

Gibbs sampling in AR models with random walk priors

Wolfgang Polasek, Song Jin

Institute for Statistics and Econometrics, University of Basel, CH-4051 Basel

Abstract: The paper analyses univariate autoregressive AR(p) models with tightness prior. The framework of the model is the conjugate normal linear model where the prior distribution is assumed to be a random walk process. The deviation from the prior distribution is measured by the tightness (hyper-) parameter λ . It is shown how the estimation of the starting values can be incorporated into the Gibbs sampling scheme. We demonstrate this approach with simulated and economic time series. It is found that for typical economic sample size the sampling fluctuations influence the posterior distribution considerably and informative prior distributions seem to be useful, especially for prediction.

1. Introduction

Bayesian VAR models have been used increasingly for macroeconomic forecasting in the last decade, because the forecasting properties have been found to outperform the corresponding classical models. The assumption of a tightness or smoothness structure for the lag coefficients implies a so-called hierarchy of prior parameters, also called hyperparameters. In the approach used by Litterman(1986) these hyperparameters for the tightness model are fixed by the modeler. Using the Gibbs sampler and a proper prior distribution for all parameters (Gelfand and Smith (1990)), we can find the posterior distribution of all parameters by simulation which utilises all so-called full conditional distributions (f.c.d.). In a conjugate Normal-Wishart framework the full conditional distributions can be derived in closed form which also allows an efficient generation of random numbers.

Smoothness models have been proposed by Shiller (1973), where the specification of the hyperparameters was left open. We will show that the Gibbs sampler can be used in a similar way as in the tightness models to find the posterior distribution of all parameters. The Gibbs sampler has been applied for time series model by e.g. Chib(1993), and Marriot et al. (1992). Gibbs sampling has the big advantage that one can impose many additional complication to a basic model, like outliers, heteroscedasticity, or errors in variables (Polasek (1994)).

In the next section we introduce the univariate tightness model for the AR(p) model (Polasek 1993). Then we demonstrate with simulated data the behavior of the starting values and the importance of the prior values for these parameters. As a real example, we estimated the Swiss consumption time series.

2. The tightness autoregression model: B-AR(p)

2.1 Introduction to the tightness autoregression model: B-AR(p)

Let \mathbf{y} be a vector of a univariate time series of length T , and we want to estimate an autoregressive process of lag p :

$$\begin{aligned}\mathbf{y} &= \alpha \mathbf{1}_n + \mathbf{y}_{-1}b_1 + \mathbf{y}_{-2}b_2 + \dots + \mathbf{y}_{-p}b_p + \epsilon \\ &= \mathbf{X}\mathbf{b} + \epsilon\end{aligned}\quad (1)$$

where $\mathbf{b} = (\alpha, b_1, \dots, b_p)$ and the $n \times (p+1)$ regressor matrix $\mathbf{X} = [\mathbf{1}_n : \mathcal{L}_p \mathbf{y}]$ consists of the constant and the past of the left hand variable \mathbf{y} . \mathcal{L}_p is the matrix lag operator $\mathcal{L}_p \mathbf{y} = [\mathbf{y}_{-1} : \dots : \mathbf{y}_{-p}]$.

The univariate autoregressive tightness model has the following hierarchical linear model structure

$$\mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I}_T), \quad (2)$$

$$(\mathbf{b}, \sigma^{-2}) \sim NoGa(\mathbf{b}_*, \lambda \mathbf{H}_*, \sigma_*^2, n_*), \quad (3)$$

