# On Quantile Estimation and MCMC Convergence

S.P. Brooks and G.O. Roberts[†]

**Abstract**

In this paper we examine the method of Raftery and Lewis (1992) for estimating the convergence of MCMC samplers, when functionals of interest are in the form of parameter quantiles.

This method has proved to be one of the most popular methods for diagnosing convergence of MCMC algorithms. However, it is commonly mis-applied to problems where quantiles are not of primary interest. We show how the method can be misleading in this case, and that it can seriously underestimate the true length of the burn-in.

We provide a number of examples, comparing the convergence rate of the chain in respect of a particular quantile, with that of the true convergence rate of the original chain. In particular we show how, in the case of the independence sampler, the two convergence rates are identical if the quantile of interest is chosen to be at an extreme of an appropriately re-ordered state space.

**Keywords**: Markov chain Monte Carlo; Convergence Diagnosis; Independence Sampler

## 1 Introduction

Markov Chain Monte Carlo (MCMC) methods have revolutionised the computation of Bayesian statistics and there is a considerable amount of work being performed in this area, see Gilks *et al* (1996). Arguably, one of the most difficult problems in the implementation of MCMC methods is associated with deciding whether or not stationarity has been achieved.

In practice, we base inference upon only a portion of the sampler output, discarding observations during an initial transient or *burn-in* phase. The remaining observations are then assumed to have been generated under the stationary density. In the absence of any general techniques for *a priori* prediction of the length of this burn-in period, it is necessary to carry out some form of statistical analysis in order to assess convergence to stationarity. There are many methods which have appeared in the literature, but the two most popular are those of Raftery and Lewis (1992) and Gelman and Rubin (1992). See Brooks and Roberts (1997) for a review of existing methods.

In this paper, we focus on the method of Raftery and Lewis (1992), and the fact that it is commonly mis-applied in the literature. In particular, we examine how badly the method can go wrong if inappropriately applied, and provide a result to show that, at least for the independence sampler, the method can be accurate at least for a suitably chosen partition of the state space.

[†]School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, and Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge, CB2 1SB. The authors wish to express their gratitude to Bill Fitzgerald, Adrian Raftery and Richard Smith, for their insightful comments and suggestions.

## 2 The Original Method

Raftery and Lewis (1992) consider the problem of calculating the number of iterations necessary to estimate a posterior quantile from a single run of a Markov chain. Their diagnostic procedure is based upon the estimation of a particular quantile for some scalar functional $\theta(\boldsymbol{X})$. The quantile $q$ corresponds to a value $u$ such that

$$\mathbb{P}(\theta(\boldsymbol{X}) \leq u) = q,$$

the probability being taken with respect to the stationary measure of the Markov chain. We use this value of $u$ to partition the state space into two distinct sets,

$$S_u = \{\boldsymbol{X} : \theta(\boldsymbol{X}) \leq u\} \text{ and } S_u^c = \{\boldsymbol{X} : \theta(\boldsymbol{X}) > u\}, \tag{1}$$

and assume that, for large $s$, the semi-Markov process $Z_s^t = I_{(\theta^{1+(t-1)s} \leq u)}$ is approximated by a 2-state Markov chain with transition probability matrix given by

$$P_u = \begin{pmatrix} 1-\alpha_u & \alpha_u \\ \beta_u & 1-\beta_u \end{pmatrix},$$

where $\alpha_u = \mathbb{P}(Z_s^t = 1 | Z_s^{t-1} = 0)$ and $\beta_u = \mathbb{P}(Z_s^t = 0 | Z_s^{t-1} = 1)$. We shall refer to this 2-state chain as the *reduced* chain for the original $\theta$ process.

Now suppose that we require that $\mathbb{P}(Z_s^t = i | Z_s^0 = j)$ be within $\epsilon$ of $\pi_i$, $\forall\, i, j = 0, 1$, then Raftery and Lewis (1992) show that this condition is satisfied when

$$t = t^* = \frac{\log\left(\frac{\epsilon(\alpha_u+\beta_u)}{\max(\alpha_u,\beta_u)}\right)}{\log|\rho_u|},$$

where $\rho_u = 1 - \alpha_u - \beta_u$ is the convergence rate of the reduced chain. Thus, $n_0 = t^* s$ is the length of the required burn-in period.

