Article:
Triantafyllopoulos, K. and Harrison, P.J. (2008) Posterior mean and variance approximation for regression and time series problems. Statistics, 42 (4). pp. 329-350. ISSN: 0233-1888

https://doi.org/10.1080/02331880701864978

Posterior mean and variance approximation for regression and time series problems

K. Triantafyllopoulos* P.J. Harrison†

February 5, 2008

Abstract

This paper develops a methodology for approximating the posterior first two moments of the posterior distribution in Bayesian inference. Partially specified probability models, which are defined only by specifying means and variances, are constructed based upon second-order conditional independence, in order to facilitate posterior updating and prediction of required distributional quantities. Such models are formulated particularly for multivariate regression and time series analysis with unknown observational variance-covariance components. The similarities and differences of these models with the Bayes linear approach are established. Several subclasses of important models, including regression and time series models with errors following multivariate t, inverted multivariate t and Wishart distributions, are discussed in detail. Two numerical examples consisting of simulated data and of US investment and change in inventory data illustrate the proposed methodology.

Some key words: Bayesian inference, conditional independence, regression, time series, Bayes linear methods, state space models, dynamic linear models, Kalman filter, Bayesian forecasting.

1 Introduction

Regression and time series problems are important problems of statistical inference, which appear widely in many science fields, as for example in econometrics and in medicine. Regression has been discussed in many textbooks (Mardia *et al.*, 1979, Chapter 6; Srivastava and Sen, 1990); from a Bayesian standpoint Tiao and Zellner (1964), Box and Tiao (1973), Mouchart and Simar (1984), Pilz (1986), Leonard and Hsu (1999, Chapter 5) and O'Hagan and Forster (2004, Chapter 9) discuss a variety of parametric regression models, where the residuals follow normal or Student t distributions. Recent work on non-normal responses includes regression models in the type of generalized linear models (GLMs) (McCullagh and Nelder, 1989) and time series models in the type of dynamic GLMs (Fahrmeir and Kaufmann, 1987, 1991; Fahrmeir, 1992; West and Harrison, 1997, Chapter 12; Fahrmeir and Tutz, 2001, Chapter 8; Kedem and Fokianos, 2002; Godolphin and Triantafyllopoulos, 2006). Hartigan (1969) and Goldstein (1976) develop Bayesian inference for a general class of linear regression problems, in which the parameters or states of the regression equation are estimated

*Department of Probability and Statistics, Hicks Building, University of Sheffield, Sheffield S3 7RH, UK, email: k.triantafyllopoulos@sheffield.ac.uk

†University of Warwick, Coventry, UK

by minimizing the posterior expected risk. Goldstein (1979, 1983), Wilkinson and Goldstein (1996) and Wilkinson (1997) propose modifications to the Bayes linear estimators to allow for variance estimation in regression and time series problems. Such considerations are useful in practice because they allow inference to a range of problems that otherwise the modeller would need to resort to Monte Carlo estimation (Gamerman, 1997) or to other simulation based methods (Kitagawa and Gersch, 1996). West and Harrison (1997, Chapter 4) and Wilkinson (1997) discuss how the above mentioned regression estimation can be applied to a sequential estimation problem, which is necessary to consider in time series analysis.

In this paper we propose a modelling framework that allows approximate calculation of the first two moments of the posterior distribution in Bayesian inference. This is motivated by situations when a model may be partially specified in terms of its first two moments, or its probability distribution may be difficult to specify (or it may be specified with uncertainty). Partially specified prior posterior (PSPP) models are developed for dynamic situation in which a modeller is reluctant to specify a full probability model and yet requires a facility for approximate prior/posterior updating on mean and variance/covariance components of that model. The basic idea is that a linear function $\phi(X, Y)$ of two random vectors, X, Y, is second-order independent of the observed value of Y. Then in learning, no matter what value of Y is observed, the mean and the variance of $\phi(X, Y)$ takes exactly the same value. A further requirement is that the mean and variance of $X|Y = y$ can be deduced by the mean and variance of $\phi(X, Y)$. We show that for a class of regression models, linear Bayes methods are equivalent to PSPP, while we describe situations where PSPP can provide more effective estimation procedures than linear Bayes. We then describe two wide classes of regression and time series models, the scaled observational precision (SOP) and the generalized SOP, both of which are aimed at multivariate application. For the former model, we give the correspondence of PSPP (based on specification of prior means and variances only) with the normal/gamma model (based on specification of the prior distribution as normal/gamma). For the latter model, we show that PSPP can produce efficient estimation, overcoming problems of existing time series models. This relates to covariance estimation for multivariate state space models when the observation covariance matrix is unknown. For this interesting model we present two numerical illustrations, consisting of simulated bivariate data and of US investment and change in inventory data.

The paper is organized as follows. PSPP models are defined in Section 2. Sections 3 and 4 apply PSPP modelling to regression and time series problems. The numerical illustrations are given in Section 5. Section 6 gives concluding comments and the appendix details the proof of a theorem of Section 2.

2 Partially specified probability modelling

2.1 Full probability modelling

In Bayesian analysis, a full probability model for a random vector Z comprises the joint distribution of all its elements. The forecast distribution of any function of Z is then just that function's marginal distribution. Learning or updating simply derives the conditional distribution of Z given the received information on the appropriate function of Z. For example, let $Z = [X'\ Y']'$, where X, Y are real valued random vectors, and the probability density function of Z be denoted by $p(.)$. X will often be the vector comprising the parameters or states of the model and Y will be the vector comprising the observations of interest. The

model is precisely defined, if a density of Y given X is specified, e.g. $p(Y|X)$ so that $p(y|X)$ is the likelihood function of X based on the single observation $Y = y$. Then the one-step forecast distribution of Y is the marginal distribution of Y

$$p(Y) = \int_{\mathcal{S}} p(X, Y)\, dX, \tag{1}$$

where $\mathcal{S}$ is the space of X, also known as parametric space. When the value y of Y is observed, the revised density of X is

$$p(X|Y = y) = \frac{p(y|X)p(X)}{p(y)}, \tag{2}$$

from direct application of the Bayes theorem.

Most Bayesian parametric regression and time series models (including linear and non-linear) adopt the above model structure and their inference involves the evaluation of integral (1) and the Bayes rule (2).

However, in many situations, the evaluation of the above integral is not obtained in closed form and the application of rule (2) does not lead to a conjugate analysis, which is usually desirable in a sequential setting such as for time series application. For such situations, it is desirable to approximate only the mean and variance of $X|Y = y$. In this paper we consider the general problem of obtaining approximations of the first two moments of $X|Y = y$, when we only specify the first two moments of X and Y alone and not their joint distribution. We achieve this by replacing the full conditional independence structure, which is based on the joint distribution of X and Y, by second order independence, which is based on means and variances of X and Y. Our motivation is generated from the Gaussian case; suppose that X and Y have a joint normal distribution, then $X - A_{xy}Y$ and Y are mutually independent and the distribution of $X|Y = y$ can be derived from the distribution of $X - A_{xy}Y$, where A_{xy} is the regression matrix of X on Y (for a definition of A_{xy} see Section 2.2). So we can define a subclass of the Bayesian models of (1) and (2), where we can replace the strict mutual independence requirement by second order independence. Details appear in our definition of prior posterior probability models that follow.

2.2 Posterior mean and variance approximation

Let $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^p$, $W \in \mathbb{R}^q$ be any random vectors with a joint distribution ($m, p, q \in \mathbb{N} - \{0\}$). We use the notation $\mathbb{E}(X)$ for the mean vector of X, $\mathrm{Var}(X)$ for the covariance matrix of X and $\mathrm{Cov}(X, Y)$ for the covariance matrix of X and Y. We use the notation $X \perp_2 Y$ to indicate that X and Y are second order independent, i.e. $\mathbb{E}(X|Y = y) = \mathbb{E}(X)$ and $\mathrm{Var}(X|Y = y) = \mathrm{Var}(X)$, for any value y of Y. Furthermore, we use the notation $X \perp_2 W|Y$ to indicate that, given Y, X and W are second order independent, i.e. $\mathbb{E}(X|W = w, Y = y) = \mathbb{E}(X|Y = y)$ and $\mathrm{Var}(X|W = w, Y = y) = \mathrm{Var}(X|Y = y)$. Details on conditional independence can be found in Whittaker (1990) or Lauritzen (1996), who discuss independence in a much more sophisticated level necessary for the development of graphical models.

Considering vectors X and Y as above, it is well known that $X - A_{xy}Y$ and Y are uncorrelated, where $A_{xy} = \mathrm{Cov}(X, Y)\{\mathrm{Var}(Y)\}^{-1}$ is the regression matrix of X on Y. In order to obtain approximations of the posterior mean $\mathbb{E}(X|Y = y)$ and the posterior covariance matrix $\mathrm{Var}(X|Y = y)$ it is necessary to go one step further and assume that

$$X - A_{xy}Y \perp_2 Y, \tag{3}$$

which of course implies that $X - A_{xy}Y$ and Y are uncorrelated. With $\mu_x = \mathbb{E}(X)$ and $\mu_y = \mathbb{E}(Y)$, the prior means of X and Y, respectively, the above assumption is equivalent to the following two postulates.

1. Given Y, the posterior mean $\mathbb{E}(X - A_{xy}Y|Y = y)$ of $X - A_{xy}Y$ does not depend on the value of y of Y, so that the value of this mean must be the same for all values of Y, and so be equal to its prior expectation $\mu_x - A_{xy}\mu_y$.

2. Given Y, the posterior covariance matrix $\mathrm{Var}(X - A_{xy}Y|Y = y)$ of $X - A_{xy}Y$ does not depend on the value y of Y, so that this posterior covariance matrix takes the same value for all values y of Y and is necessarily equal to its prior covariance matrix $\mathrm{Var}(X - A_{xy}Y)$.

Thus it is possible to approximate $\mathbb{E}(X|Y = y)$ and $\mathrm{Var}(X|Y = y)$, since from the definition of second order independence (given above), we have

$$\begin{gathered}\mathbb{E}(X - A_{xy}Y|Y = y) = \mathbb{E}(X - A_{xy}Y) \Rightarrow \mathbb{E}(X|Y = y) - A_{xy}y = \mu_x - A_{xy}\mu_y \\ \Rightarrow \mathbb{E}(X|Y = y) = \mu_x - A_{xy}(y - \mu_y), \\ \mathrm{Var}(X|Y = y) = \mathrm{Var}(X - A_{xy}Y|Y = y) = \mathrm{Var}(X - A_{xy}Y) \\ = \Sigma_x + A_{xy}\Sigma_y A'_{xy} - 2\mathrm{Cov}(X, Y)A'_{xy} = \Sigma_x - A_{xy}\Sigma_y A'_{xy}\end{gathered}$$

and so we write

$$X|Y = y \sim \{\mu_x + A_{xy}(y - \mu_y), \Sigma_x - A_{xy}\Sigma_y A'_{xy}\},$$

where $\Sigma_x = \mathrm{Var}(X)$ and $\Sigma_y = \mathrm{Var}(Y)$.

Therefore we can define models that have a prior/posterior updating facility that is based on second order independence and that can approximate the posterior mean and variance obtained from an application of the Bayes theorem when the full distributions are specified. Thus we have the following definition.

