Penalised complexity priors for stationary autoregressive processes

Sigrunn Holbek Sørbye[1] and Håvard Rue[2]

[1]Department of Mathematical Sciences, UiT The Arctic University of Norway, 9037 Tromsø, Norway. e-mail: sigrunn.sorbye@uit.no
[2]CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. e-mail: haavard.rue@kaust.edu.sa

Abstract

The autoregressive process of order p (AR(p)) is a central model in time-series analysis. A Bayesian approach requires the user to define a prior distribution for the coefficients of the AR(p) model. Although it is easy to write down some prior, it is not at all obvious how to understand and interpret the prior distribution, to ensure that it behaves according to the users prior knowledge. In this paper, we approach this problem using the recently developed ideas of penalised complexity (PC) priors. These prior distributions have important properties like robustness and invariance to reparameterisations, as well as a clear interpretation. A PC prior is computed based on specific principles, where model component complexity is penalised in terms of deviation from simple base model formulations. In the AR(1) case, we discuss two natural base model choices, corresponding to either independence in time or no change in time. The latter case is illustrated in a survival model with possible time-dependent frailty. For higher-order processes, we propose a sequential approach, where the base model for AR(p) is the corresponding AR($p-1$) model expressed using the partial autocorrelations. The properties of the new priors are compared with reference priors in a simulation study.

Keywords: AR(p), `R-INLA`, prior selection, reference prior, robustness.

1 Introduction

Autoregressive (AR) processes are widely applied to model time-varying stochastic processes, for example within finance, biostatistics and natural sciences (Brockwell and Davis, 2002; Chatfield, 2003; Prado and West, 2010). Applications also include Bayesian model formulations, often combined with Markov chain Monte Carlo computations to perform posterior and predictive inference (Albert and Chib, 1993; Chib, 1993; Barnett et al., 1996). Particularly, AR processes are useful to model underlying latent dependency structure and they make up important building blocks in complex hierarchical models, for example analysing spatial data (Lesage, 1997; Sahu et al., 2007; Sahu and Bakar, 2012).

In fitting an AR(p) process using a Bayesian approach, it is necessary to select prior distributions for all model parameters. A simple choice is to assign uniform priors to the regression coefficients (Zellner, 1971; DeJong and Whiteman, 1991), but this is not optimal neither for the first-order nor higher-order processes (Berger and Yang, 1994). Alternative approaches to derive objective priors are given by Liseo and Macaro (2013), who provide a general framework

to compute both Jeffreys and reference priors using the well-known partial autocorrelation function (PACF) parameterisation (Barndorff-Nielsen and Schou, 1973). Stationarity of the AR(p) process is equivalent to choosing the partial autocorrelations within a p-dimensional unit hypercube. In general, Jeffreys priors are invariant to reparameterisations, while reference priors are not. Liseo and Macaro (2013) recommend reference priors, at least when the order of the AR process is smaller or equal to 4. For higher-order processes, calculation of the reference prior is numerically cumbersome and requires extensions of their suggested numerical approximation.

This paper derives and investigates penalised complexity (PC) priors (Simpson et al., 2017) for the partial autocorrelations of stationary AR processes of any finite order. In general, a PC prior distribution is computed based on specific underlying principles, in which a model component is seen as a flexible parameterisation of a simple base model structure. The main idea is to assign a prior distribution to a measure of divergence from the flexible version of the component to its base model and the PC prior for the relevant parameter is derived by transformation. In the AR(1) case, we can derive two different PC priors by considering two different base models. Define an AR(1) process of length n by $x_t = \phi x_{t-1} + w_t$, $t = 1, \ldots, n$, where ϕ denotes the autocorrelation coefficient while w_t is a zero-mean Gaussian white noise process. One possible base model is to assume that the observations are independent, using white noise ($\phi = 0$) as a base model. Alternatively, we can view the limiting random walk case ($\phi = 1$) as a base model, representing no change in time. Which of these base models that represent a natural choice to derive the PC prior depends on the relevant application.

In the higher-order AR(p) case, we introduce a sequential approach to construct a PC prior for the pth partial autocorrelation, using the corresponding AR($p-1$) process as a base model. The resulting joint prior density for the partial autocorrelations is consistent under marginalisation, and each of the marginals can be adjusted according to a user-defined scaling criterion. The scaling is important and prescribes the degree of informativeness of the prior distribution. Here, we suggest to incorporate a scaling criterion using the variance of the one-step ahead forecast error, also allowing for different rates of shrinkage for each of the partial autocorrelations. The resulting prior distributions have good robustness properties and are also seen to have comparable frequentistic properties with reference priors.

The plan of this paper is as follows. PC priors and their properties are reviewed in Section 2. We derive PC priors for the coefficient of an AR(1) process in Section 3, using the two mentioned base models. PC priors are designed to prevent overfitting and this property is demonstrated for a real data example in Section 4, where an AR(1) process is used to model time-dependent frailty in a Cox proportional hazard model. Contrary to previous results (Fleming and Harrington, 2005; Yau and McGilchrist, 1998), the given data on chronic granulomatous disease do not seem to support the additional introduction of a time-varying frailty. Extension of the PC priors to higher-order AR processes is given in Section 5, including incorporation of interpretable scaling parameters to adjust the rate of shrinkage. Section 6 contains simulation results, comparing the performance of the PC and reference priors, while concluding remarks are given in Section 7.

2 Penalised complexity priors and their properties

The framework of PC priors (Simpson et al., 2017) represents a systematic and unified approach to compute prior distributions for parameters of model components with an inherit nested structure. A simple version of the model component is referred to as a base model, typically characterised by a fixed value of the relevant parameter, while the flexible version is seen as a function of the random parameter. The PC prior is computed to penalise deviation from the flexible model

to the fixed base model. This section gives a brief review on PC priors and their properties in the context of AR(p) processes.