To implement this method, we run the single chain sampler for an initial number of iterations, which we denote by

$$n = n_{min} = q(1-q)\left(\frac{\Phi^{-1}[\frac{1}{2}(1+p)]}{r}\right)^2, \tag{2}$$

where $\Phi^{-1}$ is the inverse standard normal cumulative distribution function. From these iterations we estimate $\alpha_u$ and $\beta_u$ in order to determine the number of additional runs required for each scalar functional of interest. The method can be iterated so that after the required number of iterations have been completed, we can re-evaluate $\alpha_u$ and $\beta_u$ more accurately to check our preliminary findings.

Raftery and Lewis argue that often a particular functional of interest is a quantile function. Thus, the method may be applied to the output of the chain to see whether or not the chain has converged in terms of that quantile alone. The authors go on to suggest that this method might be applied to a range of different quantiles and that the largest burn-in estimate be taken as the value to use. This has been mis-interpreted by many to mean that by applying the method to a variety of quantiles, some form of estimate of the convergence rate of the original chain may be obtained.

In this paper, we present a result which examines the extent to which the convergence properties of the original Markov chain $X$ can be implied from those of the reduced chain, and examine how well the convergence rate of the reduced Markov chain $\rho_u$ approximates $\rho$, for different values of $u$, by obtaining explicit approximations to the transition probabilities $\alpha_u$ and $\beta_u$ associated with particular $u$-values, at stationarity. It is clear from (2) that the best value of $u$, in terms of minimising $n_{min}$, is in the tails of the state space, corresponding to $q$ close to zero or one. However, is this a suitable choice in terms of the "closeness" of $\rho_u$ to $\rho$?

In order to keep the formulae tractable, we concentrate on one particularly simple MCMC sampler, namely the Independence Sampler, though any conclusions we draw may well be more widely applicable to more complex samplers. We shall generally take $\theta$ to be the identity function for scalar $X$, and also take the thinning parameter $s = 1$, so that $Z_s^t = Z^t = I_{(X^t \leq u)}$.

# 3 A General Result for the Independence Sampler

In the case of the Independence Sampler, where the normalised density $\pi$ is known, if we let $w(x) = \pi(x)/q(x)$ and $w^* = \sup_x\{w(x)\}$, then if $w^* < \infty$, Liu (1994) shows that the convergence rate of the chain is given by $1 - 1/w^*$. Thus, in the case of the Independence Sampler, we have an explicit value for $\rho$, with which we can compare our estimator, $\hat{\rho}$. This Independence Sampler result is further discussed in Smith and Tierney (1996).

**Theorem 3.1** *Let $X^1, X^2, \ldots$ denote the output from an Independence Sampler with proposal $q$ and which converges to the stationary density $\pi$ at rate $\rho$. Further, let $w(x) = \pi(x)/q(x)$, and take a partition of the state space $S$, of the form*

$$S_\epsilon = \{x : w(x^*) - w(x) \leq \epsilon\},$$

*for some $\epsilon > 0$, where $x^*$ maximises $w$ over the state space, $S$. Then, if $\rho_\epsilon$ denotes the convergence rate of the reduced chain induced by the partition $S_\epsilon$, then $\rho_\epsilon \to \rho$, as $\epsilon \to 0$.*

**Proof**: The convergence rate of the reduced chain is given by $\rho_\epsilon = 1 - \alpha_\epsilon - \beta_\epsilon$, where

$$\alpha_\epsilon = \frac{\mathbb{P}(X^{t+1} \in S_\epsilon | X^t \in S_\epsilon^c)}{\mathbb{P}(X^t \in S_\epsilon)} \text{ and } \beta_\epsilon = \frac{\mathbb{P}(X^{t+1} \in S_\epsilon^c | X^t \in S_\epsilon)}{\mathbb{P}(X^t \in S_\epsilon^c)}.$$

At stationarity,

$$\alpha_\epsilon = \frac{\int_{S_\epsilon}\int_{S_\epsilon^c} q(y)\pi(x)\min\left(1, \frac{w(y)}{w(x)}\right)dxdy}{\int_{S_\epsilon^c}\pi(x)dx} = \int_{S_\epsilon} q(y)dy.$$

Similarly,

$$\beta_\epsilon = \frac{\int_{S_\epsilon} q(y)dy \int_{S_\epsilon^c}\pi(x)dx}{\int_{S_\epsilon}\pi(x)dx}$$

and therefore

$$\rho_\epsilon = 1 - \int_{S_\epsilon} q(y)dy - \frac{\int_{S_\epsilon} q(y)dy \int_{S_\epsilon^c} \pi(x)dx}{\int_{S_\epsilon} \pi(x)dx}.$$