$$\lambda^{-1} \sim Ga(\lambda_*, l_*), \quad (4)$$

$$\mathbf{y}_- \sim N[\mathbf{y}_*^*, \mathbf{\Psi}_*]. \quad (5)$$

λ is the tightness parameter and the $(p+1) \times (p+1)$ prior covariance matrix $\mathbf{H}_* = \text{diag}(d_0^*, \mathbf{D}_*)$ includes the unknown variance component for the intercept and the known tightness structure for the p past lags $\mathbf{D}_* = \text{diag}(1, 1/2, \dots, 1/p)$. The diagonal structure is adopted for simplicity for the precision matrix $\mathbf{H}_*^{-1} = \text{diag}(1/d_0^*, 1, 2, \dots, p)$. The $p \times 1$ vector \mathbf{y}_- contains the p starting values and we assume that prior information about the starting values is available by the normal distribution $N[\mathbf{y}_*^*, \mathbf{\Psi}_*]$. By introducing the starting values as parameters which are estimated, we can now use all $n=T$ observations of the observed time series instead the $n=T-p$ values, if we condition on the p first observed time series values.

The joint distribution for the data $\mathbf{Y} = (\mathbf{y}, \mathbf{X})$ and the parameters $\theta = (\mathbf{b}, \sigma^2, \lambda, \mathbf{y}_-)$ is given by the product

$$\begin{aligned}p(\theta, \mathbf{Y}) &\propto N[\mathbf{y} | \mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I}_n] N[\mathbf{b} | \mathbf{b}_*, \lambda \mathbf{D}_*] Ga[\lambda^{-1} | \lambda_*, l_*] \\ &\quad \cdot Ga[\sigma^{-2} | \sigma_*^2, n_*] N[\mathbf{y}_- | \mathbf{y}_*^*, \mathbf{\Psi}_*].\end{aligned}\quad (6)$$

Let θ^c be the symbol for the complementary parameters in a full conditional distribution. Then the full conditional distributions for θ^c can be derived in blocks as follows:

2.3.1 For the regression coefficients:

$$p(\mathbf{b} | \mathbf{Y}, \theta^c) \propto (\sigma^{-2})^{n/2} \exp\left\{-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{b})' (\mathbf{y} - \mathbf{X}\mathbf{b})\right\}$$

$$\begin{aligned} & \cdot \exp\left\{-\frac{1}{2}(\mathbf{b} - \mathbf{b}_*)' \lambda^{-1} \mathbf{H}_*^{-1} (\mathbf{b} - \mathbf{b}_*)\right\} \\ \propto & N[\mathbf{b} | \mathbf{b}_{**}, \mathbf{H}_{**}] \end{aligned} \quad (7)$$

a normal distribution with the moments

$$\mathbf{H}_{**}^{-1} = \mathbf{H}_*^{-1} \lambda^{-1} + \sigma^{-2} \mathbf{X}' \mathbf{X}, \quad (8)$$

$$\mathbf{b}_{**} = \mathbf{H}_{**} [\lambda^{-1} \mathbf{H}_*^{-1} \mathbf{b}_* + \sigma^{-2} \mathbf{X}' \mathbf{y}]. \quad (9)$$

2.3.2 The f.c.d. for the residual precision:

$$\begin{aligned} p(\sigma^{-2} | \mathbf{Y}, \theta^c) & \propto (\sigma^{-2})^{n/2} \exp\left\{-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{b})' (\mathbf{y} - \mathbf{X}\mathbf{b})\right\} \\ & \quad \cdot (\sigma^{-2})^{n_*/2-1} \exp\left\{-\frac{1}{2} n_* \sigma_*^2 / \sigma^2\right\} \\ & \propto Ga[\sigma^{-2} | \sigma_{**}^2, n_{**}] \end{aligned} \quad (10)$$

a gamma distribution with the parameters $n_{**}=n_*+n$, and

$$n_{**} \sigma_{**}^2 = n_* \sigma_*^2 + (\mathbf{y} - \mathbf{X}\mathbf{b})' (\mathbf{y} - \mathbf{X}\mathbf{b}) \quad (11)$$

