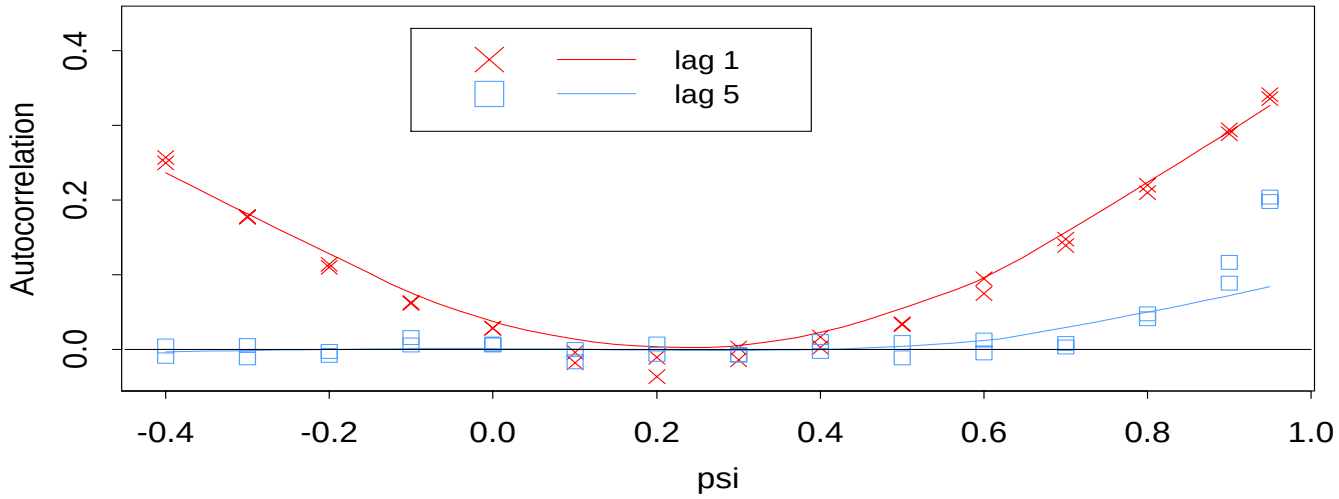
(Rafterty and Lewis, 1992, 1996)

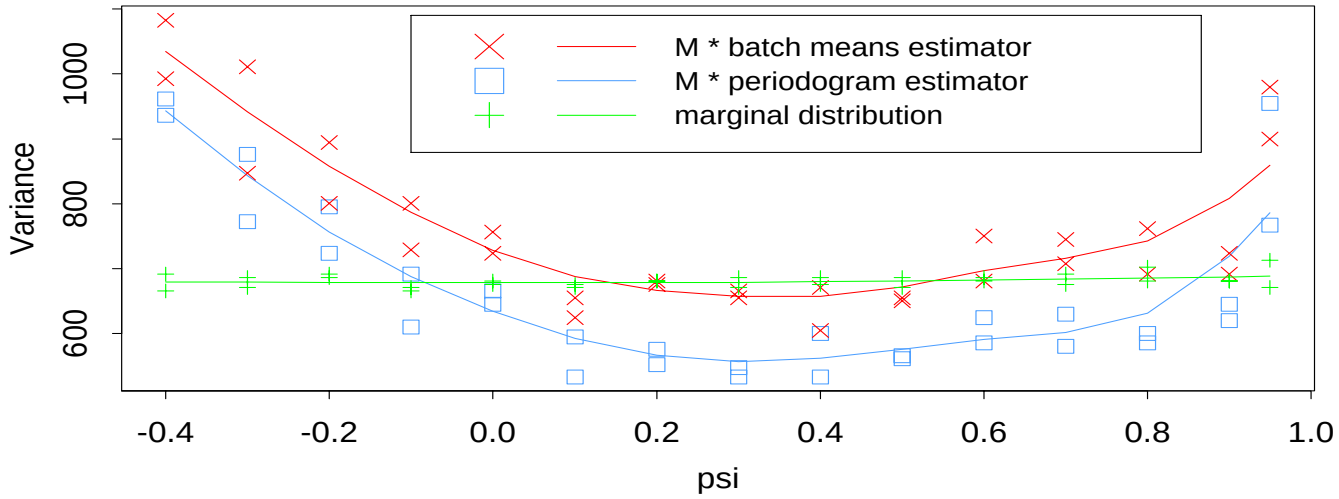
Analysis of efficiency for $\beta = 0.01$