2.1 A brief review on the principles underlying PC priors

The informativeness of PC priors is specified in terms of four main principles, stated in (Simpson et al., 2017). These principles are useful both to compute priors in a unified way and to understand their properties. The principles, summarised below, express support to Occam's razor, penalisation of model complexity using the Kullback-Leibler divergence, a constant rate penalisation and user-defined scaling.

1. Let $\pi(\boldsymbol{x} \mid \xi)$ denote the density of a model component $\boldsymbol{x}$ where we want to assign a prior distribution to the parameter ξ. A simpler structure of this model component is characterised by a density $\pi(\boldsymbol{x} \mid \xi = \xi_0)$, where ξ_0 is a fixed value. This model is referred to as a base model and in accordance with the principle of parsimony expressed by Occam's razor, the prior for ξ should be designed to give proper shrinkage to ξ_0. This implies that model simplicity is preferred over model complexity, and the prior will prevent overfitting.

2. For simplicity, we use the short-hand notation $f_1 = \pi(\boldsymbol{x} \mid \xi)$ and $f_0 = \pi(\boldsymbol{x} \mid \xi)$ to denote the flexible and base model, respectively. In order to characterise the complexity of f_1 compared with f_0, a measure of complexity between these two densities is computed. Specifically, a PC prior is derived using the Kullback-Leibler divergence (Kullback and Leibler, 1951),
$$\text{KLD}(f_1 \parallel f_2) = \int f_1(x) \log\left(\frac{f_1(x)}{f_0(x)}\right) dx,$$
which measures the information lost when the flexible model f_1 is approximated with the simpler model f_0. For zero-mean multi-normal densities, calculation of the Kullback-Leibler divergence simplifies to performing simple matrix computations on the covariance matrices as
$$\text{KLD}(f_1 \parallel f_0) \quad = \quad \frac{1}{2}\left(\text{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) - n - \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right)\right)$$
where $f_i \sim N(0, \boldsymbol{\Sigma}_i)$, $i = 0, 1$, while n is the dimension. To facilitate interpretation, the Kullback-Leibler divergence is transformed to a unidirectional distance measure
$$d(\xi) = d(f_1 \parallel f_0) = \sqrt{2\text{KLD}(f_1 \parallel f_0)}. \tag{1}$$
This is not a distance metric in the ordinary sense, but a quantity which is interpretable as a measure of distance from the flexible model f_1 to the base model f_0.

3. In choosing a prior distribution for the distance measure $d(\xi)$, it is natural to assume that the mode should be located at the base model while the density decays as the distance from the base model increases. The PC prior is derived based on a principle of constant rate penalisation,
$$\frac{\pi(d(\xi) + \delta)}{\pi(d(\xi))} = r^{\delta}, \quad d(\xi), \delta \geq 0, \tag{2}$$
where $r \in (0, 1)$. This implies that the relative change in the prior distribution for $d(\xi)$ is independent of the actual distance. Consequently, $d(\xi)$ has the exponential density, $\pi(d(\xi)) = \lambda \exp(-\lambda d(\xi))$, where $\lambda = -\ln(r)$. The corresponding PC prior for ξ follows

by a standard change of variable transformation. Other prior choices for the distance could also be investigated but a constant rate penalisation seems reasonable as it would be complicated to properly characterise different decay rates for different distances, see Simpson et al. (2017) for further discussion.

4. The rate λ characterises the shrinkage properties of the prior and it is important that this parameter can be chosen (implicitly) in an intuitive and interpretable way, for example by a user-defined probability statement for the parameter of interest. Simpson et al. (2017) suggest to determine λ by incorporating a probability statement of tail events, e.g.

$$P(Q(\xi) > U) = \alpha, \tag{3}$$

where U represents an assumed upper limit for an interpretable transformation $Q(\xi)$, while α is a small probability. Other scaling suggestions might be just as reasonable, depending on the specific application. In deriving PC priors for the partial autocorrelations of AR(p) processes, the rate parameter is derived by imposing that the variance of the one-step ahead forecast error should stabilise as the order of the process increases.

2.2 Important properties of PC priors in the context of AR processes

The given four principles provide a strategy to calculate prior distributions for model parameters in a systematic way, rather than turning to ad-hoc prior choices still often made in Bayesian literature. Also, the principles can be helpful to interpret the assumed prior information and how this influences posterior results.

A first important property of PC priors is invariance to reparameterisations. This follows automatically as the prior is derived based on a measure of divergence between models, which does not depend on the specific model parameterisation. We consider the invariance property to be particularly useful in the case of autoregressive processes, as these are typically parameterised either in terms of the regression coefficients, or by using the partial autocorrelations. The great benefit of using the partial autocorrelations is that these give an unconstrained set of parameter values, ensuring a positive definite correlation matrix. In contrast, the valid parameter space for the regression coefficients is rather complicated, especially for higher-order processes ($p > 3$).

Second, the PC priors are designed to shrink towards well-defined base models. In the setting of autoregressive processes, this implies that the priors will prevent overfitting, for example in terms of selecting an unnecessarily high order of the process. In addition, the base model can be chosen to reflect different simple structures of a model component, depending on the given application. For an AR(1) process, it is relevant to assume either no dependency, or no change in time, as simple base model formulations. For higher-order processes, we could also choose no correlation as a base model but this might cause too much shrinkage in many applications. As an alternative, we introduce a new sequential approach which defines a sequence of base models, reflecting the additional complexity in increasing the order of the fitted AR process. It is important to realise that the rate parameter in (2) plays a very important role as it governs shrinkage to the base model.

Third, PC priors are computationally simple and already implemented within the `R-INLA` framework (Rue et al., 2009; Martins et al., 2013), for different latent Gaussian model components. The priors are designed to have a clear interpretation as the informativeness of the priors is adjusted by user-defined scaling. Here, we will take advantage of this to allow for different rates of shrinkage for priors assigned to partial autocorrelations of different lags. In contrast, objective priors simply aim to incorporate as little information to the inference as possible.