2.3.3 The full conditional distribution for the tightness parameter:

$$\begin{aligned} p(\lambda | \mathbf{Y}, \theta^c) & \propto |\lambda \mathbf{H}_*|^{-1/2} \exp\left\{-\frac{1}{2} (\mathbf{b} - \mathbf{b}_*)' \lambda^{-1} \mathbf{H}_*^{-1} (\mathbf{b} - \mathbf{b}_*)\right\} \\ & \quad \cdot \lambda^{-l_*/2+1} \exp\left\{-\frac{1}{2} \lambda^{-1} l_* \lambda_*\right\} \\ & \propto Ga[\lambda^{-1} | \lambda_{**}, l_{**}] \end{aligned} \quad (12)$$

is a gamma distribution with the parameters $l_{**}=l_*+p+1$, and

$$l_{**} \lambda_{**} = l_* \lambda_* + (\mathbf{b} - \mathbf{b}_*)' \mathbf{H}_*^{-1} (\mathbf{b} - \mathbf{b}_*) \quad (13)$$

2.3.4 The distribution of the starting values \mathbf{y}_- :

The first elements of an AR(p) process suffer from the ‘starting value problem’. In Bayesian terms starting values can be simply viewed as unknown parameters which can be estimated from the data. The convenient feature of the Gibbs sampler is, that the full conditional distribution of the starting values can be also expressed in closed form as a normal distribution where we can draw samples. So the ‘starting value problem’ adds just another iteration step in the Gibbs sampler.

For an AR(p) process we have p starting values which we collect in the $p \times 1$ vector $\mathbf{y}'_-(y_{-1}, \dots, y_{-p})$. The first p elements of the AR(p) regression model are denoted as $\mathbf{y}_0=(y_1, \dots, y_p)$ and

$$\begin{aligned}
\mathbf{y}_0 &= \mathbf{X}_0 \mathbf{b} + \epsilon_0. \\
&= \alpha \mathbf{1}_p + \begin{pmatrix} y_{-1} & y_{-2} & y_{-3} & \dots & y_{-p} \\ y_0 & y_{-1} & y_{-2} & \dots & y_{-p+1} \\ \dots & \dots & \dots & \dots & \dots \\ y_{p-2} & y_{p-3} & \dots & y_0 & y_{-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_p \end{pmatrix} + \epsilon_0 \quad (14)
\end{aligned}$$

This can be written also as

$$\begin{aligned}
y_0 &= \alpha \mathbf{1}_p + \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ y_0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ y_{p-2} & y_{p-3} & \dots & y_0 & 0 \end{pmatrix} \mathbf{b} \\
&\quad + \mathbf{b} y_{-1} + \dots + \begin{pmatrix} b_p \\ 0 \\ \dots \\ 0 \end{pmatrix} y_{-p} + \epsilon_0 \\
&= \alpha \mathbf{1}_p + \mathbf{Y}^\circ \mathbf{b} + \mathbf{A} \mathbf{y}_- + \epsilon_0 \quad (15)
\end{aligned}$$

where $\mathbf{a} = \alpha \mathbf{1}_p + \mathbf{Y}^\circ \mathbf{b}$ does not depend on the starting values, and the \mathbf{Y}° and the matrix \mathbf{A} contain known data and regression coefficients, respectively:

$$\mathbf{Y}^\circ = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ y_0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ y_{p-2} & y_{p-3} & \dots & y_0 & 0 \end{pmatrix}, \quad (16)$$

$$\mathbf{A} = \begin{pmatrix} b_1 & b_2 & \dots & b_{p-1} & b_p \\ b_2 & b_3 & \dots & b_p & 0 \\ \dots & \dots & \dots & \dots & \dots \\ b_p & 0 & \dots & 0 & 0 \end{pmatrix} \quad (17)$$