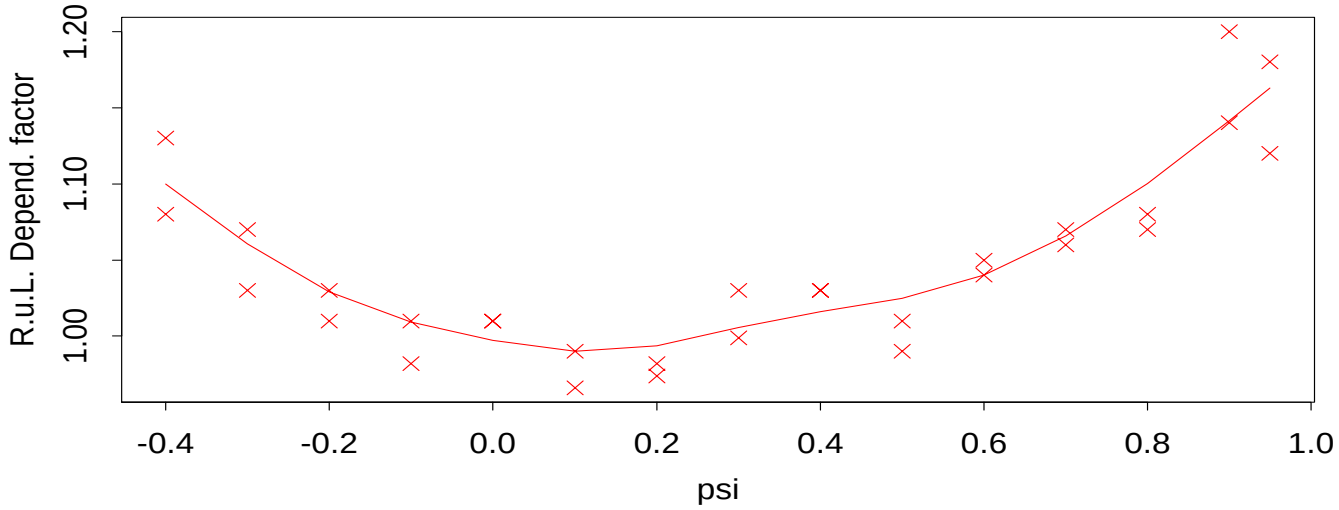


Top graph: Empirical lag-1 and lag-5 autocorr. of $\left(g(\mathbf{x}^{(m)})\right)$.
Middle graph: $M * \widehat{var}(\bar{g}_M(\mathbf{x}))$, estimated by two methods, and $\widehat{var}(g(\mathbf{x}))$ as reference.
Bottom graph: Rafterty & Lewis' Dependence factor.