Definition 1. *Let X and Y be any vectors of dimensions m and p respectively and assume that it exists the joint distribution of $Z = [X'\ Y']'$. Let A_{xy} be the regression matrix of X on Y. A first order partially specified prior posterior probability model for $(X; Y)$ (notation: PSPP(1)), is defined such that: (a) $X - A_{xy}Y \perp_2 Y$ and (b) for any value y of Y, the mean vector and the covariance matrix of $X|Y = y$ are obtainable from the mean vector and the covariance matrix of $X - A_{xy}Y$.*

We note that if X and Y have a joint normal distribution, then second order independence is guaranteed and in particular $X - A_{xy}Y$ and Y are mutually independent, which is much stronger than property (3). In this case $\mathbb{E}(X|Y = y)$ and $\mathrm{Var}(X|Y = y)$ are the exact posterior moments, produced by an application of Bayes rule (2). It follows that the approximation of the first two moments reflects on the approximation of postulate (3). Thus the approximations of $\mathbb{E}(X|Y = y)$ and $\mathrm{Var}(X|Y = y)$ will be so accurate as the condition (3) is satisfied. The question is: as we depart from normality, how justified are we to apply (3)? In order to answer this question and to support the adoption of (3), we give the next result, which states that Bayes linear estimation is equivalent to mean and variance estimation employing assumption (3).

Theorem 1. *Consider the vectors X and Y as above. Under quadratic loss, $\mu_x + A_{xy}(Y - \mu_y)$ is the Bayes linear estimator if and only if $X - A_{xy}Y \perp_2 Y$.*

The proof of this result is given in the appendix. Thus, if one is happy to accept the assumptions of Bayes linear optimality, she has to employ (3). Next we give three illustrative examples that show assumption (3) may be approximately satisfied.

Example A: checking postulate (3) for the multivariate Student t distribution

Let $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^p$ be random vectors with a joint Student t distribution with n degrees of freedom (Gupta and Nagar, 1999, §4.2). For example the marginal density of X is the Student t distribution $X \sim \mathcal{T}_m(n, \mu_x, C_{11})$ with density function

$$p(X) = \frac{\pi^{-p/2} n^{n/2} \Gamma\{(n+p)/2\}}{\Gamma(n/2)|C_{11}|^{1/2}} \left\{n + (X-\mu_x)' C_{11}^{-1}(X-\mu_x)\right\}^{(n+p)/2},$$

for $\mu_x = \mathbb{E}(X)$ and $\mathrm{Var}(X) = nC_{11}/(n-2)$, where $\Gamma(.)$ denotes the gamma function and $|\cdot|$ denotes determinant.

Write

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{T}_{m+p}\left\{n, \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix}\right\},$$

for some known parameters μ_x, μ_y, C_{11}, C_{12}, and C_{22}. The regression coefficient of X on Y is $A_{xy} = C_{12}C_{22}^{-1}$ so that

$$\begin{bmatrix} X - A_{xy}Y \\ Y \end{bmatrix} \sim \mathcal{T}_{m+p}\left\{n, \begin{bmatrix} \mu_x - A_{xy}\mu_y \\ \mu_y \end{bmatrix}, \begin{bmatrix} C_{11} - A_{xy}C_{22}A_{xy}' & 0 \\ 0 & C_{22} \end{bmatrix}\right\}.$$

Now for any value y of Y, the conditional distribution of $X - A_{xy}Y$ given $Y = y$ is

$$X - A_{xy}Y | Y = y \sim \mathcal{T}_m\left\{n+p, \mu_x - A_{xy}\mu_y, (C_{11} - A_{xy}C_{22}A_{xy}')\left[1 + n^{-1}(y-\mu_y)C_{22}^{-1}(y-\mu_y)'\right]\right\}.$$

Thus for any $n > 0$, $\mathbb{E}(X - A_{xy}Y|Y=y) = \mathbb{E}(X - A_{xy}Y)$, while for the variance, for $n > 2$, it is $\mathrm{Var}(X - A_{xy}Y|Y=y) \approx n(n-2)^{-1}(C_{11} - A_{xy}C_{22}A_{xy}') = \mathrm{Var}(X - A_{xy}Y)$. For large n postulate $X - A_{xy}Y \perp_2 Y$ is thought to be satisfactory.

Example B: checking postulate (3) for the inverted multivariate Student t distribution

The inverted Student t distribution is discussed in Dickey (1967), in Gupta and Nagar (1999, §4.4) and it is generated from a multivariate normal and a Wishart distribution as follows. Suppose that $X^* \sim \mathcal{N}_p(0, I_p)$ and $\Sigma \sim \mathcal{W}_p(n+p-1, I_p)$, for some $n > 0$, where $\mathcal{W}_p(n+p-1, I_p)$ denotes a Wishart distribution with $n+p-1$ degrees of freedom and parameter matrix I_p; this distribution belongs to the orthogonally invariant and residual independent family of distributions, discussed in Khatrie *et al.* (1991) and Gupta and Nagar (1999, §9.5). For a vector μ and a covariance matrix C we define $X = n^{1/2}C^{1/2}\{\Sigma + X^*(X^*)'\}^{-1/2}X^* + \mu$, where $C^{1/2}$ denotes the symmetric square root of C. Then the density of X is

$$p(X) = \frac{\Gamma\{(n+p)/2\}}{\pi^{p/2}\Gamma(n/2)|C|^{1/2} n^{(p+n-2)/2}} \left\{n - (X-\mu)'C^{-1}(X-\mu)\right\}^{n/2-1}.$$

This density defines the inverted multivariate Student t distribution and the notation used is $X \sim \mathcal{IT}_p(n, \mu, C)$.

Following a similar thinking as in Example A we have that

$$X - A_{xy}Y \sim \mathcal{IT}_m(n, \mu_x - A_{xy}\mu_y, C_{11} - A_{xy}C_{22}A'_{xy})$$

and conditioning on $Y = y$ (Gupta and Nagar, 1999, §4.4) we obtain

$$X - A_{xy}Y|Y = y \sim \mathcal{IT}_m\{n, \mu_x - A_{xy}\mu_y, (C_{11} - A_{xy}C_{22}A'_{xy})[1 - n^{-1}(y - \mu_y)'C_{22}^{-1}(y - \mu_y)]\}.$$

So we conclude that for large n the mean and variance of $X - A_{xy}Y|Y = y$ and $X - A_{xy}Y$ are approximately the same and thus $X - A_{xy}Y \perp_2 Y$.

Example C: checking postulate (3) for the Wishart distribution

Suppose that $\Sigma = (\Sigma_{i,j})_{i,j=1,2}$ follows a Wishart distribution $\Sigma \sim \mathcal{W}_2(n, S)$ with density

$$p(\Sigma) = \left\{2^n \Gamma_2(n/2)|S|^{n/2}\right\}^{-1} |\Sigma|^{(n-3)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(S^{-1}\Sigma)\right\},$$

where exp(.) denotes exponent, tr(.) denotes the trace of a square matrix, $S = (S_{ij})_{i,j=1,2}$, $n > 0$ are the degrees of freedom and $\Gamma_2(x) = \sqrt{\pi}\Gamma(x)\Gamma(x - 1/2)$ denotes the bivariate gamma function. Let $X = \Sigma_{12}$ and $Y = \Sigma_{22}$ and assume that we observe $Y = y$ so that $\mathbb{E}(Y) = nS_{22} \approx y$. From the expected values of the Wishart distribution (Gupta and Nagar, 1999, §3.3.6), we can write

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \left\{ n \begin{bmatrix} S_{12} \\ S_{22} \end{bmatrix}, n \begin{bmatrix} S_{11}S_{22} + S_{12}^2 & 2S_{12}S_{22} \\ 2S_{12}S_{22} & 2S_{22}^2 \end{bmatrix} \right\},$$

which, with $A_{xy} = S_{12}/S_{22}$, yields $\mathbb{E}(X - A_{xy}Y) = 0$ and $\mathrm{Var}(X - A_{xy}Y) = n(S_{11}S_{22} - S_{12}^2)$.

From Gupta and Nagar (1999, §3.3.4), the posterior distribution of $X|Y = y$ is $X|Y = y \sim \mathcal{N}\{S_{12}y/S_{22}, (S_{11} - S_{12}^2/S_{22})y\}$ leading to $\mathbb{E}(X - A_{xy}Y|Y = y) = 0 = E(X - A_{xy}Y)$ and $\mathrm{Var}(X - A_{xy}Y|Y = y) = \mathrm{Var}(X|Y = y) = (S_{11} - S_{12}^2/S_{22})y = (S_{11}S_{22} - S_{12}^2)y/S_{22} = \mathrm{Var}(X - A_{xy}Y)$. Thus we can establish that $X - A_{xy}Y \perp_2 Y$.

Examples A and B show that PSPP(1) modelling can be regraded as approximation to the true posterior mean and variance, corresponding to the full probability model assuming the distribution of these examples.

Returning to Definition 1, there are situations where the prior mean vectors and covariance matrices of X and Y are available, conditional on some other parameters, the typical example being when the moments of X and Y are given conditional on a covariance matrix V. Then, as V is usually unknown, the purpose of the study is to approximate the posterior mean vector and covariance matrix of $X|Y = y$ as well as to approximate the posterior mean vector and covariance matrix of V. In such situations postulate (3) reads $X - A_{xy}Y \perp_2 Y|V$ and another postulate for V is necessary in order to approximate the moments of $X|Y = y$, unconditionally of V. Regression problems of this kind are met frequently in practice, as V can represent an observation variance or volatility, which estimation is beneficial to accounting for the uncertainty of predictions. We can then extend Definition 1 to accommodate for the estimation of V.

Definition 2. *Let X, V and Y be any vectors of dimensions m, r and p respectively and assume that it exists the joint distribution of $Z = [X'\ V'\ Y']'$. Let A_{xy} be the regression matrix of X on Y, given V and let B_{vy} the regression matrix of V on Y. A second order partially specified prior posterior probability model for $(X, V; Y)$ (notation: PSPP(2)), is defined such that: (a) $X - A_{xy}Y \perp_2 Y|V$ and $V - B_{vy}Y \perp_2 Y$ and (b) for any value y of Y, the mean vector and the covariance matrix of $X|V, Y = y$ and $V|Y = y$ are obtainable from the mean vector and the covariance matrices of $X - A_{xy}Y$ and $V - B_{vy}Y$, respectively.*

An example of PSPP(2) model is the scaled observational precision model, which is examined in detail in Sections 3 and 4. Next we discuss the differences of PSPP(2) and Bayes linear estimation when V is a scalar variance.