However, since $S_\epsilon \to \{x^*\}$ as $\epsilon \to 0$, it is clear that $\int_{S_\epsilon} q(y)dy \to 0$, $\int_{S_\epsilon^c} \pi(x)dx \to 1$ and

$$\frac{\int_{S_\epsilon} q(y)dy}{\int_{S_\epsilon} \pi(x)dx} \longrightarrow \frac{q(x^*)}{\pi(x^*)} = \frac{1}{w(x^*)} \text{ as } \epsilon \to 0.$$

□

Thus, by taking $\theta \equiv w$ in (1), or equivalently any function $\theta$ which results in an identical (or exact reciprocal) ordering of the state space under $w$, we can obtain an estimate of the convergence rate which is arbitrarily close to the true value, by choosing a partition point arbitrarily close to $w^*$, ie; by setting $u = w(x^*) - \epsilon$, for some $\epsilon > 0$.

Hence, by appropriate re-parameterisation of the state space, it is possible to obtain an arbitrarily accurate estimate of the convergence rate of the original chain, via the method of Raftery and Lewis. This is particularly useful, in the case where $\pi$ is known only up to a multiplicative constant, in which case the exact convergence rate will not be known since the convergence rate calculation method of Smith and Tierney (1996) may not be applied.

# 4 Examples

## 4.1 A Normal Example

Here we use the Independence Sampler in order to simulate from $\pi \sim N(0,1)$, and take proposal $q \sim N(0,\sigma^2)$, with $\sigma > 1$ in order to ensure that $q(\cdot)$ has heavier tails than $\pi(\cdot)$. Given that $X^t = x$, we generate an observation $y$, from the proposal $q(y)$, and accepts it as the new state of the chain with probability

$$\begin{aligned} \alpha(x,y) &= \min\left(1, \frac{\pi(y)q(x)}{\pi(x)q(y)}\right) \\ &= \begin{cases} \exp\left\{-\frac{1}{2}(y^2 - x^2)(1 - \frac{1}{\sigma^2})\right\} & |y| > |x| \\ 1 & \text{else} \end{cases}. \end{aligned} \tag{3}$$

Now asymptotically, $X^t$ has distribution $\pi$, so that at stationarity, it can be shown that

$$\hat{\alpha}_u = \frac{1}{\Phi(u)} \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}\sigma} \left[\sigma e^{-\frac{1}{2}x^2} \Phi\left(\frac{|x|}{\sigma}\right) - e^{-\frac{1}{2}x^2/\sigma^2} \Phi(|x|)\right] dx + \frac{\Phi\left(\frac{u}{\sigma}\right)}{\Phi(u)} \left[1 - \Phi(u)\right].$$

Similarly,

$$\hat{\beta}_u = \frac{1}{1 - \Phi(u)} \int_{u}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \left[\sigma e^{-\frac{1}{2}x^2} \Phi\left(\frac{|x|}{\sigma}\right) - e^{-\frac{1}{2}x^2/\sigma^2} \Phi(|x|)\right] dx + \Phi(u) \left[1 - \Phi\left(\frac{u}{\sigma}\right)\right]$$

leading to an explicit expression for $\hat{\rho}_u = 1 - \hat{\alpha}_u - \hat{\beta}_u$.

Recall that for the Independence Sampler, if

$$w^* = \sup_x \{w(x)\}, \text{ where } w(x) = \frac{\pi(x)}{q(x)},$$

and the normalisation constant for $\pi$ is known, then the true convergence rate is given by $1 - 1/w^*$. In this case, $w^* = \sigma$, so the true convergence rate is given by $\rho = 1 - 1/\sigma$.

**Figure 1 about here**

We can plot $\hat{\rho}_u$ against $u$ for a range of $u$-values, and these are reproduced in Figure 1, for different values of $\rho$ (or equivalently, $\sigma$). Here, the optimal value of $u$ is at the mode ($u = 0,\ q = 0.5$), with the accuracy of the estimator $\hat{\rho}_u$ rapidly deteriorating as $u$ moves into the tails. However, for reasonable $u$-values, the underestimation is no worse than by a factor of 2. We can see that the estimate consistently underestimates the length of the burn-in and that this worsens as we take quantiles further into the tails.