3 PC priors for AR(1) using two different base models

A first-order autoregressive process can be defined by

$$x_t = \phi x_{t-1} + w_t, \quad w_t \sim N(0, \kappa^{-1}), \quad t = 2, \ldots, n,$$

where x_1 is assumed to be normally distributed with mean 0 and marginal precision $\tau = \kappa(1 - \phi^2)$, and the variables $\{w_t\}_{t=1}^n$ define a white noise process. The AR(1) model represents an important special case of general autoregressive processes, in which the dependency structure is completely specified by the autocorrelation coefficient ϕ. Using the framework of penalised complexity priors, ϕ is viewed as a flexibility parameter reflecting deviation from simple fixed base model formulations. In this section, we derive PC priors for ϕ both using no autocorrelation ($\phi = 0$) and no change in time ($\phi = 1$) as base models, and we suggest how these priors can be scaled. A real-data application using the latter base model is included in Section 4.

Note that we also use a penalised complexity prior for the precision parameter τ. Following Simpson et al. (2017), this prior is derived using infinite precision as a base model, which gives the type-2 Gumbel distribution

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp(-\lambda\tau^{-1/2}), \quad \lambda > 0. \tag{4}$$

The rate λ is inferred using the probability statement $P(1/\sqrt{\tau} > U) = \alpha$, where α is a small probability. The prior is scaled by specifying an upper limit U for the marginal standard deviation $1/\sqrt{\tau}$, in which the corresponding rate is $\lambda = -\log(\alpha)/U$. To make an intuitive choice for U, one can consider the marginal standard deviation after the precision τ is integrated out. For example, if $\alpha = 0.01$ this standard deviation is $0.31U$ (Simpson et al., 2017).

3.1 Base model: No dependency in time

In general, the correlation matrix of the first-order autoregressive process is $\boldsymbol{\Sigma}_1 = \left(\phi^{|i-j|}\right)$. Choosing no autocorrelation ($\phi = 0$) as a base model, the resulting process is white noise with correlation matrix equal to the identity matrix, $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$. By simple matrix calculations, the distance function (1) is seen to equal $d(\phi) = \sqrt{(1-n)\log(1-\phi^2)}$. Using the principle of constant rate penalisation (2), an exponential prior is assigned to $d(\phi)$ with rate $\lambda = \theta/\sqrt{n-1}$. The resulting prior distribution is invariant to n and by the ordinary transformation of variable formula, the PC prior for the one-lag autocorrelation is

$$\pi(\phi) = \frac{\theta}{2}\exp\left(-\theta\sqrt{-\ln(1-\phi^2)}\right)\frac{|\phi|}{(1-\phi^2)\sqrt{-\ln(1-\phi^2)}}, \qquad |\phi| < 1, \theta > 0. \tag{5}$$

The rate parameter θ is important as it influences how fast the prior shrinks towards the white noise base model. To infer θ, we need a sensible criterion which facilitates the interpretation of this parameter. Simpson et al. (2017) suggest to use a probability statement for an interpretable transformation of the parameter of interest, for example in terms of tail events as defined by (3). When the base model is $\phi = 0$, a reasonable alternative is to define such a tail event as large absolute correlations being less likely, i.e.

$$\text{Prob}(|\phi| > U) = \alpha.$$

This implies that $\theta = -\ln(\alpha)/\sqrt{-\ln(1-U^2)}$. The interpretation of this criterion is intuitive in the first-order case, but we find it difficult to use in practice for higher-order processes. An alternative scaling idea is presented in Section 5.2, where we consider the variance of the one-step forecast error as the order of the AR process is increased. We recommend the latter approach, as this is more intuitively implemented for general AR(p) processes.

3.2 Base model: No change in time

An alternative base model for the AR(1) process is to assume that the process does not change in time ($\phi = 1$). This represents a limiting random walk case, being a non-stationary and singular process. Consequently, a limiting argument is needed to derive the PC prior for ϕ.

Let $\mathbf{\Sigma}_1 = \left(\phi^{|i-j|}\right)$ and $\mathbf{\Sigma}_0 = \left(\phi_0^{|i-j|}\right)$ where ϕ_0 is close to 1 and $\phi < \phi_0$. In this case, the Kullback-Leibler divergence is

$$\text{KLD}(f_1(\phi) \parallel f_0) = \frac{1}{2}\left(\frac{1}{1-\phi_0^2}(n - 2(n-1)\phi_0\phi + (n-2)\phi_0^2) - n - (n-1)\ln\left(\frac{1-\phi^2}{1-\phi_0^2}\right)\right).$$

Considering the limiting value as $\phi_0 \to 1$, the distance

$$d(\phi) = \lim_{\phi_0 \to 1} \sqrt{2\text{KLD}(f_1(\phi) \parallel f_0)} = \lim_{\phi_0 \to 1} \sqrt{\frac{2(n-1)(1-\phi)}{1-\phi_0^2}} = c\sqrt{1-\phi}, \quad |\phi| < 1,$$

for a constant c that does not depend on ϕ. Since $0 \leq d(\phi) \leq c\sqrt{2}$, we assign a truncated exponential distribution to $d(\phi)$ with rate $\theta = \lambda/c$ and the resulting PC prior for ϕ is

$$\pi(\phi) = \frac{\theta \exp\left(-\theta\sqrt{1-\phi}\right)}{\left(1 - \exp\left(-\sqrt{2}\theta\right)\right)\ 2\sqrt{1-\phi}}, \quad |\phi| < 1. \tag{6}$$

Again, we need to suggest an intuitive criterion to scale the prior in terms of θ. This case requires separate consideration, as it cannot be seen as a special case of the approach in Section 5. One option is to make use of (3), and determine (U, α) in terms of the probability statement $\text{Prob}(\phi > U) = \alpha$. The solution to this equation is given implicitly by

$$\frac{1 - \exp\left(-\theta\sqrt{1-U}\right)}{1 - \exp\left(-\sqrt{2}\theta\right)} = \alpha,$$

provided that α is larger than the lower limit $\sqrt{(1-U)/2}$.