Goldstein (1979, 1983), Wilkinson and Goldstein (1996) and Wilkinson (1997) examine some variants of this problem by considering variance modifications of the basic linear Bayes rule, considered in Hartigan (1969) and in Goldstein (1976). Below we give a basic description of the proposed estimators and we indicate the similarities and the differences of the proposed PSPP models and of the Bayes linear estimators. Consider a simple regression problem formulated as $Y|X, V \sim (X, V)$, $X \sim \{\mathbb{E}(X), \mathrm{Var}(X)\}$, where Y is a scalar response variable, X is a scalar regressor variable and $\mathbb{E}(X)$, $\mathrm{Var}(X)$ are the prior mean and variance of X. If V is known the posterior mean $\mathbb{E}(X|V, Y = y)$ can be approximated by the Bayes linear rule

$$\mu = \frac{\mathbb{E}(X)V + y\mathrm{Var}(X)}{V + \mathrm{Var}(X)} = \mathbb{E}(X) + A_{xy}\{y - \mathbb{E}(X)\}, \tag{4}$$

with related posterior expected risk

$$R(\mu) = \frac{\mathrm{Var}(X)V}{\mathrm{Var}(X) + V} = \mathrm{Var}(X)(1 - A_{xy}),$$

where $A_{xy} = \mathrm{Var}(X)/\{\mathrm{Var}(X) + V\}$ is the regression coefficient of X on Y, conditional on V. As it is well known $R(\mu)$ is the minimum posterior expected risk, over all linear estimators for $\mathbb{E}(X|Y = y)$, and in this sense μ attains Bayes linear optimality. If one assumes that the distributions of $Y|X, V$ and X are normal distributions, then μ gives the exact posterior mean $\mathbb{E}(X|V, Y = y)$ and $R(\mu)$ gives the exact posterior variance $\mathrm{Var}(X|V, Y = y)$. However, in practice in many problems, V is not known, and ideally the modeller wishes to estimate V and provide an approximation to the mean and variance of $X|Y = y$, unconditionally of V. Suppose that in addition to the above modelling assumptions, in order to estimate V, a prior mean $\mathbb{E}(V)$ and prior variance $\mathrm{Var}(V)$ of V are specified, namely $V \sim \{(\mathbb{E}(V), \mathrm{Var}(V)\}$. Goldstein (1979, 1983) suggest to estimate V with the Bayes linear rule

$$V^* = \frac{\mathbb{E}(V)\mathrm{Var}(Y^*) + y^*\mathrm{Var}(V)}{\mathrm{Var}(Y^*) + \mathrm{Var}(V)}, \tag{5}$$

where y^* is an observation from Y^*, a statistic that is unbiased for V, and $\mathrm{Var}(Y^*)$ is specified *a priori*. Then the Bayes rule μ is replaced by the rule μ^*, where V in μ is replaced by its estimate V^*. One can see that the revised regression matrix A^*_{xy} becomes

$$A^*_{xy} = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + V^*} = \frac{\mathrm{Var}(X)\mathrm{Var}(Y^*) + \mathrm{Var}(X)\mathrm{Var}(V)}{\mathrm{Var}(X)\mathrm{Var}(Y^*) + \mathrm{Var}(X)\mathrm{Var}(V) + \mathbb{E}(V)\mathrm{Var}(Y^*) + y^*\mathrm{Var}(V)}$$

and so the variance modified Bayes rule for $\mathbb{E}(X|Y = y)$ is $\mu^* = \mathbb{E}(X) + A^*_{xy}\{y - \mathbb{E}(X)\}$.

From Theorem 1, it is evident that the Bayes rule (4) is equivalent to $X - A_{xy}Y \perp_2 Y|V$. The Bayes rule (5) corresponds to the postulate $V - B_{vy}Y \perp_2 Y$, although the latter does not establish the equivalence of the PSPP models and Bayes linear estimation methods, since it can be verified that μ^* and V^* are not the same as in the PSPP modelling approach (see Section 3). In addition, the roles of Y^* and y^* are not fully understood; for example one question is how y and y^* are related and how one can determine y^* from y, especially when y is a vector of observations. The main problem experienced in the variance modified Bayes linear estimator μ^* is that the related expected risk $R(\mu^*)$ can not easily be determined and the work in this direction (Goldstein, 1979, 1983) has led to either intuitive evaluation for $R(\mu^*)$ or it has led to imposing even more restrictions to the model in order to obtain an analytic formula for $R(\mu^*)$. Although, both of these approaches can work in regression problems, they are not appropriate for time series problems, where sequential updating is required and thus an accurate evaluation of that risk is necessary. On the other hand the PSPP approach combines the two postulates, $X - A_{xy}Y \perp_2 Y|V$ and $V - B_{vy}Y \perp_2 Y$, using conditional expectations. It should be noted that the PSPP treatment is free of most of the assumptions made to the variance modified Bayes linear system so that approximate estimation of the posterior $\mathrm{Var}(X|Y)$ be given. The PSPP models are developed mainly for multivariate regression and time series problems and they are aimed to situations that either a fully Bayesian model is not available, or computationally intensive calculations, such as Monte Carlo methods, are undesirable, or a model can only be specified via means and variances.

3 The scaled observational precision model

3.1 Main theory

The scaled observational precision (SOP) model is a conjugate regression model, which illustrates the normal dynamic linear model with observational variances, see for example West and Harrison (1997, §4.5). This model is widely used in practice because it is capable to handle the practical problem of unknown observation variances. Here we construct a PSPP(2) model and we compare it with the usual conjugate SOP model.

Let V be a scalar variance, $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^p$ with

$$Z = \left[\begin{array}{c} X \\ Y \end{array} \right] \Bigg| V \sim \left\{ \left[\begin{array}{c} \mu_x \\ \mu_y \end{array} \right], V \left[\begin{array}{cc} \Sigma_x & A_{xy}\Sigma_y \\ A_{yx}\Sigma_x & \Sigma_y \end{array} \right] \right\},$$

for some known μ_x, μ_y, Σ_x and Σ_y.

Assuming $X - A_{xy}Y \perp_2 Y|V$, the partially specified posterior is

$$X|V, Y = y \sim \{\mu_x + A_{xy}(y - \mu_y), V(\Sigma_x - A_{xy}\Sigma_y A'_{xy})\}.$$

Let T be a, generally non-linear, function of Y, often taken as

$$T = (Y - \mu_y)'\Sigma_y^{-1}(Y - \mu_y).$$

Define K to be a α times the variance of $T|V$, for some $\alpha > 0$, and $A_{v\tau}$ to be the regression coefficient of V on T, conditional on K. We assume $V - A_{v\tau}T \perp_2 Y, K$ with forecast

$$T|V, K \sim (V, K/\alpha) \quad \text{and} \quad \mathrm{Cov}(T, V|K) = \mathrm{Var}(V|K),$$

where $V|K \sim (\widehat{V}, K/\eta)$, which is η/α times as precise as the conditional distribution of T, for some known $\widehat{V}, \alpha, \eta$, with

$$\left[\begin{array}{c} V \\ T \end{array} \right] \Bigg| K \sim \left\{ \left[\begin{array}{c} \widehat{V} \\ \widehat{V} \end{array} \right], \frac{K}{\eta} \left[\begin{array}{cc} 1 & 1 \\ 1 & (\eta+\alpha)/\alpha \end{array} \right] \right\}.$$

Given the observation $T = \tau$, and using $V - A_{v\tau}T \perp_2 Y, K$ with $A_{v\tau} = \alpha/(\eta+\alpha)$ we have

$$\begin{aligned} \mathbb{E}(V|K, T=\tau) &= \mathbb{E}(V|K) + \frac{\alpha}{\eta+\alpha}\left[\tau - \mathbb{E}(T|K)\right] = \frac{\eta\widehat{V} + \alpha\tau}{\eta+\alpha}, \\ \mathrm{Var}(V|K, T=\tau) &= \mathrm{Var}(V|T=\tau) - \mathrm{Cov}(V,T|K)\{\mathrm{Var}(T|K)\}^{-1}\mathrm{Cov}(T,V|K) \\ &= \frac{K}{\eta} - \frac{K^2}{\eta^2}\frac{\eta\alpha}{K(\eta+\alpha)} = \frac{K}{\eta}\left(1 - \frac{\alpha}{\eta+\alpha}\right) = \frac{K}{\eta+\alpha} \end{aligned}$$

so that

$$V|K, T=\tau \sim \left(\frac{\eta\widehat{V} + \alpha\tau}{\eta+\alpha}, \frac{K}{\eta+\alpha}\right). \tag{6}$$

Hence using conditional expectations, it follows that

$$X|Y=y \sim \left\{ \mu_x + A_{xy}(y-\mu_y), \frac{\eta\widehat{V} + \alpha\tau}{\eta+\alpha}(\Sigma_x - A_{xy}\Sigma_y A_{yx}) \right\}, \tag{7}$$

where $\tau = (y-\mu_y)'\Sigma_y^{-1}(y-\mu_y)$.

3.2 Comparison with the conjugate normal/gamma model

Now consider the relationship of the above model with standard normal conjugate models. A typical normal conjugate model with unknown scalar variance V, postulates the distribution of Z given V as

$$Z = \left[\begin{array}{c} X \\ Y \end{array} \right] \Bigg| V \sim \mathcal{N}_{mp}\left\{ \left[\begin{array}{c} \mu_x \\ \mu_y \end{array} \right], V \left[\begin{array}{cc} \Sigma_x & A_{xy}\Sigma_y \\ A_{yx}\Sigma_x & \Sigma_y \end{array} \right] \right\},$$

with the distribution of V as an inverse gamma so that $\nu s/V \sim \chi^2_\nu$. Here $\mathcal{N}_{mp}(.,.)$ denotes the mp-dimensional normal distribution and χ^2_ν denotes the chi-squared distribution with ν degrees of freedom. Writing $T = (Y-\mu_y)'\Sigma_y^{-1}(Y-\mu_y)$, the conditional distribution of T given V can be easily derived from the distribution of $TV^{-1}|V$ which is $TV^{-1}|V \sim \chi^2_p$. Then the posterior distribution of V^{-1} given $Y=y$ is

$$p\left(\frac{1}{V}\Big|T=\tau\right) = \frac{p(\tau|V)p(1/V)}{p(\tau)} \propto \left(\frac{1}{V}\right)^{(\nu+p)/2-1} \exp\left(-\frac{\nu s+\tau}{2V}\right),$$

from which it is deduced that, given $Y=y$, $(\nu s+\tau)V^{-1}|Y=y \sim \chi^2_{\nu+p}$. The posterior distribution of $X|Y=y$ is a multivariate Student t distribution based upon $\nu+p$ degrees of freedom with

$$X|Y=y \sim \mathcal{T}_m\left\{\nu+p, \mu_x + A_{xy}(y-\mu_y), \frac{\nu s+\tau}{\nu+p}(\Sigma_x - A_{xy}\Sigma_y A_{yx})\right\}, \tag{8}$$

$$\frac{\nu s+\tau}{V}\Big|Y=y \sim \chi^2_{\nu+p}, \quad \tau = (y-\mu_y)'\Sigma_y^{-1}(y-\mu_y). \tag{9}$$

Note that, if $\widehat{V} = \nu s/(\nu + p - 3), \eta = \nu + p - 3$, and $\alpha = 1$, then the posterior mean vector and covariance matrix of (7) and (8) are identical. However, this is not consistent with the conjugate model since from the prior assumption $\nu s/V \sim \chi^2_\nu$ it is

$$\mathbb{E}(V|s) = \frac{\nu s}{\nu - 2} \neq \widehat{V}, \quad (\nu > 2),$$

for any $p > 1$.

If we want to adopt the same prior for $\widehat{V} = \nu s/(\nu - 2)$ in both the PSPP and the conjugate models, then the respective posterior means for V will differ, i.e.

$$\mathbb{E}(V|Y = y, \text{PSPP model}) - \mathbb{E}(V|Y = y, \text{conjugate model}) = \frac{(p-1)\nu s}{(\nu - 2)(\nu + p - 2)},$$

where we have used $\eta = \nu + p - 3$ and $\alpha = 1$ as before. Note that if Y is a scalar response, e.g. $p = 1$, then the two variance estimates are identical. So the respective posterior variances of equations (7) and (8) will differ accordingly only when $p > 1$.