In this example,

$$w(x) = \sigma \exp\left\{-\frac{1}{2}x^2\left(1 - \frac{1}{\sigma^2}\right)\right\}$$

so that the $x$'s and $w$'s have different orderings, due to the quadratic power of $x$ in the exponent, and the result of section 3 does not apply. However, note that if we take $\theta(x) = |x|$, and a proposal distribution taken to be $N(0, \sigma^2)$ with the same constraint, then this is equivalent to taking $\pi$ to be the standard normal distribution, constrained to take only non-negative values. It is easy to see that the acceptance probability is the same as that given in (3) and the estimators for $\alpha$ and $\beta$ simplify to

$$\hat{\alpha}_u = \frac{(2\Phi\left(\frac{u}{\sigma}\right) - 1)(2 - 2\Phi(u))}{2\Phi(u) - 1} \text{ and } \hat{\beta}_u = 2\Phi\left(\frac{u}{\sigma}\right) - 1.$$

**Figure 2 about here**

Figure 2 plots the corresponding value of $\hat{\rho}_u$ for values of $u$, and we can see that the convergence rate estimate approaches the correct value as $u$ approaches zero, since the $\theta$'s and the $w$'s have identical orderings, and the value of $x$ maximising $w(x)$ is $x^* = 0$. We can compare this with Figure 1, which showed the corresponding graph with $\theta$ set to be the identity function.

# 5 A Gamma Example

Here we take the Independence Sampler, with a Gamma target and proposal distribution, so that

$$\pi(\cdot) \sim ?\,(\alpha, \beta) \text{ and } q(\cdot) \sim ?\,(\alpha, \lambda) \qquad \lambda > \beta \text{ and } \alpha \in \mathbb{N},$$

where we take $\lambda > \beta$ to ensure that the proposal density has "heavier" tails than $\pi$.

Given that the chain is in state $x$ at time $t$, a candidate observation $y$, from $q(y)$, is accepted as the next state of the chain with probability

$$\alpha(x,y) \;=\; \min\left(1, \exp[\{x-y\}\{\frac{1}{\beta}-\frac{1}{\lambda}\}]\right)$$

We calculate $\hat{\alpha}_u$ and $\hat{\beta}_u$ as follows.

$$\hat{\alpha}_u \;=\; \frac{\mathbb{P}(X^{t+1} > u,\ X^t \le u)}{\mathbb{P}(X^t \le u)},$$

where

$$\mathbb{P}(X^t \le u) \;=\; 1 - \sum_{j=1}^{\alpha} \left(\frac{u}{\beta}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\beta}, \text{ and}$$

$$\mathbb{P}(X^{t+1} > u, X^t \le u) \;=\; \left(1 - \sum_{j=1}^{\alpha} \left(\frac{u}{\lambda}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\lambda}\right) \left(\sum_{j=1}^{\alpha} \left(\frac{u}{\beta}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\beta}\right).$$

Therefore

$$\hat{\alpha}_u = \frac{\left(1 - \sum_{j=1}^{\alpha} \left(\frac{u}{\lambda}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\lambda}\right) \left(\sum_{j=1}^{\alpha} \left(\frac{u}{\beta}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\beta}\right)}{1 - \sum_{j=1}^{\alpha} \left(\frac{u}{\beta}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\beta}}.$$

Similarly,

$$\hat{\beta}_u = \frac{\left(1 - \sum_{j=1}^{\alpha} \left(\frac{u}{\lambda}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\lambda}\right) \left(\sum_{j=1}^{\alpha} \left(\frac{u}{\beta}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\beta}\right)}{\sum_{j=1}^{\alpha} \left(\frac{u}{\beta}\right)^{j-1} \frac{1}{(j-1)!} e^{-u/\beta}}.$$

Thus, we obtain an explicit expression for $\hat{\rho}_u = 1 - \hat{\alpha}_u - \hat{\beta}_u$. Note that $\sum_{j=1}^{\alpha} \left(\frac{u}{\nu}\right)^{j-1} \frac{1}{(j-1)!}$ is the $\alpha$th order Taylor series expansion of $e^{u/\nu}$, and we can use this result to show that $\hat{\rho}_u \to 1 - \left(\frac{\beta}{\lambda}\right)^{\alpha}$ as $u \to 0$. As before, we use the result of Liu (1994) to calculate the true rate of convergence, in this case

$$w(x) = \left(\frac{\lambda}{\beta}\right)^{\alpha} e^{-x(1/\beta - 1/\lambda)}$$

so that $w^* = \left(\frac{\lambda}{\beta}\right)^{\alpha}$ and $\rho = 1 - \left(\frac{\beta}{\lambda}\right)^{\alpha}$. Thus, $\hat{\rho}_u \to \rho$ as $u \to 0$.