Analysis of efficiency for $\beta = 0.1$

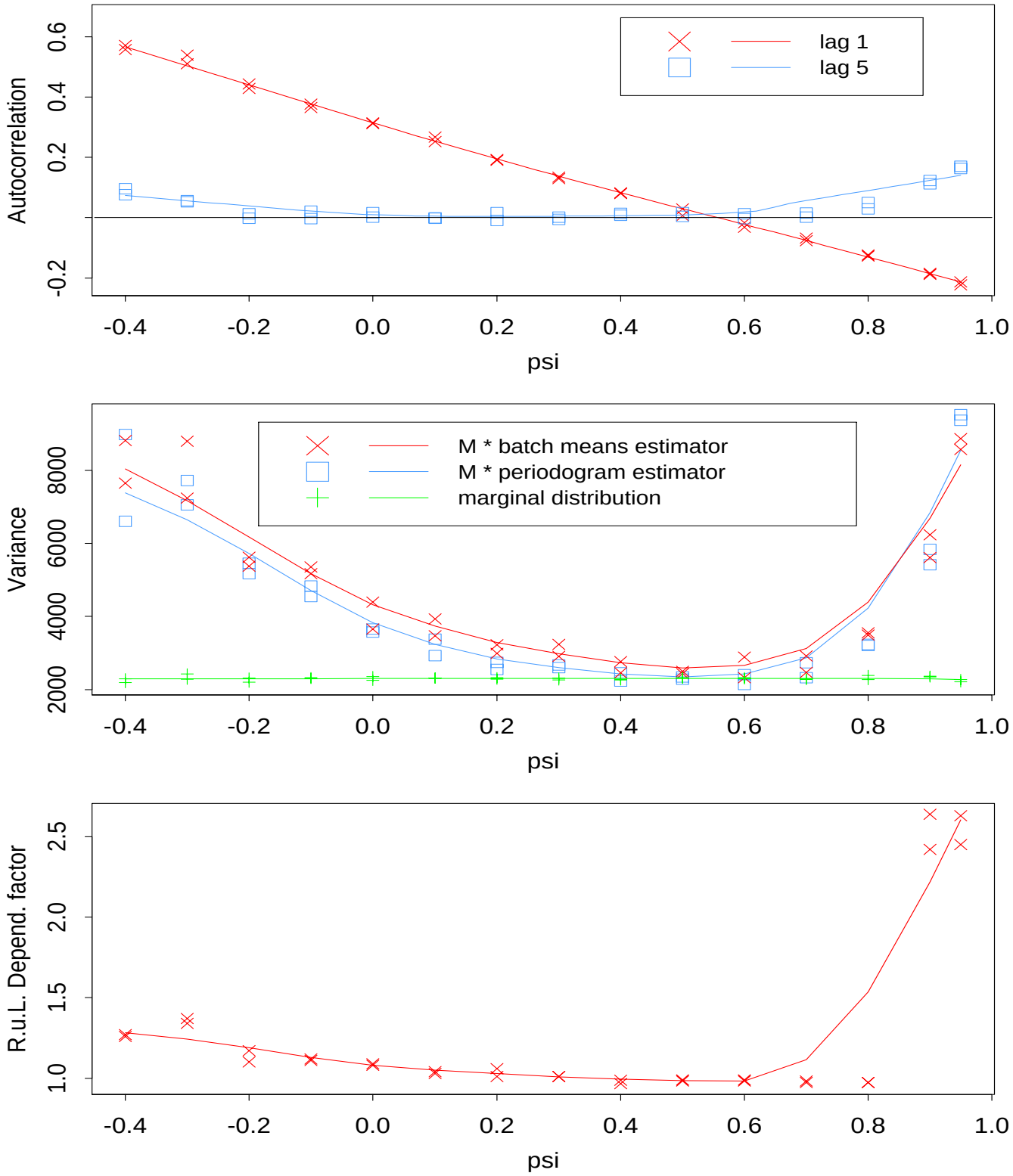






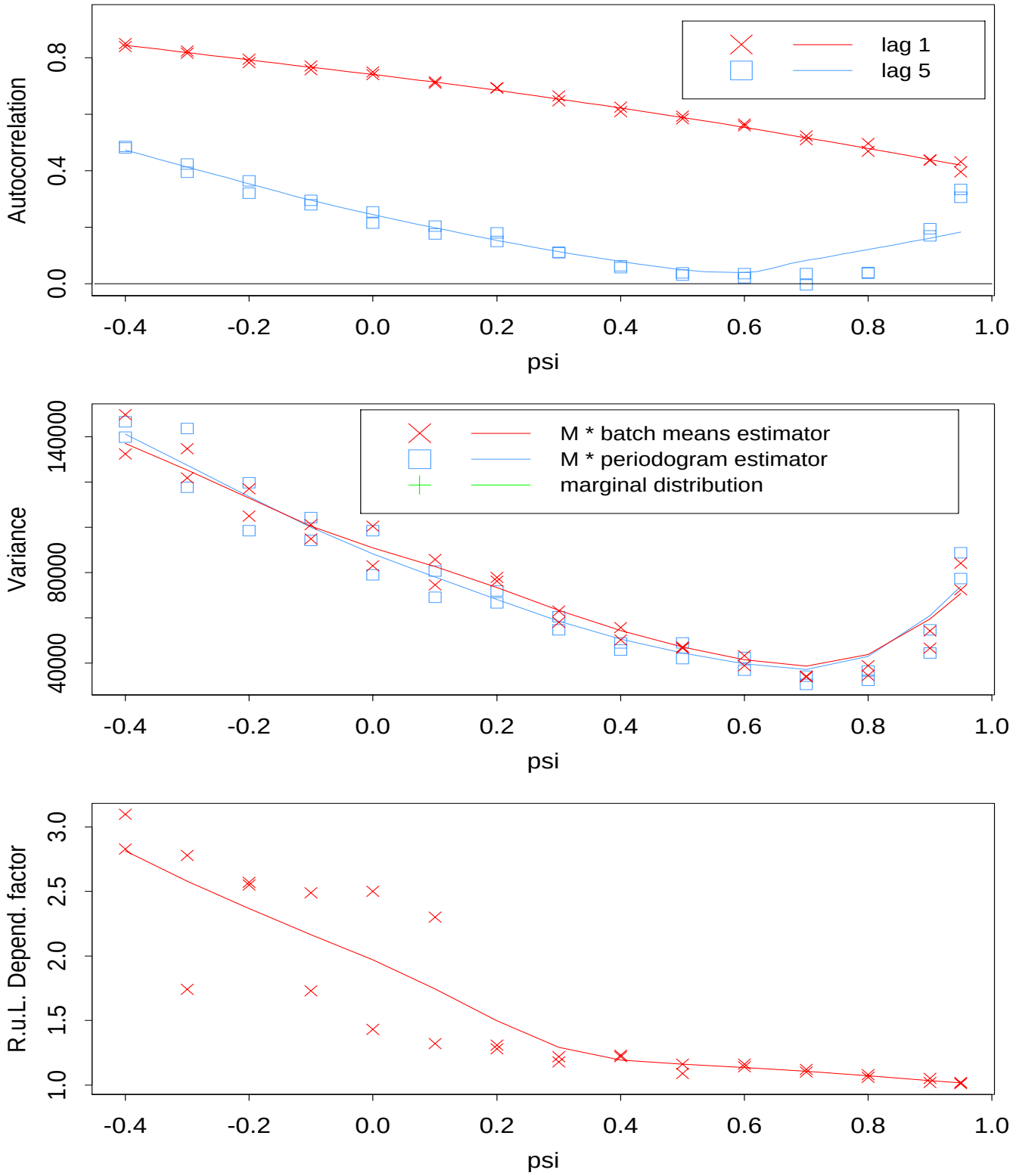
Top graph: Empirical lag-1 and lag-5 autocorr. of $\left(g(\mathbf{x}^{(m)})\right)$.
Middle graph: $M * \widehat{var}(\bar{g}_M(\mathbf{x}))$, estimated by two methods, and $\widehat{var}(g(\mathbf{x}))$ as reference.
Bottom graph: Rafferty & Lewis' Dependence factor.