Note that \mathbf{A} and \mathbf{Y}° are constructed in such a way that they have complementary roles for the decomposition of observed data (\mathbf{Y}°), regression coefficients (\mathbf{A}) and starting values (\mathbf{y}_-). Using the prior information about the starting values $\mathbf{y}_- \sim N[\mathbf{y}_-^*, \mathbf{\Psi}_*]$, we find the f.c.d. posterior distribution for \mathbf{y}_- to be

$$\begin{aligned}
p(\mathbf{y}_- | \mathbf{Y}, \theta^c) &\propto (\sigma^{-2})^{n/2} \exp\left\{-\frac{1}{2\sigma^2} (\mathbf{y}_0 - \mathbf{X}_0 \mathbf{b})' (\mathbf{y}_0 - \mathbf{X}_0 \mathbf{b})\right\} \\
&\quad \cdot \exp\left\{-\frac{1}{2} (\mathbf{y}_- - \mathbf{y}_-^*)' \mathbf{\Psi}_*^{-1} (\mathbf{y}_- - \mathbf{y}_-^*)\right\}. \quad (18)
\end{aligned}$$

Now the first quadratic form can be written by the above transformation as $(\mathbf{y}_0 - \mathbf{X}_0 \mathbf{b})' (\mathbf{y}_0 - \mathbf{X}_0 \mathbf{b})$ with $\mathbf{a} = \mathbf{Y}^\circ \mathbf{b} + \alpha \mathbf{1}_n$. Therefore the full conditional can be obtained as

$$p(\mathbf{y}_- | \mathbf{Y}, \theta^c) = N[\mathbf{y}_-^{**}, \mathbf{\Psi}_{**}] \quad (19)$$

a normal distribution with the parameters

$$\Psi_{**}^{-1} = \Psi_*^{-1} + \sigma^{-2} \mathbf{A}' \mathbf{A}, \quad (20)$$

$$\mathbf{y}_-^{**} = \Psi_{**}^{-1} [\Psi_*^{-1} \mathbf{y}_-^* + \sigma^{-2} \mathbf{A}' (\mathbf{y}_0 - \mathbf{a})]. \quad (21)$$

The posterior mean for the starting values is the matrix weighted average between the prior location \mathbf{y}_-^* and the first p time series values \mathbf{y}_0 which are adjusted by the vector \mathbf{a} containing the information ('back-casting') of the AR model. The starting values will now be used to construct the $n \times (p+1)$ \mathbf{X} matrix: $\mathbf{X} = [\mathbf{1}_T : \mathcal{L}_p \mathbf{y}^o]$ consists of the constant and the past of the left hand variable \mathbf{y} .

The Gibbs sampler can be invoked with the starting values which are based on simple OLS estimates: $\mathbf{b}^{(0)} = \mathbf{b}_{OLS} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$, $\sigma^{2(0)} = \sigma_{OLS}^2$, and $\lambda^{(0)} = (\mathbf{b}_{OLS} - \mathbf{b}_*)' \mathbf{H}_*^{-1} (\mathbf{b}_{OLS} - \mathbf{b}_*)$, where the \mathbf{X} matrix contains no starting values. As prior values for the starting values \mathbf{y}_- we suggest to take the first observation, i.e. $\mathbf{y}_-^* = y_1 \mathbf{1}_p$, and a tightness variance as well: $\text{Var}(\mathbf{y}_-) = \text{Var}(\mathbf{y}) \mathbf{D}_*$, where $\mathbf{D}_* = \text{diag}(1, 1/2, \dots, 1/p)$. Note that a small Ψ_* puts more weight on the prior mean \mathbf{y}_-^* and reduces sampling fluctuations for large p .

In case of centered time series, i.e. $\tilde{y}_t = y_t - \bar{y}$, $t=1, \dots, n$, which improves the convergence of the Gibbs sampler in time series models, we suggest to take as prior mean for the starting values $\mathbf{y}_-^* = (y_1 - \bar{y})/p \mathbf{1}_p$. We call this shrinkage prior for the starting values. This prior puts more weight on the prior starting values which should be close to zero. This approach has been found to produce posterior distributions for the starting values which are less variance inflated if the lag length becomes large.