3.3 The PC priors versus the reference prior

The two alternative PC priors for the first-lag coefficient of an AR(1) process are illustrated in Figure 1, using rate parameter $\theta = 2$ in (5) and (6). As already explained, these two priors are designed to give shrinkage towards $\phi = 0$ and $\phi = 1$ respectively. In a given analysis, the aim would not be to compare these two priors but choose the base model that is more suitable, either reflecting no correlation or no change in time. Unless we have specific reasons to assume the latter, it seems natural to use no correlation as a base model for AR(1). This is also the case which will be extended to general pth order processes in Section 5.

Figure 1 also displays the reference prior defined by $\pi(\phi) = \frac{1}{\pi}(1-\phi^2)^{-1/2}$, $|\phi| < 1$ (Barndorff-Nielsen and Schou, 1973; Berger and Yang, 1994; Liseo and Macaro, 2013). In general, reference priors are designed to give objective Bayesian inference in the sense of being least informative in a certain information-theoretic sense (Berger et al., 2009). This implies that the data are given a maximum effect on the posterior estimates. In general, the reference prior is calculated to maximise a measure of divergence from the posterior to the prior. In the given AR(1) case, the reference prior for ϕ is calculated to maximise an asymptotic version of the expected Kullback-Leibler divergence, in practice performed using an asymptotic version of

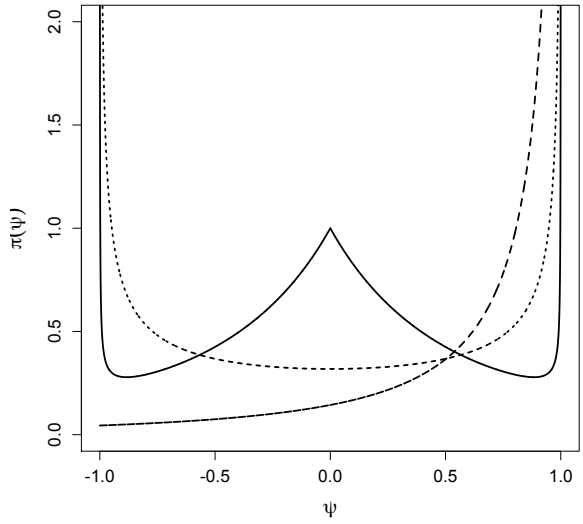


Figure 1: The PC priors for the coefficient ϕ of AR(1), using $\phi = 0$ (solid thick line) and $\phi = 1$ (dashed line) as base models. The rate parameters θ in (5) and (6) are set equal to 2 in both cases. For comparison we also include the reference prior for ϕ (dotted line).

the Fisher information matrix (Barndorff-Nielsen and Schou, 1973; Liseo and Macaro, 2013). The resulting reference prior is seen to be similar to the Jeffreys prior which is defined (up to a constant) by the square root of the determinant of the Fisher information matrix (Liseo and Macaro, 2013). Using a small rate parameter, the PC prior with base $\phi = 0$ will be quite similar to the reference/Jeffreys prior but for increasing rate parameters, the effect of shrinkage to 0 is increased. Note that a PC prior using $\phi = -1$ as the base model can be derived similarly as using the $\phi = 1$ base model.

4 Application: Modeling time-varying frailty with AR(1)

To demonstrate the use of the PC prior for the lag-one autocorrelation, we consider an example of a Cox proportional hazard model with time-varying frailty. The Cox proportional hazard model is a popular type of survival model that can be fitted to recurrent event data. It assumes that the time-varying hazard for the ith subject can be expressed as $h(t; i) = h_0(t) \exp(\eta_i)$, where the combined risk variable η_i in most cases depends on subject-specific covariates $\boldsymbol{z}_i$ and contributions from random effects/frailty. The function $h_0(t)$ is the baseline hazard, see Fleming and Harrington (2005) for further details and applications of the model. In the given example, our main focus is on the inclusion of a subject-specific and possibly time-dependent frailty term in η_i.

4.1 Dependent Gaussian random effects

A full Bayesian analysis of the Cox proportional hazard model requires a model for the baseline hazard. A natural choice is to consider the log baseline hazard as a piecewise constant function on small time intervals, and impose smoothness to penalise deviations from a constant, see for

example Fahrmeir and Tutz (2001, Sec 8.1.1) and Rue and Held (2005, Sec. 3.3.1). Let $[0, T]$ be the time interval of interest, and divide that interval into n equidistant (for simplicity) intervals $0 < t_1 < t_2 < \cdots < t_{n-1} < T$. Let h_j, $j = 1, \ldots, n$ denote the log baseline hazard in the jth interval. The first order random walk (RW1) model imposes smoothing among neighbour h_i's,

$$\pi(\boldsymbol{h} \mid \tau) \propto (\tau\tau^*)^{(n-1)/2} \exp\left(-\frac{\tau\tau^*}{2}\sum_{j=2}^{n}(h_j - h_{j-1})^2\right).$$

This is a first-order intrinsic Gaussian Markov random field with a covariance matrix on the form $\tau^{-1}R$, where the correlation matrix R is singular and of rank $n-1$. The parameter τ^* is a positive scaling constant which is added such that the generalised variance (the geometric mean of the diagonal elements of R^{-1}), is 1. This is needed to make the model invariant to the size of n and to unify the interpretation of τ, which then represents the precision of the (marginal) deviation from the null space of R, see Sørbye and Rue (2014) and Simpson et al. (2017) for further details. To separate the baseline hazard from the intercept, we impose the constraint $\sum_i h_i = 0$. The base model is a constant (in time) baseline hazard, which corresponds to infinite smoothing, $\tau = \infty$. The resulting penalised complexity prior for τ is given by (4).

An interesting extension to the commonly used subject specific frailty model is to allow the frailty term to depend on time (Yau and McGilchrist, 1998), leading to a time-dependent combined risk variable $\eta_i(t)$. Anticipating a positive correlation in time, it is natural to model this time dependent risk using a continuous-time Ornstein-Uhlenbeck process or its discrete time version given by AR(1). The stationary AR(1) model for subject i's specific frailty is given by

$$v_{it} \mid \{v_{is}, s < t\} \sim \mathcal{N}(\phi v_{i,t-1}, 1/(\tau_v(1-\phi^2))),$$

parameterised so that τ_v is the marginal precision and ϕ is the lag-one correlation. For this model component, the natural base model (keeping the marginal precision constant) is a time-constant frailty, in which we use the PC prior for ϕ in (6). For a fixed correlation ϕ, the base model for the precision τ_v is the constant zero which gives the type-2 Gumbel prior in (4).