From the posterior distribution of $1/V$ we have that

$$\text{Var}(V|Y = y, \text{conjugate model}) = \frac{2(\tau + \nu s)^2}{(\nu + p - 2)^2(\nu + p - 4)} \tag{10}$$

while, from equation (6), the respective posterior variance for the PSPP model is

$$\text{Var}(V|K, Y = y, \text{PSPP model}) = \frac{K}{\nu + p - 2}, \tag{11}$$

where we have used $\alpha = 1$ and $\eta = \nu + p - 3$. If we choose $K = 2(\tau + \nu s)^2/\{(\nu + p - 2)(\nu + p - 4)\}$, then the two variances will be the same. Note that, irrespectively of the choice of K (given that K is bounded), as the degrees of freedom ν tend to infinity, the variances of both equations (10) and (11) converge to zero and so as $\nu \to \infty$, V concentrates about its mean asymptotically degenerating.

3.3 Application to time series modelling I

The above ideas can be applied to time series modelling when interest is placed on the estimation of the observation or measurement variance. Consider, for example, the p-dimensional time series vector Y_t, which at a particular time t sets

$$Y_t = B_t X_t + \epsilon_t, \quad \epsilon_t \sim (0, VZ), \quad X_t = C_t X_{t-1} + \omega_t, \quad \omega_t \sim (0, VW), \tag{12}$$

where B_t is a known $p \times m$ design matrix, C_t is a known $m \times m$ transition matrix and the innovation error sequences $\{\epsilon_t\}$ and $\{\omega_t\}$ are individually and mutually uncorrelated. The $p \times p$ and $m \times m$ covariance matrices Z and W are assumed known, while the scalar variance V is unknown. Initially we assume

$$X_0|V \sim (m_0, VP_0) \quad \text{and} \quad V \sim \left(\widehat{V}_0, \frac{K_0}{\eta_0}\right),$$

for some known m_0, P_0, $\widehat{V}_0$, K_0 and η_0. It is also assumed that *a priori*, X_0 is uncorrelated with $\{\epsilon_t\}$ and $\{\omega_t\}$. Denote with y^t the information set comprising the observations $y_1, y_2, \ldots, y_t$.

Then the PSPP model described above, applies at each time t with $\mu_x = C_t m_{t-1}$, $\mu_y = f_t = B_t C_t m_{t-1}$, $\Sigma_x = R_t = C_t P_{t-1} C_t' + W$ and $\Sigma_y = Q_t = B_t R_t B_t' + Z$, where m_{t-1} and P_{t-1} are calculated with the same way at time $t-1$, starting with $t = 1$. Given y^{t-1}, the regression matrix of X_t on Y_t is $A_{xy} = A_t = R_t B_t' Q_t^{-1}$, which is independent of V. It follows that $V|y^t \sim (\widehat{V}_t, K_t/\eta_t)$. With $\alpha = 1$, it is $K_t = K_{t-1}$ and $\eta_t = \eta_{t-1} + 1$ so that

$$\eta_t \widehat{V}_t = \eta_{t-1} \widehat{V}_{t-1} + e_t' Q_t^{-1} e_t,$$

where $e_t' Q_t^{-1} e_t = \tau_t$ and $e_t = y_t - f_t$ is the 1-step forecast error vector. The above estimate $\widehat{V}_t$ approximates the variance estimate of the conjugate dynamic linear model (West and Harrison, 1997, §4.5), which, assuming a prior $\eta_{t-1}\widehat{V}_{t-1} V^{-1}|y^{t-1} \sim \chi^2_{\eta_{t-1}}$, arrives at the posterior $(\eta_{t-1}\widehat{V}_{t-1} + \tau_t)V^{-1}|y^t \sim \chi^2_{\eta_{t-1}+p}$ so that $\mathbb{E}(V|y^t) = \eta_t \widehat{V}_t/(\eta_t + p - 3) \approx \widehat{V}_t$. The variance of $V|y^t$ in the conjugate model is

$$\mathrm{Var}(V|y^t) = \frac{2\eta_t^2 \widehat{V}_t^2}{(\eta_t - 2)^2(\eta_t - 4)},$$

whereas the respective variance in the PSPP model is $\mathrm{Var}(V|y^t) = K/\eta_t$, with $K = K_0$. Although these two variances differ considerably, in the sense that in the conjugate model the variance of $V|y^t$ is a function of the data y^t and in the PSPP model the variance of $V|y^t$ is only a function of time t and on the prior K_0, it can be seen that as $t \to \infty$, both variances converge to zero and so in both cases $V|y^t$ concentrates about its mean $\widehat{V}_t$ asymptotically degenerating.

In the PSPP model, the posterior mean vector and covariance matrix of $X_t|y^t$ are given by $X_t|y^t \sim (m_t, \widehat{V}_t P_t)$, where $m_t = C_t m_{t-1} + A_t e_t$ and $P_t = R_t - A_t Q_t A_t'$. These approximate the respective mean vector and covariance matrix produced by the conjugate model, which, under the inverted gamma prior, results to the posterior Student t distribution: $X_t|y^t \sim \mathcal{T}_m(\eta_t, m_t, \widehat{V}_t P_t)$.

4 The generalized observational precision model

4.1 Main theory

The generalization of the SOP model of Section 3 when V is a $p \times p$ variance-covariance matrix is not available and only special forms of conjugate SOP models are known (West and Harrison, 1997, Chapter 16). The problem is that since the dimensions of X and Y are different, it is not possible to scale the covariance matrix of $X|V$ by V, because X has dimension m and V is a $p \times p$ matrix. This problem is discussed in detail in Barbosa and Harrison (1992) and Triantafyllopoulos (2007). Next we propose a generalization of the SOP model, in which, given V, we avoid to scale the covariance matrices of X and Y by V. This setting is more natural than the setting of the SOP, which considers the somewhat mathematically convenient variance scaling.

Let V be a $p \times p$ covariance matrix, $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^p$ with

$$Z = \left[\begin{array}{c} X \\ Y \end{array} \right] \Bigg| V \sim \left\{ \left[\begin{array}{c} \mu_x \\ \mu_y \end{array} \right], \left[\begin{array}{cc} \Sigma_x & A_{xy}(\Sigma_y + V) \\ (\Sigma_y + V)A_{xy}' & \Sigma_y + V \end{array} \right] \right\},$$

for some known μ_x, μ_y, Σ_x and Σ_y, not depending on V. Note that now we cannot gain a scaled precision model. Even if we assume prior distributions for $Z|V$ and V, we can not

obtain the marginal distributions $X|Y=y$ and $V|Y=y$ in closed form, since the covariance matrices of X and Y are not scaled by V.

Assuming $X - A_{xy}Y \perp_2 Y|V$, conditional on V, the partially specified posterior is

$$X|V, Y=y \sim \{\mu_x + A_{xy}(y-\mu_y), \Sigma_x - A_{xy}(\Sigma_y + V)^{-1}A'_{xy}\}. \tag{13}$$

Define $T=(Y-\mu_y)(Y-\mu_y)'-\Sigma_y$ and denote with $\text{vech}(V)$ the column stacking operator of a lower portion of the symmetric positive definite matrix V. Given V, the forecast of T is

$$\text{vech}(T)|V,K \sim \left\{\text{vech}(V), \frac{K}{\alpha}\right\} \quad \text{and}$$
$$\text{Cov}\{\text{vech}(V), \text{vech}(T)\} = \frac{K}{\eta} = \text{Var}\{\text{vech}(V|K)\},$$

where α, η are known positive scalars and K is a known $\{p(p+1)/2\}\times\{p(p+1)/2\}$ covariance matrix. With $\widehat{V}$ the prior estimate of V and $I_{p(p+1)/2}$ the $\{p(p+1)/2\}\times\{p(p+1)/2\}$ identity matrix, we have

$$\left[\begin{array}{c}\text{vech}(V)\\ \text{vech}(T)\end{array}\right] \Bigg| K \sim \left\{\left[\begin{array}{c}\text{vech}(\widehat{V})\\ \text{vech}(\widehat{V})\end{array}\right], \frac{K}{\eta}\left[\begin{array}{cc} I_{p(p+1)/2} & I_{p(p+1)/2}\\ I_{p(p+1)/2} & (\eta+\alpha)\alpha^{-1}I_{p(p+1)/2}\end{array}\right]\right\}.$$

The regression matrix of $\text{vech}(V)$ on $\text{vech}(T)$ is $A_{v\tau}=\alpha(\eta+\alpha)^{-1}I_{p(p+1)/2}$. Assuming now that $\text{vech}(V) - A_{v\tau}\text{vech}(T) \perp_2 T|K$ we obtain the posterior mean and covariance of V as

$$\mathbb{E}\{\text{vech}(V)|K, T=\tau\} = \text{vech}(\widehat{V}) + \frac{\alpha}{\eta+\alpha}\left\{\text{vech}(\tau) - \text{vech}(\widehat{V})\right\}$$

and

$$\text{Var}\{\text{vech}(V)|K, T=\tau\} = \text{Var}\{\text{vech}(V)|K\} + A_{v\tau}\text{Var}\{\text{vech}(T)|K\}A'_{v\tau} = \frac{K}{\eta+\alpha}$$

so that

$$\text{vech}(V)|K, T=\tau \sim \left\{\frac{\text{vech}(\eta\widehat{V}+\alpha\tau)}{\eta+\alpha}, \frac{K}{\eta+\alpha}\right\}, \tag{14}$$

from which we see that the posterior mean of V can be written as

$$\mathbb{E}(V|K, T=\tau) = \widehat{V} + \frac{\alpha}{\eta+\alpha}\left(\tau - \widehat{V}\right) = \frac{\eta\widehat{V}+\alpha\tau}{\eta+\alpha}.$$

We note that in general the regression matrix A_{xy} in (13) will be a function of V^{-1} and this adds more complications to the calculation of the mean and covariance matrix of $X|Y=y$. However, if we impose the assumption that $\text{Cov}(X,Y|V) = A\text{Var}(Y)$, where A is a known $m\times p$ matrix not depending on V, then $A_{xy}=A$ is independent of V and so we get

$$X|Y=y \sim \left\{\mu_x + A_{xy}(y-\mu_y), \Sigma_x - \frac{1}{\eta+\alpha}A_{xy}\left(\Sigma_y + \eta\widehat{V} + \alpha\tau\right)A'_{xy}\right\}, \tag{15}$$

where $\tau=(y-\mu_y)(y-\mu_y)'-\Sigma_y$. Given that K is bounded, as $\eta\to\infty$, the covariance matrix of $\text{vech}(V)|K, T=\tau$ converges to the zero matrix and so $V|K, T=\tau$ concentrates about its mean $\mathbb{E}(V|K, T=\tau)$ asymptotically degenerating. This can be a theoretical validation of the proposed procedure for the accuracy of the estimator of V, $\mathbb{E}(V|K, T=\tau) = (\eta\widehat{V}+\alpha\tau)/(\eta+\alpha)$.

4.2 Application to linear regression modelling

A typical linear regression model sets

$$Y = BX + \epsilon, \quad \epsilon \sim (0, V), \quad X \sim (\mu_x, \Sigma_x), \tag{16}$$

where Y is a p-dimensional vector of response variables, B is a known $p \times m$ design matrix and ϵ is a p-dimensional error vector, which is uncorrelated with the random m-dimensional vector X. The mean vector μ_x and the covariance matrix Σ_x are assumed known and $\Sigma_y = B\Sigma_x B'$ so that $\mathrm{Var}(Y) = B\Sigma_x B' + V$. The covariance matrix of X and Y is $\mathrm{Cov}(X, Y) = \Sigma_x B'$ and so the assumption $\mathrm{Cov}(X, Y) = A\{\mathrm{Var}(Y)\}^{-1}$, does not hold, since $\mathrm{Var}(Y)$ is a function of V. Thus the posterior mean vector and covariance matrix of equation (15) do not apply, since now A_{xy} is stochastic in V. In order to resolve this difficulty next we propose an approximation that will allow computation of equation (13).