Analysis of efficiency for $\beta = 0.2$



Top graph: Empirical lag-1 and lag-5 autocorr. of $\left(g(\mathbf{x}^{(m)})\right)$.
Middle graph: $M * \widehat{var}(\bar{g}_M(\mathbf{x}))$, estimated by two methods, and $\widehat{var}(g(\mathbf{x}))$ as reference.
Bottom graph: Rafterty & Lewis' Dependence factor.

Analysis of efficiency for $\beta = 0.24$



Top graph: Empirical lag-1 and lag-5 autocorr. of $\left(g(\mathbf{x}^{(m)})\right)$.
Middle graph: $M * \widehat{var}(\bar{g}_M(\mathbf{x}))$, estimated by two methods, and $\widehat{var}(g(\mathbf{x}))$ as reference.
Bottom graph: Rafferty & Lewis' Dependence factor.

Main results

- Antithetic GS more efficient than ordinary GS ($\psi = 0$) for $\beta = 0.1, 0.2, 0.24$.

- Most efficient if lag-1 autocorrelation $= 0$.

- Negative lag-1 autocorrelation doesn't give further reduction of $var(\bar{g}_M(\mathbf{x}))$.

- For $\beta = 0.01, 0.1, 0.2$:

$$\min_{\psi}(M \cdot var(\bar{g}_M(\mathbf{x}))) \approx var(g(\mathbf{x})).$$

Properties of diagn.-tools for efficiency

- MCMC Variance: Batch means tend to give systematically higher estimates than periodogram estimator.

- In most situations they correspond with dependence factor.

- In some situation strong differences.

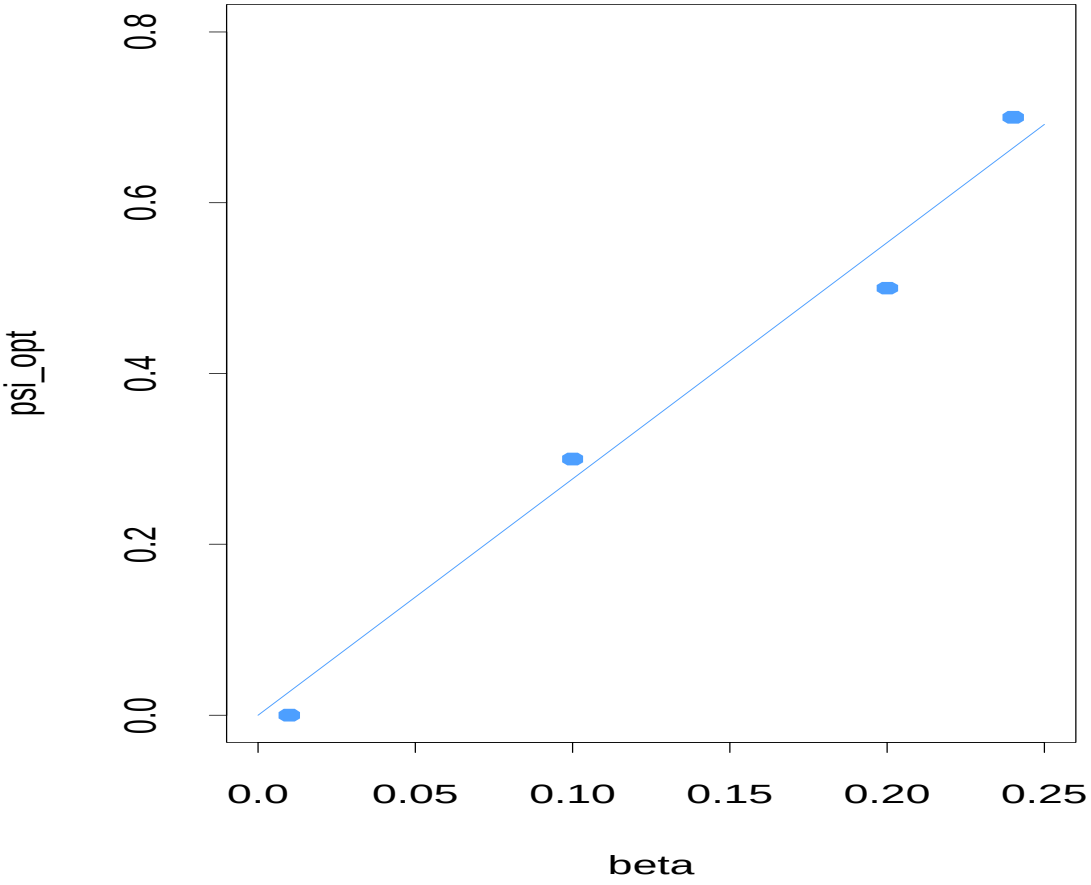
- Without higher order autocorrelations, positive lag-1 autocorrelation is captured by these methods.