2.2 Simulation results

We used simulated time series to test the inference procedure. The function is a simple random walk series, i.e. $y_t = y_{t-1} + \epsilon_t$, $t=1, \dots, n$, where $\epsilon_t \sim N(0, 0.01)$, with the starting values $y_0 = 1$. This random walk series with $n=200$ observations is called 'RW-200'. In Figure 1, besides the Gibbs sampling output of the tightness model we show in the histograms of the marginal posterior distribution, the analytically calculated posterior distribution of an usual normal-gamma model - as solid line overlay plot - where we take a conjugate normal distribution $(\mathbf{b}, \sigma^{-2}) \sim \text{NoGa}(\mathbf{b}_*, \lambda_* \mathbf{H}_*, \sigma_*^2, n_*)$, i.e. the interim step of a tightness prior distribution is omitted.

We can see that for the short time series, e.g. $n=200$, the prior information for λ and σ^2 is important for the convergence behavior. The Gibbs sampler produces an inflated posterior distribution if the prior information is not close to the true values, especially for short series. One should carefully choose the prior parameters. If the process converges to the 'wrong posterior distribution', the variances of the regression coefficients, the prediction and the starting values are larger than in the correct case. Thus we suggest a prior sensitivity analysis: We search over a range of prior values and take the one where the posterior variance is the smallest. If a time series is long

enough, e.g. 1000 observations, the prior information is less important. The Gibbs sampler in a long series converges easily to the right parameters.

2.4 An Example for Swiss Macro-Economic Time Series

A simple B-AR(p) model for quarterly consumption (in real prices) is analysed for Switzerland for the period 1966.4-1988.4. We start the analysis with a univariate AR(4) process. The prior parameters are set to $l_* = 1$, $\lambda_* = 0.04$, $n_* = 1$, and $\sigma_*^2 = 0.01$. The convergence of the Gibbs sampler could be achieved very fast. The posterior distribution for the parameters of the univariate AR(4) model for real Swiss consumption is shown in the Figure 2. The first and third columns of the figure contain the descriptive statistics of the marginal distribution of the parameters (including the OLS values if available) while the second and fourth ones show the histograms.

3. Conclusions

The estimation of hyperparameters in econometric models like in the Bayesian VAR model is an unsolved problem, because it requires heavy computations in non-standard form. The paper has demonstrated that the Gibbs sampler solves the estimation problem of the hyperparameters quite elegantly if the tightness prior information can be specified by conjugate distributions. Furthermore, it is possible to extend the approach to hierarchical tightness and smoothness models, which will be reported in a separate paper. A further advantage is that all small sample parameter distribution and predictive distributions can be simulated.

References:

- Chib S. (1993): Bayes regression with autoregressive errors: A Gibbs sampling approach. *Journal of Econometrics*, 58, 275-294.
- Gelfand A.E., Smith A.F.M. (1990): Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.
- Littermann R.B. (1986): A statistical approach to economic forecasting. *Journal of Business and Economic Statistics*, 4, 1-24.
- Marriott J., Ravishanker N., Gelfand A.E., Pai J. (1992): Bayesian analysis of ARMA processes: Complete sampling based inference under full likelihood. mimeo, University of Connecticut.
- Polasek W. (1993): Gibbs sampling in VAR models with tightness priors. mimeo, University of Basel.
- Polasek W. (1994): Gibbs sampling in B-VAR models with latent variables. WWZ-discussion papers Nr. 9415, University of Basel.
- Shiller R.J. (1973): A distributed lag estimator derived from smoothness priors. *Econometrica*, 41, 775-788