**Figure 3 about here**

Figure 3 plots $\hat{\rho}_u$ against $q$, the proportion of the density below $u$, for four different values of $\rho$, and fixed $\alpha = \beta = 1$. These plots confirm that the best convergence rate estimate is approached as $u \to 0$, which does not correspond to the mode of $\pi$. Similar

plots can be produced for other $\alpha$ and $\beta$ values. We can also see that the slower the rate of convergence, the worse the estimator becomes, as we observed in the normal example of the previous section.

Clearly, for most of the range of $q$ (and thus $u$) the estimated length of the burn-in period is acceptable, but for $u$-values in the upper tail the accuracy of the estimate quickly deteriorates, and the corresponding burn-in estimate tends to infinity as we look at quantiles further into the upper tail of the distribution. This means that the estimated length of the burn-in period tends to zero as $q$ tends to 1.

This example illustrates that the optimal $u$ need not necessarily be at the mode of $\pi$ and can be explained by the result of section 3, as follows. Here,

$$w(x) = \left(\frac{\lambda}{\beta}\right)^{\alpha} \exp\left\{-x(1/\beta - 1/\lambda)\right\}, \tag{4}$$

so that the $x$'s have an exact reciprocal ordering to the $w$'s. We also have that

$$w^* = \left(\frac{\lambda}{\beta}\right)^{\alpha}$$

corresponding to $x = 0$ in (4), so that by taking $u$ arbitrarily close to zero, we get an arbitrarily close approximation to the true convergence rate.

# 6 Discussion

The results of this paper, appear to suggest that the estimator $\hat{\rho}_u$ consistently underestimates the true convergence rate. This may be partially explained by looking at our estimator in terms of the capacitance ideas described by Sinclair and Jerrum (1988), who partition the state space into two distinct sets, $S$ and $S^c$. They then define the *capacitance* of the chain by

$$\phi = \inf_S \frac{\int_S \int_{S^c} \pi(x)\mathcal{K}(x,y)dydx}{\min\left\{\pi(S), \pi(S^c)\right\}}.$$

Cheeger's inequality may then be used to provide bounds on the convergence rate of the chain, $\rho$, given by

$$1 - 2\phi \leq \rho \leq 1 - \frac{\phi^2}{2}, \tag{5}$$

so that $1 - \phi^2/2$ provides an upper bound on the convergence rate of the original chain.

Now, by taking a quantile $u$ and in reducing the original chain to only two states, we introduce a partition of the state space into two distinct sets $S_u$ and $S_u^c$, say. The capacitance associated with such a partition is given by $\phi_u = \max(\alpha_u, \beta_u)$, with $\phi_u \geq \phi$, since $\phi$ is the infimum over all possible partitions and $\phi_u$ is restricted to partitions of the form $S = [u, \infty)$.

The reduced, two-state, chain will have convergence rate $\rho_u$ say, and will similarly be bounded by

$$1 - 2\phi_u \leq \rho_u \leq 1 - \frac{\phi_u^2}{2}.$$

Clearly, $\rho \geq 1 - 2\phi \geq 1 - 2\phi_u$, so that we also obtain a bound on the convergence rate of the *original* chain via the capacitance of the reduced chain.

Note also, that the interval for $\rho_u$ has both upper and lower bounds below the corresponding bounds for the convergence rate of the original chain, given in (5). Whilst this provides us with no hard and fast rule, it is clear that the proximity of $\rho_u$ (and therefore our estimator $\hat{\rho}_u$) to $\rho$ will depend upon $u$ and that, in general, we might expect the value of $\rho_u$ to lie below that of $\rho$. The examples presented in this paper appear to support this conclusion. In the case of the Independence Sampler, it is possible to get equality between $\rho_u$ and $\rho$, but there is no obvious generalisation of this result to other samplers.

Thus, some care should be taken in selecting a suitable $q$ (or $u$) value, for the problem at hand. In particular, it is clear that the routine use of $q = 0.025$, suggested both in the literature and within the S-Plus and Fortran code that the original authors distribute, should not be adopted as a general rule, since it may lead to a strong underestimate of the true length of the burn-in period. When interested in a number of quantiles, it may be most sensible to calculate $\hat{\rho}_u$ for a number of different $u$-values and take the largest of the resulting estimated burn-in lengths, but in the case where quantiles themselves are not of interest, this method should be used with caution.