- If higher order autocorrelations exist, methods give different results.

- Negative lag-1 autocorrelation doesn't reduce MCMC-variance
  $\Rightarrow$ Higher order autocorrelations.

Relationship of ψ_{opt} and β

- ψ_{opt} was chosen by eye for each β.



- Linear relationship

$$\widehat{\psi_{opt}} = 2.767 \cdot \beta$$

at least over the range $0.01 \leq \beta \leq 0.24$.

Convergence diagnostics

Geweke's Z-Score

Two disjoint subchains are selected:
((g_1) = first 10%, (g_2) = last 50%)

Variance of subchains is estimated as before by using spectral density.

If g is stationary

$$Z = \frac{\bar{g}_{M_1} - \bar{g}_{M_2}}{\sqrt{\frac{1}{M_1}\widehat{\tilde{f}_{g_1}(0)} + \frac{1}{M_2}\widehat{\tilde{f}_{g_2}(0)}}} \xrightarrow{\mathcal{D}} N(0,1),$$

holds for $M \to \infty$.

If difference is too large, $\bar{g}_{M_1}$ is assumed to be biased.

(Geyer, 1992, Brooks & Roberts, 1996)

Convergence diagnostics

Heidelberger & Welch

Also based on ergodic mean of subchains.

Let $S_0 = 0,\ S_M = \sum_{m=1}^{M} g^{(m)},\ [r]$ be the greatest integer $\leq r$ and

$$B_M(t) = \frac{S_{[Mt]} - [Mt]\bar{g}_M}{\sqrt{M\tilde{f}_g(0)}},\quad 0 \leq t \leq 1,$$

then unter stationarity $B_M = \{B_M(t)\}_{0 \leq t \leq 1}$ is distributed approx. as a Brownian bridge.