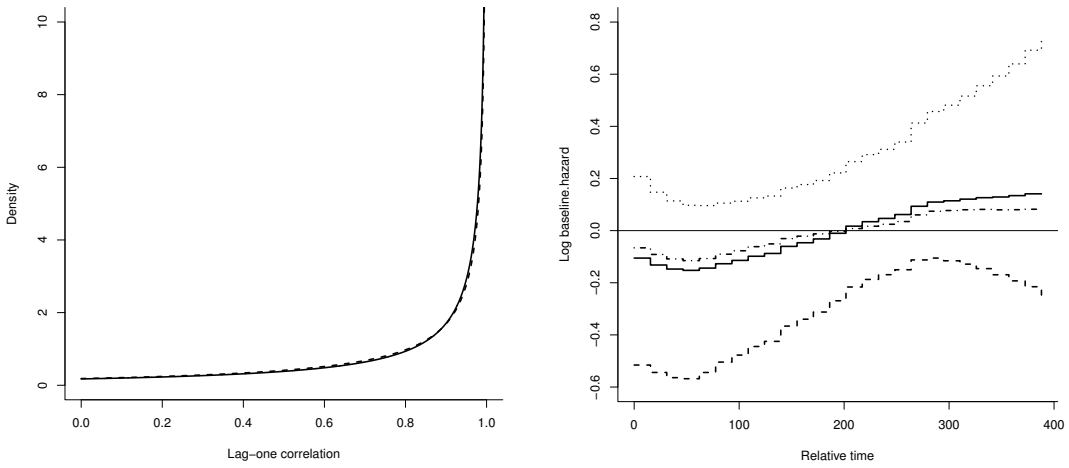
4.2 Analysis of chronic granulomatous disease data

We end this section by analysing data on chronic granulomatous disease (CGD) (Fleming and Harrington, 2005) available in `R` as the `cgd` dataset in the `survival` package. This data set consists of 128 patients from 13 hospitals with CGD. These patients participated in a double-blinded placebo controlled randomised trial, in which a treatment using gamma interferon (γ-IFN) was used to avoid or reduce the number of infections suffered by the patients. The recorded number of CGD infections for each patient ranged from zero to a maximum of seven, and the survival times are given as the times between recurrent infections on each patient. We follow Yau and McGilchrist (1998) and introduce a deterministic time dependent covariate for each patient, given as the time since the first infection (if any). Additionally, we include the covariates treatment (placebo or γ-IFN), inherit (pattern of inheritance), age (in years), height (in cm), weight (in kg), prophylac (use of prophylactic antibiotics at study entry), sex, region (US or Europe), and steroids (presence of corticosteroids) (Manda and Meyer, 2005; Yau and McGilchrist, 1998). The covariates age, height and weight were scaled before the analysis.

The computations were performed using the `R-INLA` package, by rewriting the model into a larger Poisson regression, see Fahrmeir and Tutz (2001) for a more general discussion and Martino et al. (2010) for `R-INLA` specific details. The prior specifications are as follows. We used

a constant prior for the intercept and independent zero mean Gaussian prior with low precision, i.e. 0.001, for all the fixed effects. For the log baseline hazard with $n = 25$ segments, we used the type-2 Gumbel prior with parameters $(U = 0.15/0.31, \alpha = 0.01)$ giving a marginal standard deviation for the log baseline hazard of about 0.15. This seems adequate as we do not expect the log baseline hazard to be highly variable. The time-dependent frailty was assigned a type-2 Gumbel prior for the precision with parameters $(U = 0.3/0.31, \alpha = 0.01)$ giving a marginal standard deviation of about 0.3, hence we allow for moderate subject specific variation. For the derived prior distribution (6) for ϕ, we used the parameters $(U = 1/2, \alpha = 0.75)$, which puts most of the prior density mass for high values of ϕ as $P(\phi > 1/2) = 0.75$. This corresponds to using a rate parameter $\theta \approx 1.55$ in (6).

Figure 2 (a) shows the prior (dashed) and posterior (solid) densities for the autocorrelation coefficient of the AR(1) model for the frailty. The data hardly alters the prior density at all, showing that there is not much information in the data available for this parameter, and we cannot conclude anything about the time-varying frailty. This is contrary to the findings in Manda and Meyer (2005) and Yau and McGilchrist (1998). Figure 2 (b) displays the log baseline hazard, showing an increasing trend (additional to the deterministic time dependent covariate), but the wide point-wise credible bands give no clear evidence for a time-dependent baseline hazard. With the new prior we are more confident that we do not overfit the data using the more flexible model for the log baseline hazard, as we do control the amount of deviation and its shrinkage towards it. The given conclusions are robust to changes in the parameter choices (U, α) for the different model components.



(a) (b)

Figure 2: Panel (a) displays the posterior density (solid) and prior density (dashed) for the lag-one autocorrelation ϕ in the AR(1) model for the time-dependent frailty. Panel (b) displays the log baseline hazard, mean (solid), median (dashed-dotted), lower $(0.025$, dashed) and upper $(0.975$, dotted) quantiles.

5 Deriving PC priors for higher-order AR processes

Define an autoregressive process of order p by

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \epsilon_t, \qquad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \kappa^{-1}), \tag{7}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ is an n-dimensional vector, $t = p, \ldots, n$, and κ is the precision of the innovations. The corresponding $p \times p$ correlation matrix $\boldsymbol{\Sigma}_p$ is Toeplitz (Gray, 2002) with elements that can be expressed as $\Sigma_{ij} = \sigma_{|i-j|}$, where $\sigma_0 = 1$. Although (7) is a natural parameterisation for known parameter values $\boldsymbol{\phi}_p = (\phi_1, \ldots, \phi_p)$, it is an awkward parameterisation when these are unknown, as the positive definiteness requirement of the correlation matrix makes the space of valid $\boldsymbol{\phi}_p$ complicated for $p > 3$. This implies that it is necessary to impose a number of non-linear constraints on these coefficients to define a stationary process.