In order to proceed, we will need to evaluate $\mathbb{E}\{(\Sigma_y + V)^{-1}|Y = y\}$ and $\mathrm{Var}\{\mathrm{vech}\{(\Sigma_y + V)^{-1}\}|Y = y\}$. Since we only have equation (14) and we have no information on the distribution of V, we can not obtain the above mean vector and covariance matrix. Here we choose to adopt an intuitive approach suggesting that

$$\begin{aligned}
\widetilde{V} &= \mathbb{E}\{(\Sigma_y + V)^{-1}|K, T = \tau\} \approx \{\Sigma_y + E(V|K, T = \tau)\}^{-1} \\
&= (\eta + \alpha)\left\{(\eta + \alpha)\Sigma_y + \eta\widehat{V} + \alpha\tau\right\}^{-1}, \\
\widetilde{\widetilde{V}} &= \mathrm{Var}[\mathrm{vech}\{(\Sigma_y + V)^{-1}\}|K, T = \tau] \approx \mathrm{Var}\{\mathrm{vech}(\Sigma_y + V)|K, T = \tau\} = \frac{K}{\eta + \alpha}.
\end{aligned}$$

The reasoning of this is as follows. Since $\lim_{\eta\to\infty} \mathrm{Var}\{\mathrm{vech}(V)|K, T = \tau\} = 0$, V concentrates about its mean and so we can write $V \approx \mathbb{E}(V|K, T = \tau)$, for sufficiently large η. Then $(\Sigma_y + V)^{-1} \approx \{\Sigma_y + E(V|K, T = \tau)\}^{-1}$. The covariance matrix of $\mathrm{vech}\{(\Sigma_y + V)^{-1}\}$ has been set approximately the same with the covariance matrix of $\mathrm{vech}(\Sigma_y + V)$ ensuring that for large η, both covariance matrices converge to zero.

The above problem of the specification of $\widetilde{V}$ and $\widetilde{\widetilde{V}}$ can be generally presented as follows. Suppose that M is a bounded covariance matrix and assume that $\mathbb{E}(M)$ and $\mathrm{Var}\{\mathrm{vech}(M)\}$ are finite and known. The question is, given only this information, can one obtain $\mathbb{E}(M^{-1})$ and $\mathrm{Var}\{\mathrm{vech}(M^{-1})\}$? For example one can notice that if M follows a Wishart or inverted Wishart distributions, then $\widetilde{V}$ is approximately true. Formally, if $M \sim \mathcal{W}_p(n, S)$ (M follows the Wishart distribution with n degrees of freedom and parameter matrix S, see e.g. Gupta and Nagar, 1999, Chapter 3), we have $\mathbb{E}(M) = nS$ and $\mathbb{E}(M^{-1}) = S^{-1}/(n - p - 1) = n\{\mathbb{E}(M)\}^{-1}/(n - p - 1)$, which implies $\mathbb{E}(M^{-1}) \approx \{\mathbb{E}(M)\}^{-1}$, for large n. If $M \sim \mathcal{IW}_p(n, S)$ (M follows the inverted Wishart distribution with n degrees of freedom and parameter matrix S, see e.g. Gupta and Nagar, 1999, Chapter 3), we have $\mathbb{E}(M) = S/(n - 2p - 2)$ and so $\mathbb{E}(M^{-1}) = (n - p - 1)S^{-1} = (n - p - 1)\{\mathbb{E}(M)\}^{-1}/(n - 2p - 2)$, which again implies $\mathbb{E}(M^{-1}) \approx \{\mathbb{E}(M)\}^{-1}$, for large n. Of course M might not follow Wishart of inverted Wishart distributions and in many practical situations we will not have access to the distribution of M. For general application we can verify that $\mathbb{E}(M^{-1}) \approx \{\mathbb{E}(M)\}^{-1}$, if and only if M and M^{-1} are uncorrelated. The accuracy of the choice of $\widetilde{V}$ is reflected on the accuracy of the one-step predictions, which is illustrated in Section 5.1.

We can now apply conditional expectations to obtain the mean vector and the covariance matrix of $X|Y = y$. Indeed from the above and equation (13) we have

$$\mathbb{E}(X|Y = y) = \mu_x + \mathbb{E}(A_{xy}|Y = y)(y - \mu_y) = \mu_x + \Sigma_x B'\widetilde{V}(y - \mu_y).$$

For the covariance matrix $\mathrm{Var}(X|Y=y)$ we have

$$\begin{aligned}\mathbb{E}\{\mathrm{Var}(X|V,Y=y)|Y=y\} &= \Sigma_x - \Sigma_x B' E\{(\Sigma+V)^{-1}|Y=y\}B\Sigma_x \\ &= \Sigma_x - \Sigma_x B'\widetilde{V}B\Sigma_x\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\{\mathbb{E}(X|V,Y=y)|Y=y\} &= \mathrm{Var}[\mathrm{vec}\{\Sigma_x B'(\Sigma_y+V)^{-1}(y-\mu_y)\}|Y=y] \\ &= \{(y-\mu_y)'\otimes\Sigma_x B'\}G_p\widetilde{\widetilde{V}}G_p'\{(y-\mu_y)\otimes B\Sigma_x\}.\end{aligned}$$

where $\otimes$ denotes Kronecker product, $\mathrm{vec}(\cdot)$ denotes the column stacking operator of a lower portion of a matrix and G_p is the duplication matrix, namely $\mathrm{vec}\{(\Sigma_y+V)^{-1}\} = G_p\mathrm{vech}\{(\Sigma_y+V)^{-1}\}$.

Thus the mean vector and the covariance matrix of $X|Y=y$ are

$$\begin{aligned}X|Y=y \;\sim\; \Big\{&\mu_x+\Sigma_x B'\widetilde{V}(y-\mu_y), \Sigma_x-\Sigma_x B'\widetilde{V}B\Sigma_x \\ &+[(y-\mu_y)'\otimes\Sigma_x B']G_p\widetilde{\widetilde{V}}G_p'[(y-\mu_y)\otimes B\Sigma_x]\Big\}. \qquad (17)\end{aligned}$$

We note that the mean vector and covariance matrix of $X|Y=y$ depend on the estimates $\widetilde{V}$ and $\widetilde{\widetilde{V}}$. A simple intuitive approach was employed in this section and next we give an assessment of this approach by simulation. In general, equation (17) holds where $\widetilde{V}$ and $\widetilde{\widetilde{V}}$ are any estimates of the mean vector and covariance matrix of $(\Sigma_y+V)^{-1}|Y=y$.

4.3 Application to time series modelling II

In this section we consider the state space model (12), but the covariance matrices of the error drifts ϵ_t and ω_t are $\mathrm{Var}(\epsilon_t)=V$ and $\mathrm{Var}(\omega_t)=W$. Here V is an unknown $p\times p$ covariance matrix and W is a known $m\times m$ covariance matrix. The priors are partially specified by

$$X_0\sim(m_0,P_0)\quad\text{and}\quad \mathrm{vech}(V)\sim\left\{\mathrm{vech}(\widehat{V}_0),\frac{K_0}{\eta_0}\right\},$$

for some known m_0, P_0, $\widehat{V}_0$, K_0 and η_0. It is also assumed that *a priori*, X_0 is uncorrelated with $\{\epsilon_t\}$ and $\{\omega_t\}$. Note that in contrast with model (12), the above model is not scaled by V and in fact any factorization of the covariance matrices by V would lead to restrictive forms of the model; for a discussion of this topic see Harvey (1989), Barbosa and Harrison (1992), West and Harrison, (1997, §16.4), and Triantafyllopoulos (2006a, 2007). Before we give the proposed estimation algorithm, we give a brief description of the related matrix-variate dynamic models (MV-DLMs) and the restrictions imposed in these models.

Suppose $\{Y_t\}$ is a p-dimensional vector of observations, which are observed in roughly equal intervals of time $t=1,2,3,\ldots$. Write $Y_t=[Y_{1t}\;Y_{2t}\;\cdots\;Y_{pt}]'$, where each of Y_{it} is modelled as a univariate dynamic linear model (DLM):

$$Y_{it}=B_t'X_{it}+\epsilon_{it},\quad X_{it}=C_tX_{i,t-1}+\omega_{it},\quad \epsilon_{it}\sim\mathcal{N}(0,\sigma_{ii}),\quad \omega_{it}\sim\mathcal{N}_m(0,\sigma_{ii}W_i),$$

where B_t is an m-dimensional design vector, X_{it} is an m-dimensional state vector, C_t is an $m\times m$ transition matrix and the error drifts ϵ_{it} and ω_{it} are individually and mutually

uncorrelated and also they are uncorrelated with the state prior $X_{i,0}$, which is assumed to follow the normal distribution $X_{i,0} \sim \mathcal{N}_m(m_{i,0}, P_{i,0})$, for some known $m_{i,0}$ and $P_{i,0}$. The $m \times m$ covariance matrix W_i is assumed known and the variances $\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}$ form the diagonal elements of the covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1,2,\ldots,p}$, which is assumed unknown and it is subject to Bayesian estimation under the inverted Wishart prior $\Sigma \sim \mathcal{IW}_p(n_0 + 2p, n_0 S_0)$, for some known n_0 and S_0. The model can be written in compact form as

$$Y_t' = B_t' X_t + \epsilon_t', \quad X_t = C_t X_{t-1} + \omega_t, \quad \epsilon_t \sim \mathcal{N}_p(0, \Sigma), \quad \text{vec}(\omega_t) \sim \mathcal{N}_{mp}(0, \Sigma \otimes W), \qquad (18)$$

where $B_t' = [B_{1t}' \; B_{2t}' \; \cdots \; B_{pt}']$, $X_t = [X_{1t} \; X_{2t} \; \cdots \; X_{pt}]$, $C_t = \text{diag}(C_{1t}, C_{2t}, \ldots, C_{pt})$, $\text{vec}(X_0) \sim \mathcal{N}_{mp}\{\text{vec}(m_0), \Sigma \otimes P_0\}$, for $m_0 = [m_{1,0} \; m_{2,0} \; \cdots \; m_{p,0}]$ and $P_0 = \text{diag}(P_{1,0}, P_{2,0}, \ldots, P_{p,0})$. Model (18) is termed as matrix-variate dynamic linear model (MV-DLM) and it is studied in Quintana and West (1987, 1988), Smith (1992), West and Harrison (1997, Chapter 16) Triantafyllopoulos and Pikoulas (2002), Salvador *et al.* (2003, 2004), Salvador and Gargallo (2004), and Triantafyllopoulos (2006a, 2006b); Harvey (1986, 1989) develop a similar model where Σ is estimated by a quasi likelihood estimation procedure. The disadvantage of model (18) is that $Y_{1t}, Y_{2t}, \ldots, Y_{pt}$ are restricted to follow similar patterns since the model components B_t and C_t are common for all $i = 1, 2, \ldots, p$. One can notice that the only difference between Y_{it} and Y_{jt} $(i \neq j)$, is due to the error drifts ϵ_{it}, ω_{it} and ϵ_{jt}, ω_{jt}. Thus, for example, model (18) is not appropriate to model $Y_t = [Y_{1t} \; Y_{2t}]'$, where Y_{1t} is a trend time series and Y_{2t} is a seasonal time series. It follows that when there are structural changes between Y_{it} and Y_{jt}, the MV-DLM might be thought of as restrictive and inappropriate model and its use is not recommended. When p is large one can hardly justify the "similarity" of $Y_{1t}, Y_{2t}, \ldots, Y_{pt}$. We believe that in practice the popularity of the MV-DLM is driven from its mathematical properties (fully Bayesian conjugate estimation procedures for sequential forecasting and filtering/smoothing), rather than from a data driven analysis. Although we accept that in some cases the MV-DLM can be a useful model, we would submit that in many time series problems this model is unjustifiable and the above discussion expresses our reluctance in suggesting the MV-DLM for general use for multivariate time series problems.