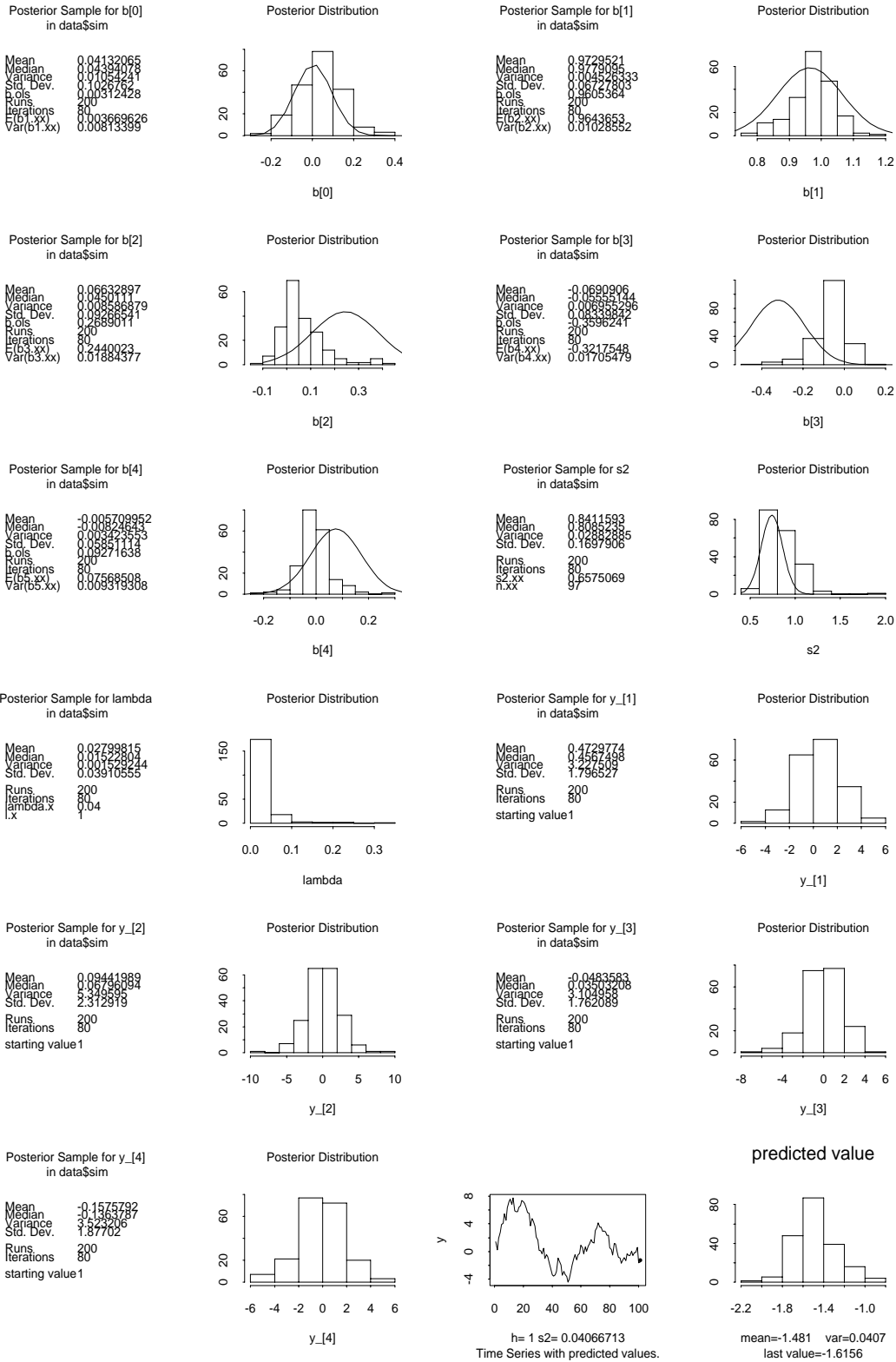


Figure 1: B-AR inference of model RW-200 with $l^* = 1.0$