A good alternative is to make use of the invariance property of the PC prior and define the prior distribution for $\boldsymbol{\phi}_p$ implicitly. The basic idea, which is commonly used when estimating AR(p) parameters, is to assign the prior to the partial autocorrelations $\boldsymbol{\psi}_p = (\psi_1, \ldots, \psi_p) \in [-1, 1]^q$, where $q = p - 1$. This gives a useful unconstrained set of parameters for this problem. Furthermore, there is a smooth bijective mapping between the partial autocorrelations and the autocorrelations in $\boldsymbol{\Sigma}_p$, given by the Levinson-Durbin recursions (Monahan, 1984; Golub and van Loan, 1996).

5.1 A sequential approach to construct PC priors

In deriving PC priors for the partial autocorrelations of an AR(p) process, we suggest to use a sequential approach, augmenting the partial autocorrelations one by one. Define $\psi_0 = 0$ and assume that $\boldsymbol{\psi}_p = (\boldsymbol{\psi}_{p-1}, \psi_p)$ for $p = 1, 2, \ldots$. We calculate the Kullback-Leibler divergence, conditional on the terms already included in the model,

$$\text{KLD}(f_1(\boldsymbol{\psi}_p) \parallel f_0(\boldsymbol{\psi}_{p-1})) = \frac{1}{2}\left(\text{tr}(\boldsymbol{\Sigma}_{p-1}^{-1}\boldsymbol{\Sigma}_p) - n - \ln\left(\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_{p-1}|}\right)\right),$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$ and f_1 and f_0 represent the densities of the AR(p) and AR($p-1$) processes, respectively. Notice that by augmenting the partial autocorrelations $\boldsymbol{\psi_{p-1}}$ with one (or several) terms, the correlation structure between the first $p-1$ elements of the corresponding AR(p) process remains unchanged. As the inverse correlation matrix of the AR($p-1$) process is a band matrix of order $2p-1$, we immediately notice that

$$\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_{p+r} = \boldsymbol{I}, \quad r = 1, 2 \ldots,$$

and $\text{tr}(\boldsymbol{\Sigma}_{p-1}^{-1}\boldsymbol{\Sigma}_p) = n$. Also,

$$\ln\left(\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_{p-1}|}\right) \;=\; \ln\left(\frac{\prod_{i=1}^{p}(1-\psi_i^2)^{n-i}}{\prod_{i=1}^{p-1}(1-\psi_i^2)^{n-i}}\right) = (n-p)\ln(1-\psi_p^2).$$

The resulting measure of distance from the AR(p) model to its base AR($p-1$), is only a function of the pth order partial autocorrelation, i.e.,

$$d(\psi_p) = \sqrt{2\text{KLD}(f_1(\boldsymbol{\psi}_p) \parallel f_0(\boldsymbol{\psi}_{p-1}))} = \sqrt{-(n-p)\ln(1-\psi_p^2)}.$$

Applying the principle of constant rate penalisation (2), an exponential density is assigned to $d(\psi_p)$ with rate $\lambda_p = \theta_p/\sqrt{n-p}$. The resulting prior density for the pth partial autocorrelation is

$$\pi(\psi_p) = \frac{\theta_p}{2}\exp\left(-\theta_p\sqrt{-\ln(1-\psi_p^2)}\right)\frac{|\psi_p|}{(1-\psi_p^2)\sqrt{-\ln(1-\psi_p^2)}}, \qquad |\psi_p| < 1, \qquad (8)$$

where the parameter $\theta_p > 0$ influences how fast the prior shrinks towards the base model.

The given formulation allows us to derive interpretable conditional priors for each of the partial autocorrelations ψ_p, given the previous parameters $\boldsymbol{\psi}_{p-1}$. In fact, $\pi(\psi_p \mid \boldsymbol{\psi}_{p-1}) = \pi(\psi_p)$ and the partial autocorrelations are seen to be consistent under marginalisation (as discussed in West (1991) in the context of kernel density estimation). Also, the marginal for an AR(q) process is not influenced by higher-order partial autocorrelations when these are 0, i.e. for $q \leq p$:

$$\pi(\boldsymbol{\psi}_q) \quad = \quad \int \pi(\boldsymbol{\psi}_p) d\boldsymbol{\psi}_{-q} = \pi(\boldsymbol{\psi}_q \mid \psi_{q+1} = 0, \ldots, \psi_p = 0).$$

5.2 Controlling shrinkage properties

The given sequential approach implies that the prior distribution for partial autocorrelations of different lags have the same functional form, but potentially different rate parameters. The next step is to determine a reasonable criterion to choose the rate θ_p in (8). Our suggestion is motivated by the conditional variance of the one-step ahead forecast error for an AR(p) with fixed p,

$$\text{Var}\left((x_{t+1} - \hat{x}_{t+1}) \mid \boldsymbol{x}_{s \leq t}, \tau\right) = \tau^{-1}(1 - \psi_1^2)(1 - \psi_2^2) \cdots (1 - \psi_p^2),$$

and the observation that often $1 - \psi_k^2$ is an non-decreasing function with k. We assume that

$$\text{E}(1 - \psi_k^2) = 1 - (1 - a)b^{k-1}, \quad a, b \in [0, 1], \quad k = 1, \ldots, p,$$

so the one-step ahead prediction, a priori, is non-decreasing with k. This reduces the prior specification into two parameters a and b, which have to be specified by the user. The parameter a represents the initial expectation $\text{E}(1 - \psi_1^2) = a$. The choice $b = 1$ induces the same shrinkage for all ψ_k while $b < 1$ gives increasing shrinkage for increasing k. For given values of a and b, the corresponding value for the rate parameter in (8) is found by solving