Returning now to the PSPP dynamic model, denote with y^t the information set comprising data $y_1, y_2, \ldots, y_t$. If at time $t-1$ the posteriors are partially specified by $X_{t-1}|y^{t-1} \sim (m_{t-1}, P_{t-1})$ and $\text{vech}(V)|y^{t-1} \sim \{\text{vech}(\widehat{V}_{t-1}), \eta_{t-1}^{-1} K_{t-1}\}$, for some known m_{t-1}, P_{t-1}, $\widehat{V}_{t-1}$, K_{t-1} and η_{t-1}, then by direct application of the theory of Section 4 we have for time t: $\mu_x = C_t m_{t-1}$, $\Sigma_x = R_t = C_t P_{t-1} C_t' + W$, $\mu_y = f_t = B_t C_t m_{t-1}$, $\Sigma_y = B_t R_t B_t'$ and $A_{xy} = A_t = R_t B_t' (B_t R_t B_t' + V)^{-1}$. The 1-step ahead forecast covariance matrix is $Q_t = \text{Var}(Y_t|y^t) = B_t R_t B_t' + \widehat{V}_{t-1}$ and so we have $Y_t|y^{t-1} \sim (f_t, Q_t)$. Given $Y_t = y_t$, the error vector is $e_t = y_t - f_t$ and so the posterior mean of $V|y^t$ is

$$\eta_t \widehat{V}_t = \eta_{t-1} \widehat{V}_{t-1} + e_t e_t' - B_t R_t B_t',$$

where we have used $\alpha = 1$. Thus it is

$$\text{vech}(V)|y^t \sim \left\{ \text{vech}(\widehat{V}_t), \frac{K_t}{\eta_t} \right\},$$

where $\eta_t = \eta_{t-1} + 1$ and $K_t = K_{t-1}$. It follows that $K_t = K_0$ and therefore as $t \to \infty$, $V|y^t$ concentrates about $\widehat{V}_t$ asymptotically degenerating. By observing that $B_t R_t B_t' = Q_t - \widehat{V}_{t-1}$ and writing the updating of $\widehat{V}_t$ recurrently, we get

$$\widehat{V}_t = \widehat{V}_{t-1} + \frac{e_t e_t' - Q_t}{\eta_t} = \widehat{V}_0 + \sum_{i=1}^{t} \frac{e_i e_i' - Q_i}{\eta_0 + i}.$$

By forming now the standardized 1-step ahead forecast errors $e_t^* = Q_t^{-1/2} e_t$, where $Q_t^{-1/2}$ denotes the symmetric square root of Q_t^{-1}, one can obtain a measure of goodness of fit, since $e_t^* \sim (0, I_p)$. This can easily be implemented, by checking whether the mean of $e_1^*(e_1^*)', e_2^*(e_2^*)', \ldots, e_t^*(e_t^*)'$ is close to I_p or equivalently by checking that, for $e_t^* = [e_{1t}^* \; e_{2t}^* \; \cdots \; e_{pt}^*]'$, the mean of each $(e_{i,1}^*)^2, (e_{i,2}^*)^2, \ldots, (e_{it}^*)^2$ is close to 1 and e_{it}^* is uncorrelated with e_{jt}^*, for all t and $i \neq j$.

Applying the procedure adopted in linear regression, we have that the posterior mean vector and covariance matrix are given by $X_t | y^t \sim (m_t, P_t)$, with

$$m_t = C_t m_{t-1} + R_t B_t' \widetilde{V}_t e_t$$

and

$$P_t = R_t - R_t B_t' \widetilde{V}_t B_t R_t + (e_t' \otimes R_t B_t') G_p \widetilde{\widetilde{V}}_t G_p' (e_t \otimes B_t R_t),$$

where

$$\widetilde{V}_t = (B_t R_t B_t' + \widehat{V}_t)^{-1} \quad \text{and} \quad \widetilde{\widetilde{V}}_t = \frac{K_0}{\eta_t}.$$

From $\eta_t = \eta_{t-1} + 1$ it follows that as $\lim_{t\to\infty} \eta_t = \infty$ it is $\lim_{t\to\infty} \widetilde{\widetilde{V}}_t = 0$ and so for large t the posterior covariance matrix P_t can be approximated by $P_t \approx R_t - R_t B_t' \widetilde{V}_t B_t R_t$. This can motivate computational savings, since there is no need to perform calculations involving Kronecker products.

5 Numerical illustrations

In this section we give two numerical examples of the state space model considered in Section 4.3.

5.1 A simulation study

We simulate 1000 bivariate time series under 3 state space models and we compare the performance of the proposed model of Section 4.3 (referred here as DLM1), of the MV-DLM discussed in 4.3 (referred here as DLM2) and of the general multivariate dynamic linear model (referred here as DLM3). Let $Y_t = [Y_{1t} \; Y_{2t}]'$ be a bivariate time series. In the first state space model we simulate 1000 bivariate time series from the model

$$Y_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} X_t + \epsilon_t, \quad X_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} X_{t-1} + \omega_t, \quad \epsilon_t \sim \mathcal{N}_2(0, V), \quad \omega_t \sim \mathcal{N}_2(0, I_2), \quad (19)$$

where X_t is a bivariate state vector and the remaining components are as in Section 4.3. Initially we assume that $X_0 \sim \mathcal{N}_2(0, I_2)$ and the covariance matrix V is

$$V = (V_{ij})_{i,j=1,2} = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix},$$

which means that the variables Y_{1t} and Y_{2t} are highly correlated. The generated time series $\{Y_t\}$ comprise two local level components, namely $\{Y_{1t}\}$ and $\{Y_{2t}\}$. We note that DLM3 is the correct model, since it is used to generate the 1000 time series.

Table 1: Performance of the PSPP dynamic model (DLM1), MV-DLM (DLM2) and the general bivariate dynamic model (DLM3) over 1000 simulated time series of two local level components (LL), one local level and one linear trend component (LT) and one local level and one seasonal component (LS). Shown are the average (over all 1000 simulated series) values of the mean square standard error (MSSE), of the mean square error (MSE), of the mean absolute error (MAE) and of the mean error (ME).

type	model	MSSE		MSE		MAE		ME	
		y_{1t}	y_{2t}	y_{1t}	y_{2t}	y_{1t}	y_{2t}	y_{1t}	y_{2t}
LL	DLM1	0.905	1.045	2.536	7.975	1.521	2.249	-0.049	-0.022
	DLM2	1.009	1.075	2.556	8.635	1.259	2.348	0.012	-0.004
	DLM3	0.998	1.022	2.342	7.894	1.208	2.238	0.013	0.008
LT	DLM1	0.913	1.057	3.407	13.017	1.399	2.784	-0.157	-0.276
	DLM2	1.113	1.075	3.835	16.105	1.552	3.170	-0.003	-0.106
	DLM3	0.996	0.993	2.569	11.221	1.274	2.614	-0.093	-0.320
LS	DLM1	1.054	0.953	2.373	7.897	1.228	2.235	0.015	0.119
	DLM2	1.186	2.829	2.450	200.963	1.259	10.755	-0.006	0.057
	DLM3	0.982	0.994	2.361	7.856	1.224	2.218	0.017	0.112

In the second state space model we simulate 1000 time series from the model

$$Y_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} X_t + \epsilon_t, \quad X_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} X_{t-1} + \omega_t, \quad \epsilon_t \sim \mathcal{N}_2(0, V), \quad \omega_t \sim \mathcal{N}_2(0, I_2),$$

and the remaining components are as in (19). The generated time series from this model are time series comprising $\{Y_{1t}\}$ as a local level component and $\{Y_{2t}\}$ as a linear trend component.

Finally, in the third state space model, we simulate 1000 time series from the model

$$Y_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} X_t + \epsilon_t, \quad X_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\pi/6) & \sin(\pi/6) \\ 0 & -\sin(\pi/6) & \cos(\pi/6) \end{bmatrix} X_{t-1} + \omega_t, \tag{20}$$

where $\epsilon_t \sim \mathcal{N}_2(0, V)$, $\omega_t \sim \mathcal{N}_3(0, I_3)$ and here X_t is a trivariate state vector with initial distribution $X_0 \sim \mathcal{N}_3(0, I_3)$ and the remaining components of the model are as in (19). The generated time series from this model are bivariate time series comprising $\{Y_{1t}\}$ as a local level component and $\{Y_{2t}\}$ as a seasonal component with period $\pi/3$. Such seasonal time series appear frequently (Ameen and Harrison, 1984; Godolphin, 2001; Harvey, 2004).

Tables 1 and 2 show the results. In Table 1 the three state space models (DLM1, DLM2 and DLM3) are compared via the mean of squared standard 1-step forecast errors (MSSE), the mean square 1-step forecast error (MSE), the mean absolute 1-step forecast error (MAE) and the mean 1-step forecast error (ME). For a discussion of these measures of goodness of fit, known also as measures of forecast accuracy, the reader is referred to general time series textbooks, see e.g. Reinsel (1997) and Durbin and Koopman (2001). In a Bayesian flavour, goodness of fit may be measured via comparisons with MCMC methods (which provide the correct posterior destinies) or via Bayes monitoring systems, such as those using Bayes factors; see West and Harrison (1997).

Table 2: Performance of estimators of the covariance matrix $V = (V_{ij})_{i,j=1,2}$, produced by the PSPP dynamic model (DLM1) and the MV-DLM (DLM2). Shown are the average (over all 1000 simulated series; see Table 1) values of each estimator for times $t = 100$, $t = 200$ and $t = 500$.

type	$V = (V_{ij})_{ij=1,2}$	DLM1 $t = 100$	DLM2	DLM1 $t = 200$	DLM2	DLM1 $t = 500$	DLM2
LL	$V_{11} = 1$	1.347	0.961	1.072	0.954	0.988	0.974
	$V_{12} = 2$	2.352	1.047	1.792	0.914	2.087	1.113
	$V_{22} = 5$	5.846	3.407	4.332	2.874	5.215	3.290
LT	$V_{11} = 1$	2.087	0.475	1.599	0.647	1.210	0.678
	$V_{12} = 2$	3.169	0.463	2.375	0.721	2.217	0.802
	$V_{22} = 5$	6.200	2.509	4.627	2.718	5.043	2.851
LS	$V_{11} = 1$	0.627	0.729	0.782	0.851	0.960	0.955
	$V_{12} = 2$	1.497	0.887	1.674	0.901	1.872	0.907
	$V_{22} = 5$	4.084	3.548	4.104	11.439	4.626	76.609

Section 4.3 details how the MSSE has been calculated. Out of the three models we know that DLM3 is the correct model, since it is used to generate the time series data. For the local level components (LL), both DLM1 and DLM2 put good performances with the DLM2 having the edge and being closer to the performance of the DLM3. This is expected, since as we noted in Section 4.3 when both time series components Y_{1t} and Y_{2t} are similar the MV-DLM (DLM2) has good performance. However, in the LT and LS time series components, where the two series Y_{1t} and Y_{2t} in each case, are not similar, we expect that the DLM2 will not perform very well. This is indeed confirmed by our simulations, for which Table 1 clearly shows that the performance of DLM1 is better than that of the DLM2. For example, for the LS component, the MSSE of the DLM1 is $[1.054\ 0.953]'$, which is close to $[1\ 1]'$, while the respective MSSE of the DLM2 is $[1.186\ 2.829]'$.