# References


Brooks, S. P. and G. O. Roberts (1997), Diagnosing Convergence of Markov Chain Monte Carlo Algorithms. Technical report, University of Cambridge

Gelman, A. and D. Rubin (1992), Inference from Iterative Simulation using Multiple Sequences. *Statistical Science* **7**, 457–511

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996), *Markov Chain Monte Carlo in Practice*. Chapman and Hall

Liu, J. S. (1994), Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling. Technical report, Harvard University

Raftery, A. E. and S. M. Lewis (1992), How Many Iterations in the Gibbs Sampler? In J. M. Bernardo, A. F. M. Smith, A. P. Dawid and J. O. Berger (eds.), *Bayesian Statistics 4*, Oxford University Press

Sinclair, A. J. and M. R. Jerrum (1988), Conductance and the Rapid Mixing Property for Markov Chains: The Approximation of the Permanent Resolved. In *Proceedings of the 20th annual ACM symposium on the Theory of Computing*

Smith, R. L. and L. Tierney (1996), Exact Transition Probabilities for the Independence Metropolis Sampler. Technical report, Statistical Laboratory, Cambridge

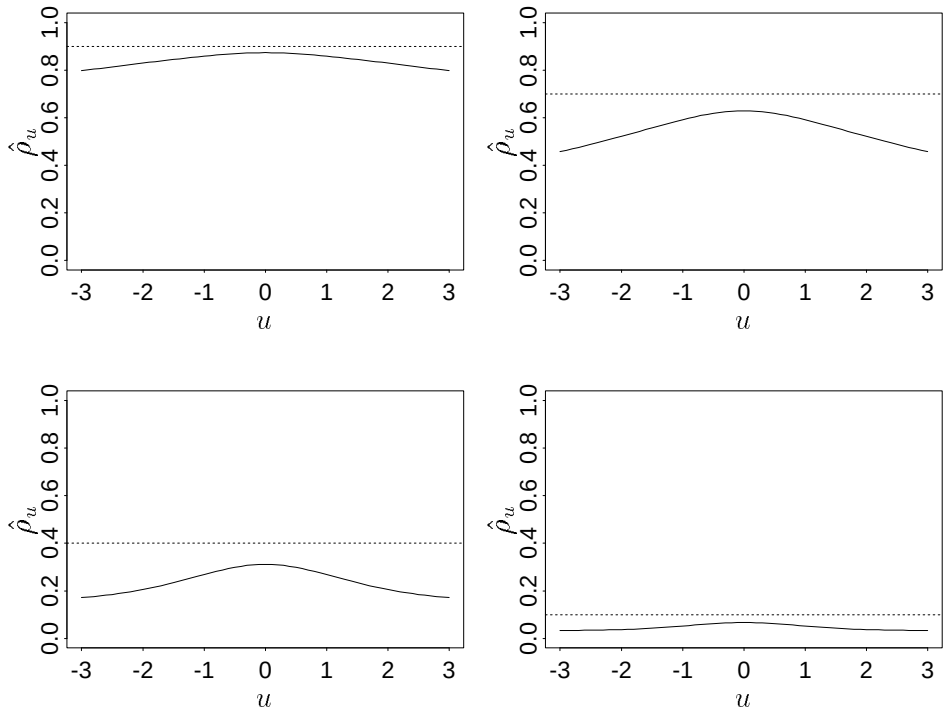


Figure 1: $\hat{\rho}_u$ (solid) for values of $u \in [-3, 3]$ in the normal example. The four plots represent the $\rho$-values 0.9, 0.7, 0.4 and 0.1 (Dashed line).