$$\text{E}(1 - \psi_k^2) = \frac{\theta_k \sqrt{\pi}}{2} \exp\left(\frac{\theta_k^2}{4} + \log\left(\text{erfc}\left(\frac{\theta_k}{2}\right)\right)\right) = 1 - (1 - a)b^{k-1} \tag{9}$$

for each $k = 1, \ldots, p$, where $\text{erfc}(z)$ denotes the complementary error function

$$\text{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2} dt.$$

6 Simulation results

To illustrate the properties of PC priors for the partial autocorrelations of autoregressive processes, we conduct a simulation study in which an AR(3) process is fitted to seven different test cases. We have chosen to present results when the length of each series is $n = 50$. In the first test example, we simply generate white noise series, while in the second and third cases we generate AR(1) processes where the first-lag autocorrelation is 0.7 and 0.9, respectively. Further, the test cases include three examples where the underlying processes are AR(2), while the last example is AR(3). The partial autocorrelations of the last four models were chosen to be equal to the test examples used in Liseo and Macaro (2013). All of the generated series are standardized to have variance one. We have chosen to fit autoregressive models of order 3 to all of the generated time series to investigate whether the the order of the underlying process is overestimated. The autoregressive models using both PC and reference priors are fitted in `R-INLA`, using the

latent model named `ar`. Further specifications for this model include the order of the process and hyperprior choices for both the precision parameter and the partial autocorrelations. The specific code to run the analysis with either the PC prior or the reference priors are available as an example case at `www.r-inla.org`.

Results including root mean squared error and coverage of credible intervals in estimating the partial autocorrelations of the different test cases are displayed in Table 1. The results are based on $m = 1000$ simulations where the average root-mean squared error is denoted by

$$\widehat{\text{rmse}}_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\psi}_i - \psi_i)^2}, \quad i = 1, 2, 3.$$

We also report frequentistic coverage, $\hat{\zeta}_i$, $i = 1, 2, 3$, of the estimated 95% highest posterior density intervals. In all test examples, the PC prior was implemented with scaling $a = b = 0.5$. By solving (9), this corresponds to using rate parameters $(\theta_1, \theta_2, \theta_3) \approx (0.87, 1.94, 3.33)$ in estimating the three partial autocorrelations. This imposes a higher rate of shrinkage to 0 as the order of the partial autocorrelations increase. In comparing the two prior distributions, we also considered the forecast error and coverage of 95% highest posterior density intervals for one-step ahead predictions. The results were very similar using the PC and reference priors, with coverage varying between 0.934 and 0.947 for all the seven cases, and these results are not shown explicitly.

As expected, the simulation results illustrate that we avoid overfitting using the PC prior. Specifically, in the first test case where the underlying process is white noise, the PC prior is seen to give both smaller root mean squared error and better frequentistic coverage compared with using the reference prior. We also notice that the use of the PC prior gives smaller error and higher coverage in estimating $\hat{\psi}_3$, for all the test cases. For the other parameters, the PC and reference priors are seen to have quite comparable performance. This implies that the PC priors seem like a promising alternative to reference priors in estimating the partial autocorrelations of AR(p) processes. The main advantage of PC priors is that these are easy to compute, also for higher-order processes. Also, the PC priors are more flexible as the rate parameters can be chosen differently for partial autocorrelations of different order.

The given approach to scale the PC prior is designed to reflect decreasing partial autocorrelations as the order of the process is increased. If we have reasons to believe that the partial autocorrelations do not decrease with higher order, we suggest to scale the prior densities for the partial autocorrelations of all lags similarly, using $b = 1$. We have chosen to report results only using $a = b = 0.5$ but we have also investigated results using several other combinations of the scaling parameters a and b. The main impression is that the PC priors are robust to different choices of a and b. Also, it is easy to understand how changes in these parameter will induce changes in the estimates. Increasing values of a and/or decreasing values of b give more shrinkage to 0. This will improve on the given results if the true partial autocorrelation is in fact 0 or close to 0. In contrast, more shrinkage will give higher error and lower coverage if the true partial autocorrelation is far away from 0. In general, we recommend that a is chosen to be less or equal to 0.5 as higher values of a might impose too must shrinkage for the first-lag partial autocorrelation. Also, values of b less than 0.5 might impose too much shrinkage for the partial autocorrelations of higher lags.

7 Discussion

An important aspect of statistical model fitting is to select models that are flexible enough to capture true underlying structure but do not overfit. Among competing models we would prefer the more parsimonious one, for example in terms of having fewer assumptions, fewer model components or a simpler structure of model components. Hawkins (2004) describes overfitting in terms of violating the principle of parsimony given by Occam's razor, the models and procedures used should contain all that is necessary for the modeling but nothing more. The given PC priors obey this principle, ensuring shrinkage to specific base models chosen to reflect prior information.

The PC priors represent a weakly informative alternative to existing prior choices for autoregressive processes, allowing for user-defined scaling to adjust the informativeness of the priors. The PC priors are computationally simple and are easily implemented for any finite order p of the AR process in software that allows the user to define their own priors. Specifically, the given PC priors are available within the `R-INLA` framework, in which AR processes can be used as building blocks within the general class of latent Gaussian models (Rue et al., 2009). This class of models have many applications, among others including analysis of temporal and spatial data. A natural extension in time series applications is to derive PC priors also for autoregressive (integrated) moving average processes. Other useful model extensions would include vector autoregressive models (Sims, 1980), frequently used to analyse multivariate time series, for example within the fields of econometrics.