Table 2 looks at the accuracy of the estimation of the covariance matrix V, for each model. For the LL components $V_{11} = 1$ is estimated better from DLM2, although for $t = 500$ the estimate from DLM1 is slightly better. For $V_{12} = 2$ and $V_{22} = 5$, DLM2 produces poor results as compared to the DLM1. For example, even for $t = 500$ the estimate of $V_{22} = 5$ of the DLM2 is only 3.290, while the estimate of the DLM1 is 5.215. This phenomenon appears to be magnified when looking at the LT and LS components, where for example even at $t = 500$ for the LT the estimate of $V_{12} = 2$ and for the LS the estimate of $V_{22} = 5$ are 0.802 and 76.609, while the respective estimates from the DLM1 are 2.217 and 4.626. The conclusion is that the DLM1 produces a consistent estimation behaviour over a wide range of bivariate time series, while the DLM2 (matrix-variate DLM) produces acceptable performance when the component time series are all similar.

It should be stated here that, the matrix-variate state space models of Harvey (1986) produce a similar performance with the DLM2; Harvey (1989) calls the above matrix-variate models as 'seemingly unrelated time series models' to indicate the similarity of the component time series. The models of Triantafyllopoulos and Pikoulas (2002) and Triantafyllopoulos (2006a, 2006b) and of many other authors (see the citations in Harvey, 1989; West and Harrison, 1997; Durbin and Koopman, 2001) can only accommodate for regression type state

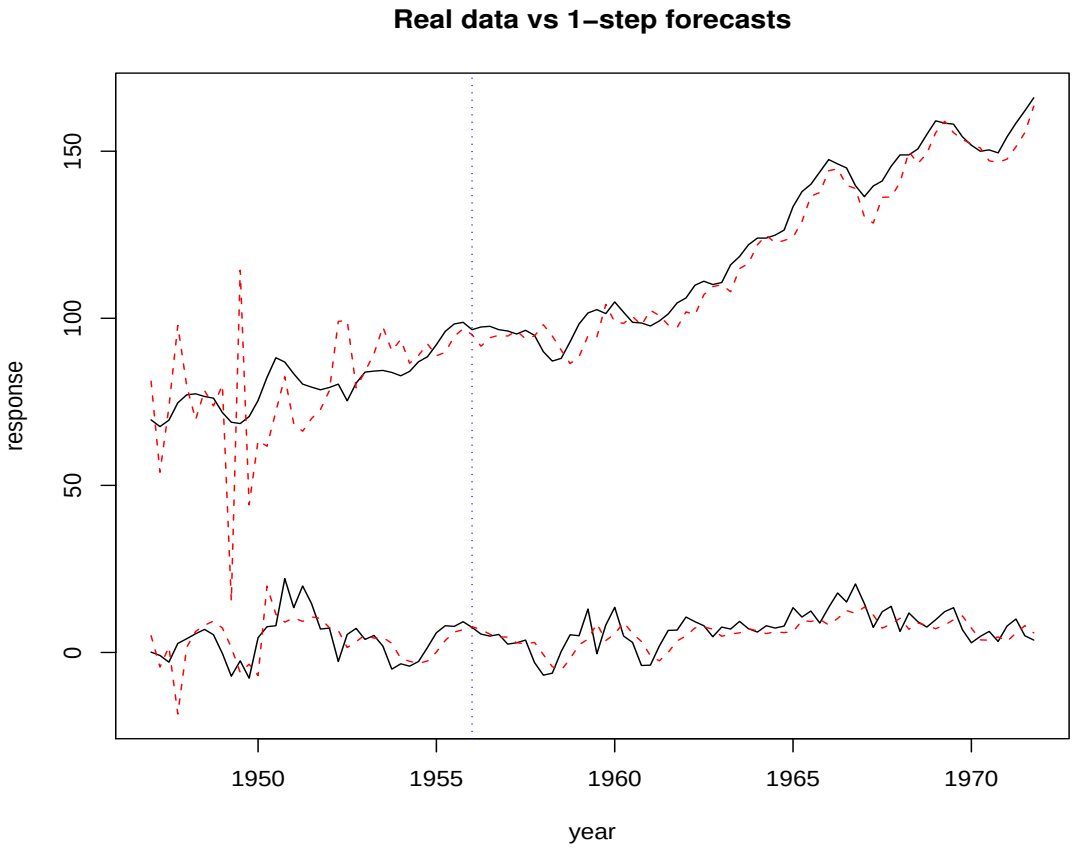

Figure 1: US Investment and Change in Inventory time series $y_t = [y_{1t}\ y_{2t}]'$ with its 1-step forecast mean $f_t = [f_{1t}\ f_{2t}]'$. The top solid line shows y_{1t} and the bottom solid line shows y_{2t}; the top dashed line shows f_{1t} and the bottom dashed line shows f_{2t}.

space models and for local level models. More general structures, such that of model (20) can only be dealt with via simulation-based methods, such as Monte Carlo simulation. For high-dimensional dynamical systems and in particular for observation covariance estimation, the proposal of PSPP state space model of Section 4.3 offers a fast and reliable approximate estimation procedure, which can be applied for a wide range of time series.

5.2 The US investment and business inventory data

We consider US investment and change in business inventory data, which are deseasonalised and they are measured quarterly into a bivariate time series (variable y_{1t}: US investment data and variable y_{2t}: US change in inventory data) over the period 1947-1971. The data are fully described and tabulated in Lütkepohl (1993) and Reinsel (1997, Appendix A). The data are plotted in Figure 1 with their forecasts, which are generated by fitting the linear trend PSPP state space model

$$Y_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} X_t + \epsilon_t, \quad X_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} X_{t-1} + \omega_t, \quad \epsilon_t \sim (0, V), \quad \omega_t \sim (0, W_t), \qquad (21)$$

where here we have not specified the distributions of ϵ_t and ω_t as normal and we have replaced the time-invariant W of Section 4.3 with a time-dependent W_t. Model (21) is a PSPP linear trend state space model, for which we choose the priors $m_0 = [80.622 \ 4.047]'$ (mean of $[Y_{1t} \ Y_{2t}]'$ for $t = 1941 - 1956$, indicated in Figure 1 by the vertical line), $P_0 = 1000 I_2$ (weakly informative prior covariance matrix or low precision $P_0^{-1} \approx 0$) and

$$V_0 = \begin{bmatrix} 66.403 & 22.239 \\ 22.239 & 46.547 \end{bmatrix},$$

which is taken as the sample covariance matrix of Y_{1t} and Y_{2t}, for the time period 1941-1955. The covariance matrix W_t measures the durability and the stability of the change or evolution of the states X_t. Here we specify W_t with 2 discount factors, δ_1 and δ_2, as follows. With G as the evolution matrix of X_t and Δ the discount matrix

$$G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \Delta = \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix},$$

we have

$$W_t = \Delta^{-1/2} G P_{t-1} G' \Delta^{-1/2} - G P_{t-1} G',$$

where R_t in the recursions of Section 4.3 is replaced by $R_t = G P_{t-1} G' + W_t$. Although this discounting specification is not advocated by West and Harrison (1997, §6.4), it has been successfully used (McKenzie, 1974, 1976; Abraham and Ledolter, 1983, Chapter 7; Ameen and Harrison, 1985; Goodwin, 1997).

The values of δ_1 and δ_2 are chosen by experimentation. The above model gave the best result with a combination of discount factors $\delta_1 = 0.2$ and $\delta_2 = 0.4$. The performance measures were MSSE $= [1.001 \ 1.101]'$, MSE $= [111.165 \ 66.941]'$, MAE $= [6.718 \ 6.855]'$ and ME $= [0.076 \ 1.725]'$. Other combinations of δ_1 and δ_2 yield less accurate results, with the usual effect that one of the two series y_{1t} and y_{2t} is accurately predicted, but the other one series is badly predicted. This problem certainly arises when $\delta_1 = \delta_2$, which clearly indicates the need of multiple discounting. Also, Figure 2 plots the observation variance, covariance and correlation estimates in the time period 1956-1970. From this plot we observe that the variability of the change in inventory time series component y_{2t} is much larger than that of y_{1t}. The estimate of the observation correlation indicates the high cross-correlation between the two series.

6 Discussion

This paper develops a method for approximating the first two moments of the posterior distribution in Bayesian inference. This work is particularly appealing in regression and time series problems when the response and parameter distributions are only partially specified by means and variances. Our partially specified prior posterior (PSPP) models offer an approximation to prior/posterior updating, which is appropriate for sequential application, such as in time series analysis. The similarities and differences with Bayes linear methods

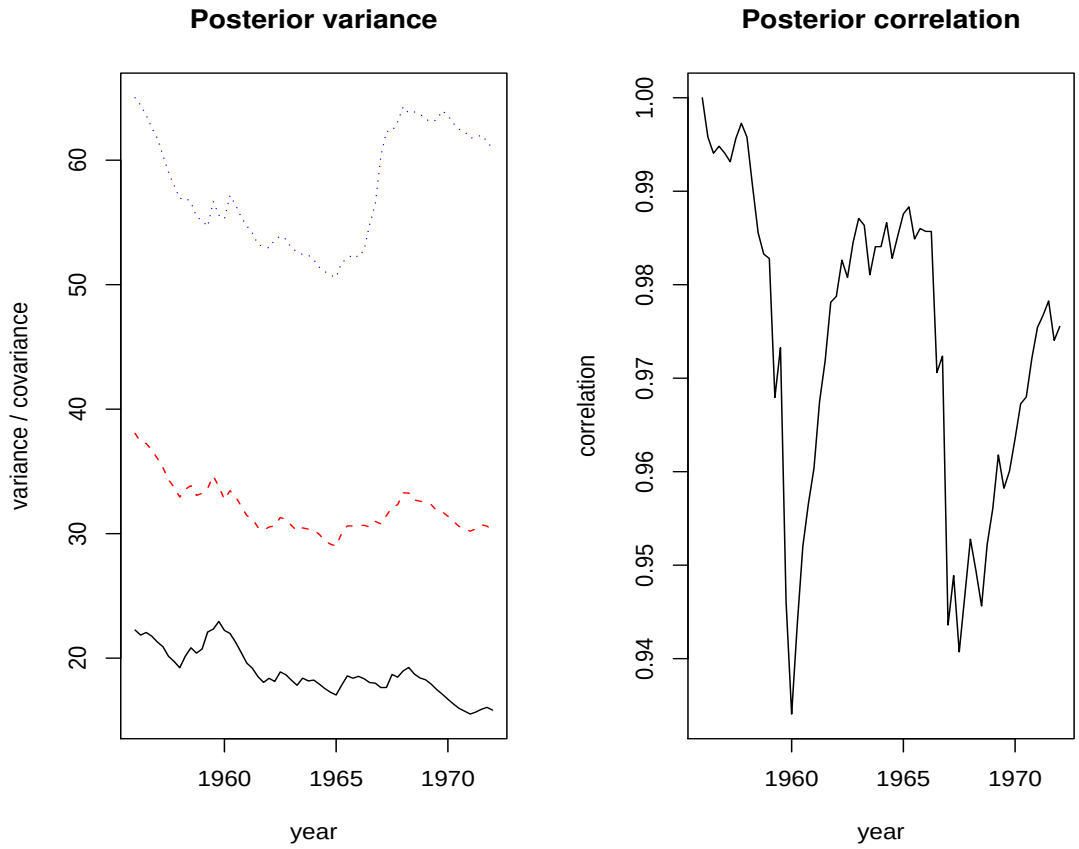

Figure 2: Posterior estimates of the observation covariance matrix $V = (V_{ij})_{i,j=1,2}$ and estimates of the correlation $\rho = V_{12}/\sqrt{V_{11}V_{22}}$. In the left panel graph, shown are: estimate of the variance V_{11} (solid line), estimate of the variance V_{12} (dashed line), and estimate of the variance V_{22} (dotted line). In the right panel graph, the solid line shows the estimate of ρ.

are indicated and, although the authors do believe that Bayes linear methods offer a great statistical tool, it is pointed out that in some problems, considered in this paper and in particular for time series data, the PSPP modelling approach can offer advantages as opposed to Bayes linear methods.