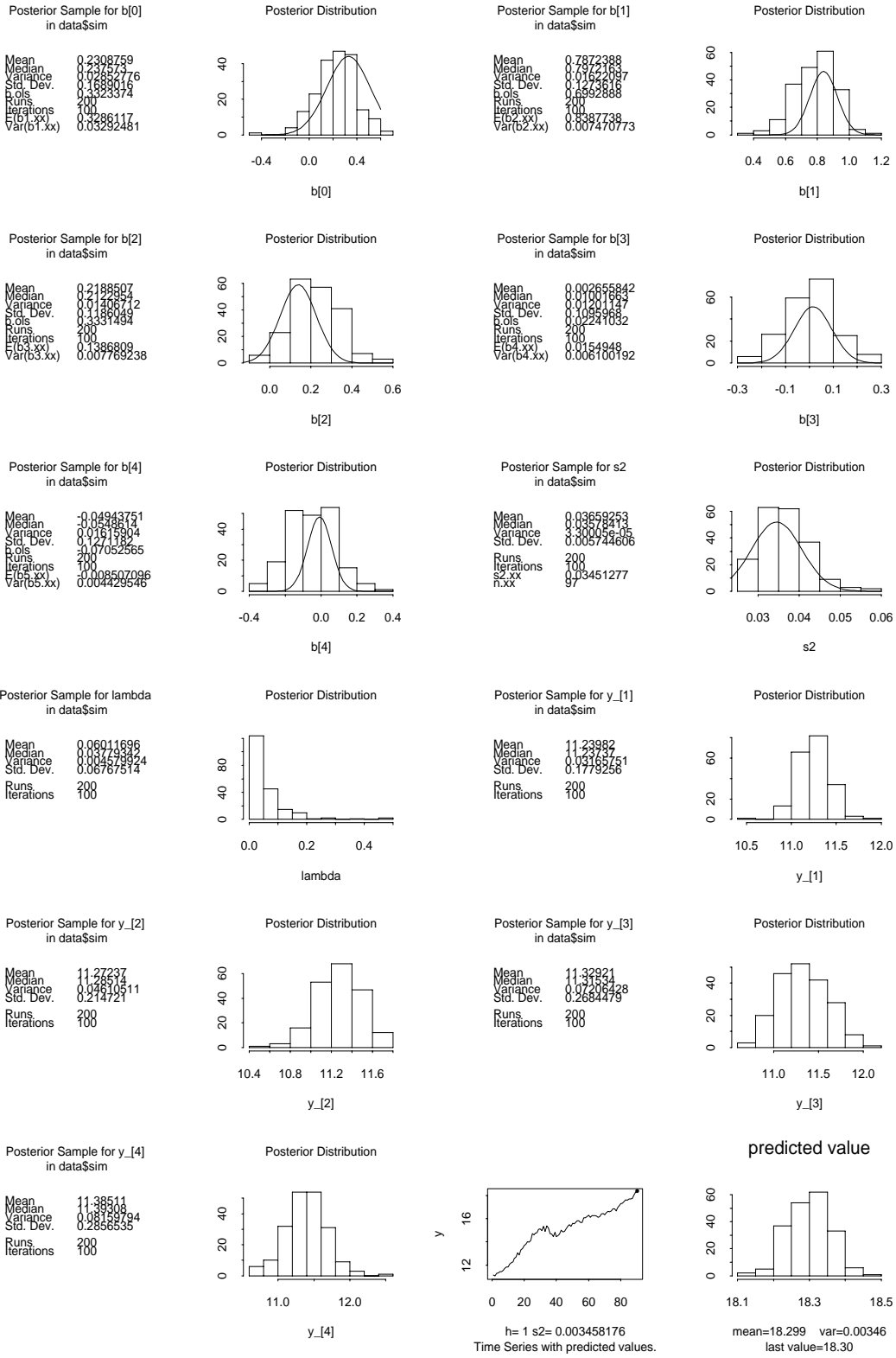


Figure 2: B-AR inference of real Swiss consumption