In this paper, we have only considered stationary AR processes and testing of stationarity has not been addressed. In the AR(1) case, a posterior estimate of the first-lag autocorrelation close to 1 might indicate that the true model is non-stationary. If we suspect this a priori, it might be natural to use the random walk as the base model. Previous controversy (Phillips, 1991) in assigning a prior density to the first-lag autocorrelation of AR(1) processes relates to whether the stationarity condition $|\phi| < 1$ is included, or not. Phillips (1991) argued that objective ignorance priors, like the Jeffreys prior, should be used for AR(1) processes if no stationarity assumptions are made, while uniform priors would give inference biased towards stationarity. One of the problem seen with Jeffreys prior is that it puts most of its probability mass on regions of the parameter space giving a non-stationary process (Liseo and Macaro, 2013). The reference prior was originally only defined for stationary process but has been extended in a symmetric way for $|\phi| > 1$ (Berger and Yang, 1994), in which it is seen to have a more reasonable shape than Jeffreys prior (Robert, 2007). A relevant future project is to include testing of stationarity and study the use of PC priors also for non-stationary AR processes.

References

Albert, J. H. and Chib, S. (1993). Bayesian inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, 11:1–15.

Barndorff-Nielsen, O. and Schou, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, 3:408–419.

Barnett, G., Kohn, R., and Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *Journal of Econometrics*, 74:237–254.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37:905–938.

Berger, J. O. and Yang, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory*, 10:461–482.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer-Verlag, New Work, 2nd edition.

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC, 6th edition.

Chib, S. (1993). Bayes regression with autoregressive errors: A Gibbs sampling approach. *Journal of Econometrics*, 58:275–294.

DeJong, D. N. and Whiteman, C. H. (1991). Reconsidering 'Trends and random walks in macroeconomic time series'. *Journal of Monetary Economics*, 28:221–254.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Science + Business Media, New York, 2nd edition.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. Wiley Series in Probability and Statistics (Book 625). John Wiley & Sons, Inc., New Jersey, 2nd edition.

Golub, G. H. and van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, 3rd edition.

Gray, R. M. (2002). Toeplitz and circulant matrices: A review. Free book available from http://ee.stanford.edu/∼gray, Department of Electrical Engineering, Stanford University.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44:1–12.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Lesage, J. P. (1997). Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, 20:113–129.

Liseo, B. and Macaro, C. (2013). Objective priors for causal AR(p) with partial autocorrelations. *Journal of Statistical Computation and Simulation*, 83:1613–1628.

Manda, S. O. M. and Meyer, R. (2005). Bayesian inference for recurrent events data using time-dependent frailty. *Statistics in Medicine*, 24:1263–1274.

Martino, S., Akerkar, R., and Rue, H. (2010). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38:514–528.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, 67:68–83.

Monahan, J. F. (1984). A note on enforcing stationarity in autoregressive-moving average models. *Biometrika*, 71:403–404.

Phillips, P. C. B. (1991). To criticize the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics*, 6:333–364.

Prado, R. and West, M. (2010). *Time Series - Modeling, Computation and Inference*. Chapman & Hall/CRC, Boca Raton.

Robert, C. R. (2007). *The Bayesian choice*. Springer Science+Business Media, LLC, New York.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71:319–392.

Sahu, S. K. and Bakar, K. S. (2012). Hierarchical Bayesian autoregressive models for large space-time data with applications to ozone concentration modelling. *Applied Stochastic Models in Business and Industry*, 28:395–415.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007). High resolution space-time ozone modeling for assessing trends. *Journal of the American Statistical Association*, 102:1221–1234.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *To appear in Statistical Science*.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.

West, M. (1991). Kernel density estimation and marginalization consistency. *Biometrika*, 78:421–425.

Yau, K. K. W. and McGilchrist, C. A. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine*, 17:1201–1213.

Zellner, A. (1971). *Introduction to Bayesian inference in econometrics*. John Wiley & Sons, Inc., New York.

Test cases	Root mean squared error $\widehat{\text{rmse}}_1$	$\widehat{\text{rmse}}_2$	$\widehat{\text{rmse}}_3$	Coverage (95%) $\hat{\zeta}_1$	$\hat{\zeta}_2$	$\hat{\zeta}_3$
PC prior ($a = b = 0.5$)						
1. $\boldsymbol{\psi} = (0, 0, 0)$	0.133	0.123	0.111	0.928	0.939	0.956
2. $\boldsymbol{\psi} = (0.7, 0, 0)$	0.106	0.118	0.103	0.912	0.961	0.968
3. $\boldsymbol{\psi} = (0.9, 0, 0)$	0.078	0.123	0.107	0.889	0.944	0.964
4. $\boldsymbol{\psi} = (0.2, 0.3, 0)$	0.174	0.150	0.106	0.888	0.882	0.968
5. $\boldsymbol{\psi} = (-0.2, -0.6, 0)$	0.070	0.123	0.108	0.953	0.921	0.959
6. $\boldsymbol{\psi} = (0.5, -0.3, 0)$	0.093	0.136	0.107	0.938	0.918	0.965
7. $\boldsymbol{\psi} = (0.5, -0.3, -0.1)$	0.092	0.146	0.118	0.937	0.892	0.950
Reference prior						
1. $\boldsymbol{\psi} = (0, 0, 0)$	0.146	0.151	0.135	0.911	0.901	0.931
2. $\boldsymbol{\psi} = (0.7, 0, 0)$	0.101	0.143	0.126	0.911	0.932	0.944
3. $\boldsymbol{\psi} = (0.9, 0, 0)$	0.076	0.149	0.131	0.895	0.908	0.939
4. $\boldsymbol{\psi} = (0.2, 0.3, 0)$	0.185	0.143	0.130	0.879	0.920	0.929
5. $\boldsymbol{\psi} = (-0.2, -0.6, 0)$	0.070	0.111	0.133	0.949	0.933	0.934
6. $\boldsymbol{\psi} = (0.5, -0.3, 0)$	0.092	0.133	0.130	0.939	0.923	0.938
7. $\boldsymbol{\psi} = (0.5, -0.3, -0.1)$	0.088	0.143	0.133	0.938	0.916	0.928

Table 1: The root mean squared error and the frequentistic coverage of 95% highest posterior density intervals for each of the estimated partial autocorrelations of AR(3) processes, using PC priors with $a = b = 0.5$ and the reference prior, respectively. The given results are averaged over 1000 simulations, and the time series length in each simulation is $n = 50$.