PSPP models are developed having in mind Bayesian inference for multivariate state space models when the observation covariance matrix is unknown and it is subject to estimation. This paper outlines the deficiency of the existing methods to tackle this problem and it is shown empirically that, for a class of important time series data, including local level, linear trend and seasonal components, PSPP generates much more accurate and reliable posterior estimators, which are remarkably fast and applicable to a wide range of time series data. US investment and change in inventory data are used to illustrate the capabilities of the PSPP state space models.

Given the similarities of the PSPP with Bayes linear methods, it is believed that the applicability of the PSPP approach goes beyond the examples considered in this paper. For example one area that is only slightly touched, is inference for data following non-normal

distributions, other than the multivariate t, the inverted multivariate t, and the Wishart distributions. In this sense a more detailed comparison of PSPP with Bayes linear methods and in particular with Bayes linear kinematics (Goldstein and Shaw, 2004), should shed more light on the performance of PSPP. It is our purpose to consider such comparisons in a future paper.

Acknowledgements

The authors are grateful to the Statistics Department at Warwick University, where this work was initiated. We are grateful to three referees for providing helpful comments.

Appendix

Proof of Theorem 1. ($\Longrightarrow$) By hypothesis $\mathbb{E}(X|Y) = \mu_x + A_{xy}(Y - \mu_y) \Rightarrow \mathbb{E}(X - A_{xy}Y|Y) = \mu_x - A_{xy}\mu_y = \text{constant}$. Furthermore $\text{Var}(X|Y) = E\{(X-\mu_x-A_{xy}(Y-\mu_y))(X-\mu_x-A_{xy}(Y-\mu_y))'|Y\} = \Sigma_x - A_{xy}\Sigma_y A'_{xy} = \text{constant} \Rightarrow \text{Var}(X - A_{xy}Y|Y) = \text{Var}(X|Y) = \text{constant}$. It follows that $X - A_{xy}Y \perp_2 Y$.

($\Longleftarrow$) The assumption $X - A_{xy}Y \perp_2 Y$ implies that $\mathbb{E}(X - A_{xy}Y|Y) = \mu$ constant $\Rightarrow \mathbb{E}(X|Y) = A_{xy}Y + \mu$, which is a linear function of Y. Given that $\mathbb{E}(X|Y)$ minimizes the quadratic prior expected risk and $\mu_x + A_{xy}(Y - \mu_y)$ minimizes this risk among all linear estimators, it follows that $\mathbb{E}(X|Y) = \mu_x + A_{xy}(Y - \mu_y)$. □

References

[1] Abraham, B. and Ledolter, A. (1983) *Statistical Methods for Forecasting*. Wiley, New York.

[2] Ameen, J.R.M. and Harrison, P.J. (1984) Discount weighted estimation. *Journal of Forecasting* **3**, 285-296.

[3] Ameen, J.R.M. and Harrison, P.J. (1985) Normal discount Bayesian models. In *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (Eds). North-Holland, Amsderdam, and Valencia University Press.

[4] Barbosa, E. and Harrison, P.J. (1992) Variance estimation for multivariate dynamic linear models. *Journal of Forecasting* **11**, 621-628.

[5] Box, G.E.P. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Massachusetts.

[6] Dickey, J.M. (1967) Matrix-variate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *Annals of Mathematical Statistics* **38**, 511-518.

[7] Durbin, J. and Koopman, S.J. (2001) *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

[8] Fahrmeir, L. (1992) Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association* **87**, 501-509.

[9] Fahrmeir, L. and Kaufmann, H. (1987) Regression models for non-stationary categorical time series. *Journal of Time Series Analysis* **8**, 147-160.

[10] Fahrmeir, L. and Kaufmann, H. (1991) On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. *Metrika* **38**, 37-60.

[11] Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn. Springer-Verlag, New York.

[12] Gamerman, D. (1997) *Markov Chain Monte Carlo - Stochastic simulation for Bayesian inference*. Chapman and Hall, New York.

[13] Godolphin, E.J. (2001) Observable trend-projecting state-space models. *Journal of Applied Statistics* **28**, 379-389.

[14] Godolphin, E.J. and Triantafyllopoulos, K. (2006) Decomposition of time series models in state-space form. *Computational Statistics and Data Analysis* **50**, 2232-2246.

[15] Goldstein, M. (1976). Bayesian analysis of regression problems. *Biometrika* **63**, 51-58.

[16] Goldstein, M. (1979). The variance modified linear Bayes estimator. *Journal of the Royal Statistical Society Series B* **41**, 96-100.

[17] Goldstein, M. (1983). General variance modifications for linear Bayes estimators. *Journal of the American Statistical Association* **78**, 616-618.

[18] Goldstein, M. and Shaw, S. (2004) Bayes linear kinematics and Bayes linear Bayes graphical models. *Biometrika* **91**, 425-446.

[19] Goodwin, P. (1997) Adjusting judgemental extrapolations using Theil's method and discounted weighted regression. *Journal of Forecasting* **16**, 37-46.

[20] Gupta, A.K. and Nagar, D.K. (1999). *Matrix Variate Distributions*. Chapman and Hall, New York.

[21] Hartigan, J.A. (1969) Linear Bayesian methods. *Journal of the Royal Statistical Society Series B* **31**, 446-454.

[22] Harvey, A.C. (1986) Analysis and generalisation of a multivariate exponential smoothing model. *Management Science* **32**, 374-380.

[23] Harvey, A.C. (1989) *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

[24] Harvey, A.C. (2004) Tests for cycles. In *State Space and Unobserved Component Models: Theory and Applications*, A.C Harvey, S.J. Koopman and N. Shephard (Eds.). Cambridge University Press, Cambridge.

[25] Horn, R.A. and Johnson, C.R. (1999) *Matrix Analysis.* Cambridge University Press, Cambridge.

[26] Kedem, B. and Fokianos, K. (2002) *Regression Models for Time Series Analysis*. Wiley, New York.

[27] Khatri, C.G., Khattree, R. and Gupta, R.D. (1991) On a class of orthogonal invariant and residual independent matrix distributions. *Sankhyā Series B* **53**, 1-10.

[28] Kitagawa, G. and Gersch, W. (1996) *Smoothness Priors Analysis of Time Series*. Springer-Verlag, New York.

[29] Lauritzen, S. (1996) *Graphical Models.* Oxford University Press, Oxford.

[30] Leonard, T. and Hsu, J.S.J. (1999) *Bayesian Methods.* Cambridge University Press, Cambridge.

[31] Lütkepohl, H. (1993) *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.

[32] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London.

[33] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, (2nd. edition). Chapman and Hall, London.

[34] McKenzie, E. (1974) A comparison of standard forecasting systems with the Box-Jenkins approach. *The Statistician* **23**, 107-116.

[35] McKenzie, E. (1976) An analysis of general exponential smoothing. *Operational Research* **24**, 131-140.

[36] Mouchart, M. and Simar, L. (1984) A note on least-squares approximation in the Bayesian analysis of regression models. *Journal of the Royal Statistical Society Series B*, **46**, 124-133.

[37] O'Hagan, A. and Forster, J.J. (2004) *Bayesian Inference*, 2nd edn. Kendall's Advanced Theory of Statistics, Vol. 2B. Arnold, London.

[38] Pilz, J. (1986) Minimax linear regression estimation with symmetric parameter restrictions. *Journal of Statistical Planning and Inference* **13**, 297-318.

[39] Pitt, M.K. and Shephard, N. (1999) Time varying covariances: a factor stochastic volatility approach (with discussion). In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds.), *Bayesian Statistics 6*, 547-570, Oxford University Press, Oxford.

[40] Quintana, J.M. and West, M. (1987). An analysis of international exchange rates using multivariate DLMs. *The Statistician* **36**, 275-281.

[41] Quintana, J.M. and West, M. (1988) Time series analysis of compositional data. In *Bayesian Statistics 3*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.). Oxford University Press, Oxford, 747-756.

[42] Reinsel, G.C. (1997) *Elements of Multivariate Time Series Analysis*, 2nd. ed. Springer-Verlag, New York.

[43] Salvador, M. and Gargallo, P. (2004). Automatic monitoring and intervention in multivariate dynamic linear models. *Computational Statistics and Data Analysis* **47**, 401-431.

[44] Salvador, M., Gallizo, J.L. and Gargallo, P. (2003). A dynamic principal components analysis based on multivariate matrix normal dynamic linear models. *Journal of Forecasting* **22**, 457-478.

[45] Salvador, M., Gallizo, J.L. and Gargallo, P. (2004). Bayesian inference in a matrix normal dynamic linear model with unknown covariance matrices. *Statistics* **38**, 307-335.

[46] Smith, J.Q. (1992) Dynamic graphical models. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds.). Oxford University Press, Oxford, 741-751.

[47] Srivastava, M. and Sen, A. (1990) *Regression Analysis: Theory, Methods and Applications*. Springer-Verlag, New York.

[48] Tiao, A.C. and Zellner, A. (1964) Bayes' theorem and the use of prior knowledge in regression analysis. *Biometrika* **51**, 219-230.

[49] Triantafyllopoulos, K. (2007) Covariance estimation for multivariate conditionally Gaussian dynamic linear models. *Journal of Forecasting* (to appear).

[50] Triantafyllopoulos, K. (2006a) Multivariate discount weighted regression and local level models. *Computational Statistics and Data Analysis* **50**, 3702-3720.

[51] Triantafyllopoulos, K. (2006b) Multivariate control charts based on Bayesian state space models. *Quality and Reliability Engineering International* **22**, 693-707.

[52] Triantafyllopoulos, K. and Pikoulas, J. (2002) Multivariate regression applied to the problem of network security. *Journal of Forecasting* **21**, 579-594.

[53] West, M. and Harrison, P.J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd edn. Springer-Verlag, New York.

[54] Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics.* Wiley, New York.

[55] Wilkinson, D.J. and Goldstein, M. (1996) Bayes' linear adjustment for variance matrices. In J.M. Bernardo et al. editors, *Bayesian Statistics 5*, 791-800, Oxford University press, Oxford.

[56] Wilkinson, D.J. (1997) Bayes linear variance adjustment for locally linear DLMs. *Journal of Forecasting* **16**, 329-342.