The Cramer-von Mises statistics

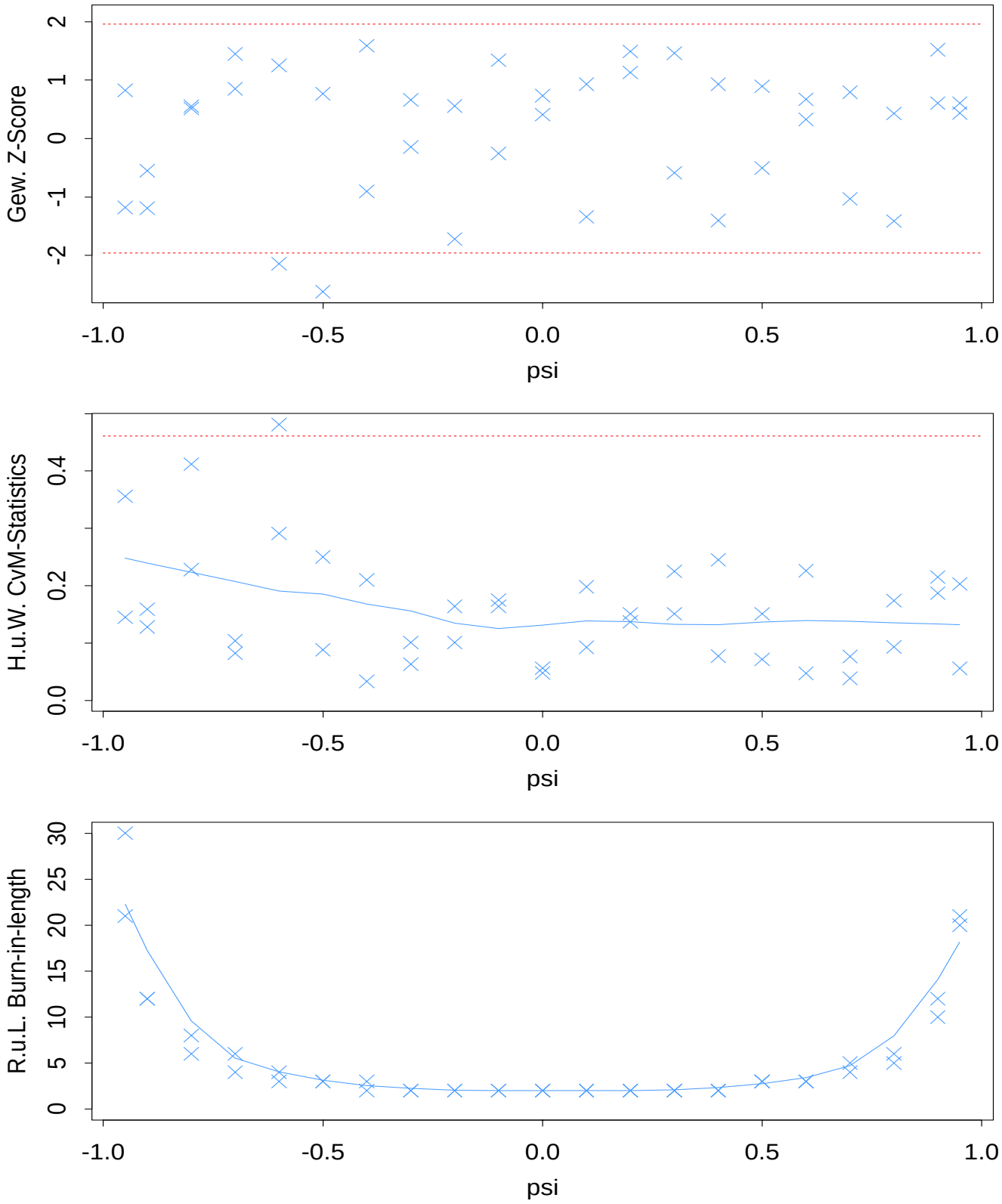
$$\int_0^1 (B_M(t))^2 dt$$

is used to test the hypothesis with a critical value at 5% level of 0.461.

If hypothesis is rejected, the first 10% are omitted.

(Heidelberger & Welch, 1983)

Convergence diagnostics for $\beta = 0.01$

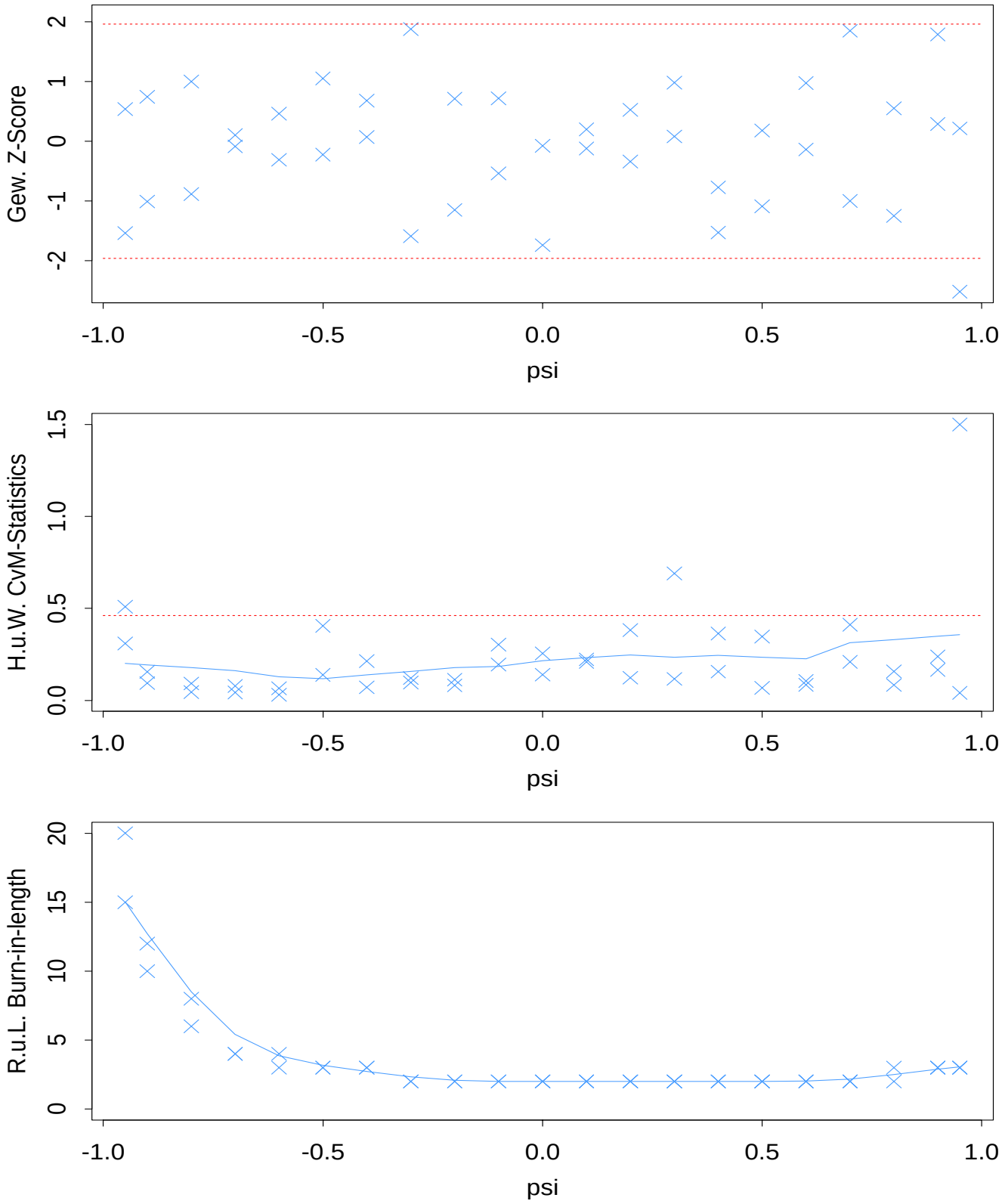


Top graph: Geweke's Z-Score.