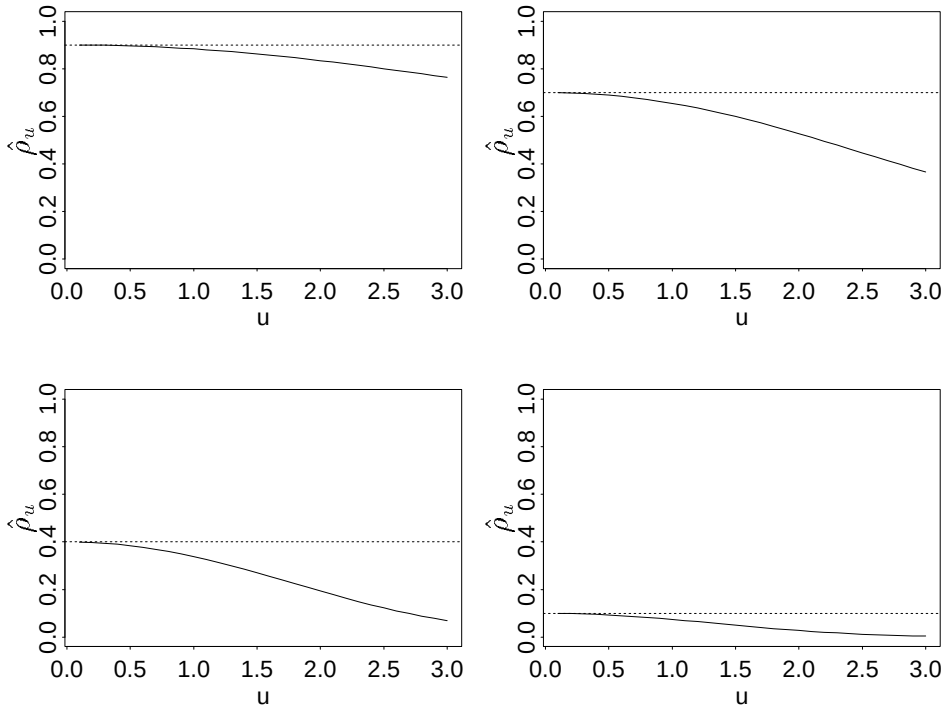


Figure 2: $\hat{\rho}_u$ (solid) for values of $q$ for the normal example, taking $\theta(x) = |x|$. The four plots represent the $\rho$-values 0.9, 0.7, 0.4 and 0.1.

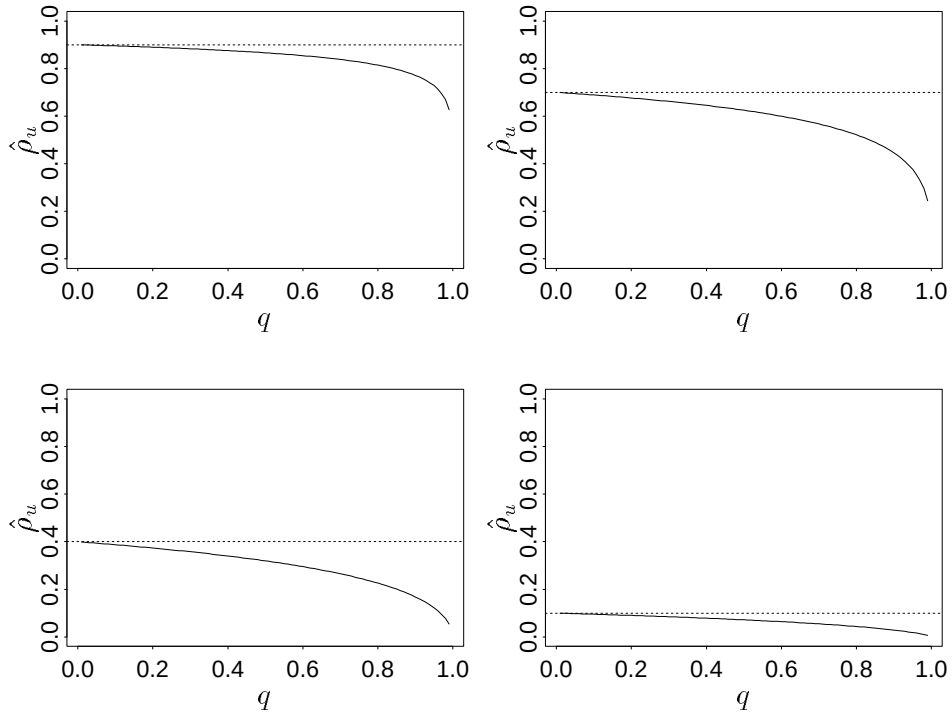


Figure 3: $\hat{\rho}_u$ (solid) for values of $q$. The four plots represent the $\rho$-values 0.9, 0.7, 0.4 and 0.1. A fixed value of $\alpha = 1$ was taken so that the Gamma is reduced to an exponential distribution.