Middle graph: Heidelberger and Welch's CvM-Statistics.

Bottom graph: Rafferty and Lewis' Burn-in-length.

Convergence diagnostics for $\beta = 0.1$

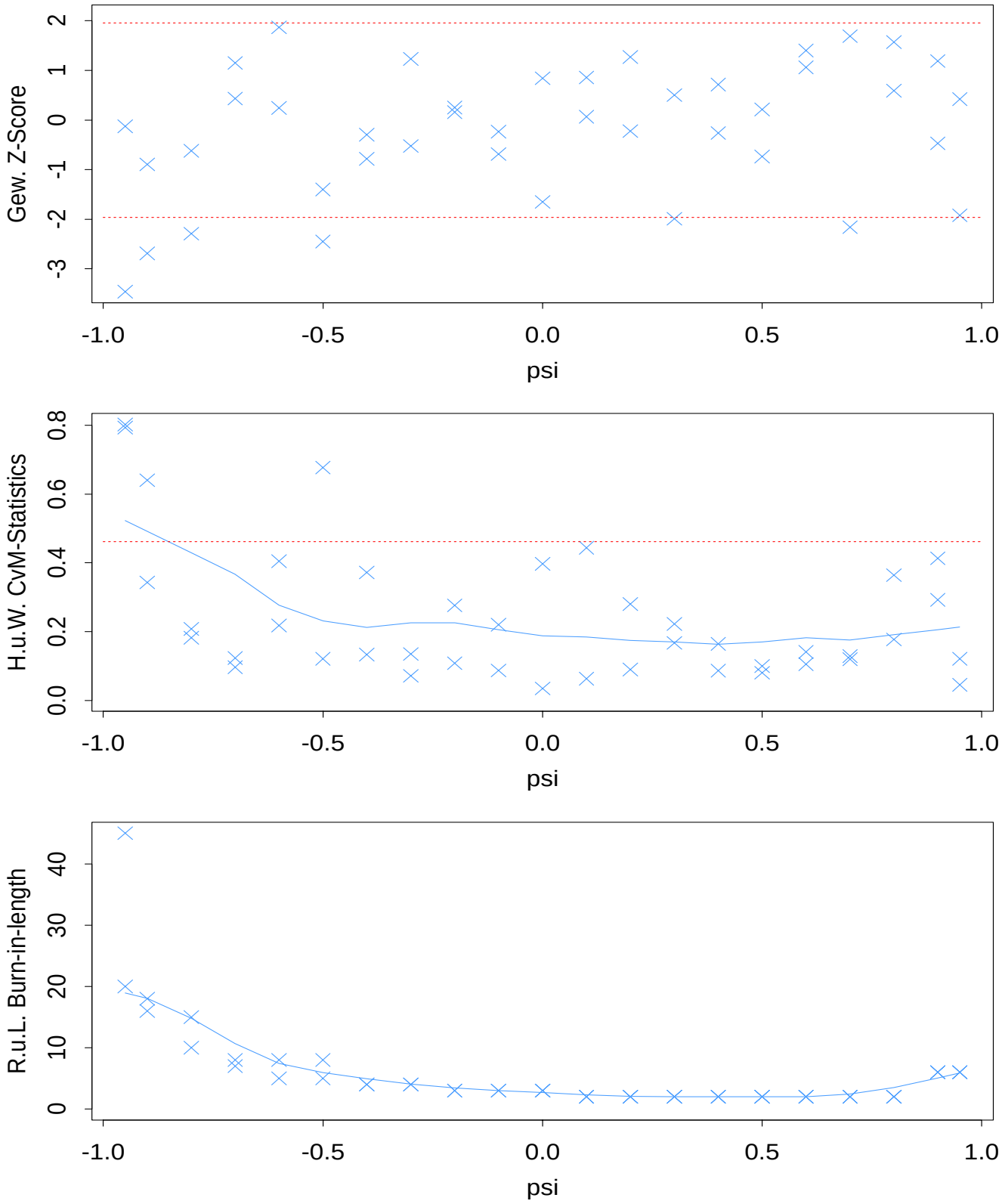


Top graph: Geweke's Z-Score.

Middle graph: Heidelberger and Welch's CvM-Statistics.

Bottom graph: Rafterty and Lewis' Burn-in-length.

Convergence diagnostics for $\beta = 0.2$

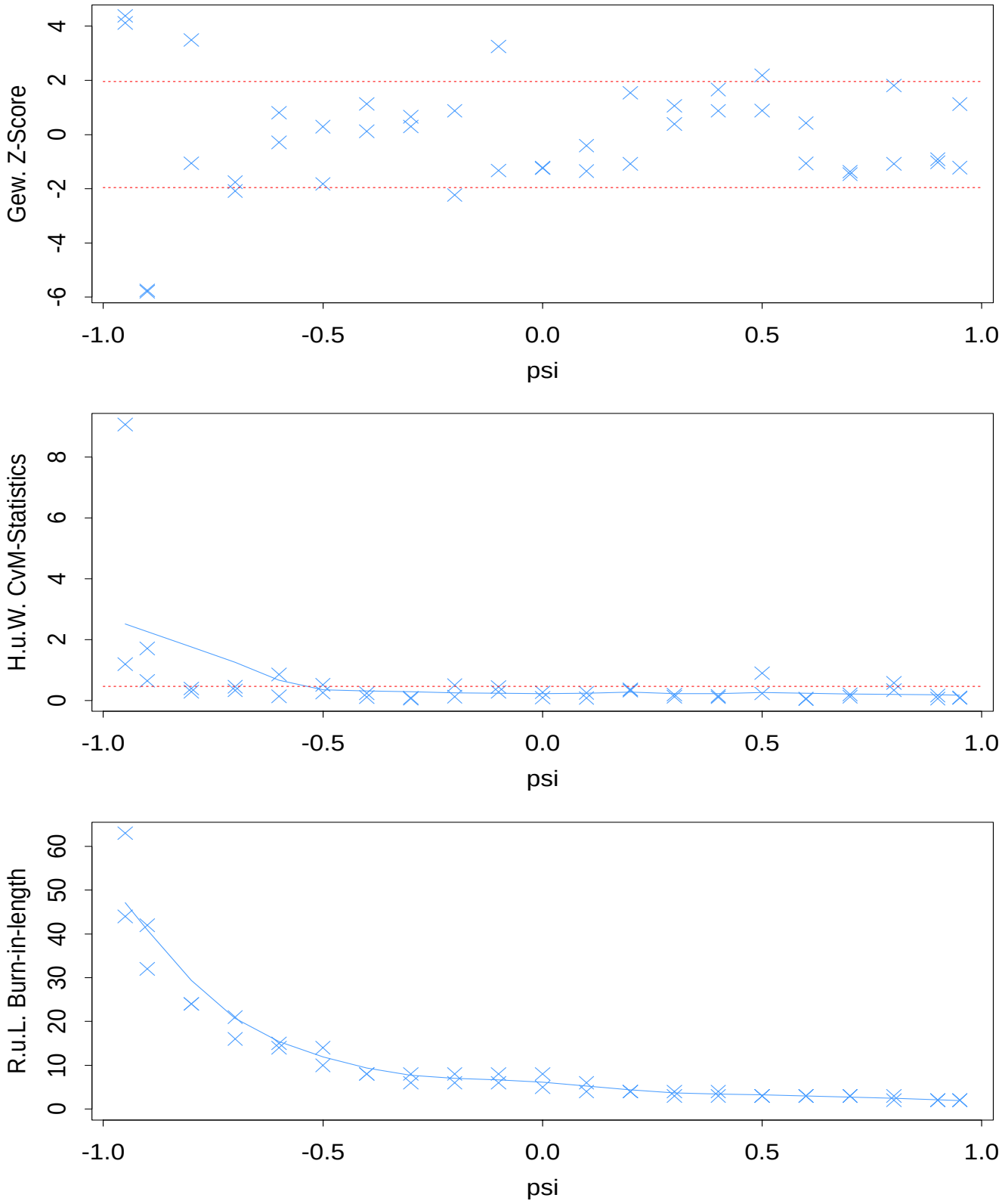


Top graph: Geweke's Z-Score.

Middle graph: Heidelberger and Welch's CvM-Statistics.

Bottom graph: Rafferty and Lewis' Burn-in-length.

Convergence diagnostics for $\beta = 0.24$



Top graph: Geweke's Z-Score.

Middle graph: Heidelberger and Welch's CvM-Statistics.

Bottom graph: Rafferty and Lewis' Burn-in-length.

Properties of convergence diagn.-tools

- Z-Score and CvM-Statistics should give similar results as both are focussed on the beginning of the trajectory.

- Cowles & Carlin (1994) found the Z-Score to be more conservative.

- In our experiments we observed high correspondence.

- Rafterty & Lewis' method gives suggestion for Burn-in-length, but no test decision.

- Suggested Burn-in-lengths are rather short.

- CvM-Statistics and Burn-in-lengths are appropriate for smoothing over the parameter space.

Conclusions

- Antithetic GS is easy to implement and provides substantial efficiency gain.

- For GMRFs antithetic GS can compete with i.i.d. sampling.

- For typical parametrization of GMRF ($\beta = 0.2$) ordinary GS has only 60% efficiency.

- A rule for the choice of the antithetic parameter was developed.

- CODA-tools are rather general and easy to apply to univariate Markov chains.

- All CODA-tools should be considered. Differences between methods can give additional information.