Scharnagl, Michael; Schumacher, Christian

Conference Paper

Finding good predictors for inflation by shotgun stochastic search

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Forecasting Methods, No. A11-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Scharnagl, Michael; Schumacher, Christian (2010) : Finding good predictors for inflation by shotgun stochastic search, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Forecasting Methods, No. A11-V2, Verein für Socialpolitik, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/37181

Finding good predictors for inflation by shotgun stochastic search*,†

Michael Scharnagl
Deutsche Bundesbank

Christian Schumacher
Deutsche Bundesbank

20 October 2009

Abstract

This paper evaluates a novel sampling algorithm, called shotgun stochastic search (S^3), for Bayesian model averaging in the context of finding predictors for inflation when the set of potential predictors is large. This is a relevant case in the forecasting literature, where often hundreds of predictors are compared with autoregressive distributed lag models for inflation. With such a large model space, standard Bayesian approaches like MCMC model composition (MC^3) tend to converge slowly. On the other hand, S^3 systematically searches in the neighborhood of good models and concentrates on regions of high posterior probability in the model space. We carry out a Monte Carlo simulations to compare the computational efficiency of S^3 to MC^3, based on standard data generating processes from the literature. When many potential predictors are available, S^3 outperforms MC^3. In an empirical exercise, we apply the two algorithms to find predictors for US inflation from a set of about one hundred indicators and their lags. S^3 absorbs posterior mass much quicker than MC^3 and makes Bayesian estimation of the standard inflation equations with many predictors computationally feasible.

JEL classification E31, E37, C52, C11

1 Introduction

Finding good indicators for inflation is a highly relevant task for the conduct of monetary policy. Many empirical exercises are based on autoregressive distributed lags models following Stock and Watson (1999, 2002), where future inflation is regressed on a small set of indicators and their lags. Recently, also Bayesian estimation techniques have been

*This paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank.

†Address of authors: Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main, Michael Scharnagl: Phone: ++49/+69-9566-2305, E-mail: michael.scharnagl@bundesbank.de, Christian Schumacher: Phone: ++49/+69-9566-2939, E-mail: christian.schumacher@bundesbank.de.

employed to assess the relative importance of indicators for inflation. For example, Jacobson and Karlsson (2004) and Eklund and Karlsson (2007) use Bayesian model averaging (BMA) for that purpose and provide inclusion probabilities to assess the information content of predictors. BMA has the general advantage of taking explicitly into account model uncertainty with respect to the proper selection of indicators and is thus tailor-made to make probability statements about the importance of indicators, see the general survey in Hoeting et al. (1999). However, as standard Bayesian simulation techniques usually have slow convergence properties, these approaches are restricted to relatively small datasets so far and taking into account large datasets of hundreds of variables seems infeasible as in Stock and Watson (1999, 2002) and De Mol et al. (2008).

Recently, Hans et al. (2007) have proposed a novel algorithm called shotgun stochastic search (S^3) to regressions with a large set of potential regressors. The key feature of S^3 is the thorough analysis of the neighborhood of a particular model, where the neighborhood is defined by the model combinations that emerge from adding, deleting or swapping a few variables from the current model in the chain. By evaluating all models in the neighborhood and deriving a proposal from them leads to a quick approach of regions with a high posterior probability. Hans et al. (2007) employ S^3 in the context of the gene expression cancer genomics. In this paper, we evaluate to what extent this new search algorithm can be a useful alternative to standard MCMC techniques such as MCMC model composition (MC^3) by Raftery et al. (1997) and Brown et al. (2002), that have been employed in the context of finding good predictors for inflation in the recent literature, see Jacobson and Karlsson (2004) and Eklund and Karlsson (2007). In particular, we want to check to what extent S^3 can be useful to find good predictors for inflation when the number of predictors is large.

We carry out a Monte Carlo analysis to compare S^3 with MCMC model composition (MC^3). The design of the MC exercise follows the recent literature and thus is based on well-known DGPs, see Fernandez, Ley, and Steel (2001). In our results, we find substantial improvements in computational efficiency by S^3 in all our DGPs chosen. If we expand the DGP to consider large datasets of potential predictors, the computational gains of S^3 are even more pronounced.

To illustrate the empirical performance of the method, we also carry out an empirical exercise based on the data from Stock and Watson (2002), which contains 131 potential predictors. We compare the relative performance of the algorithms with respect to their ability to accumulate posterior mass, when the forecast model can include the predictors and up to six lags of them. In this very large model space, S^3 accumulate posterior mass quickly. Based on these results, we make an attempt to find good predictors for US inflation in terms of inclusion probabilities. We discuss the results for different subsamples. In line with the results from the empirical literature, for example Banerjee and Marcellino (2006) and De Mol et al. (2008), we find considerable instability in the selection of indicators over time. However, by looking at group inclusion probabilities following Scharnagl and Schumacher (2007), we can at least identify clusters or groups of variables that provide stable information content for inflation. The S^3 algorithm makes such an analysis feasible.

The paper proceeds as follows: In section 2, we discuss S^3 and its relationship to MC^3.

Section 3 contains the Monte Carlo results, section 4 the empirical results, and section 5 concludes.

2 Shotgun stochastic search

2.1 Bayesian model averaging and inclusion probabilities

Consider the multiple regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is a $(T \times 1)$-dimensional vector, $\mathbf{X}$ is $(T \times k)$-dimensional and contains the observations of the predictors, and $\boldsymbol{\varepsilon}$ is a vector of residuals. Consider the problem of making inference on the determinants of $\mathbf{y}$ given data $\mathbf{y}$ and $\mathbf{X}$, when there are no restrictions with respect to the size of the model, i.e. different model dimensions are admissible. Let p be equal to the number of possible predictor variables. Then, the $(p \times 1)$-dimensional indicator vector $\boldsymbol{\gamma}_i = (\gamma_1, \ldots, \gamma_p)'$ with

$$\gamma_j = \begin{cases} 1 & \text{if variable } j \text{ is in model } \mathcal{M}_i \\ 0 & \text{else} \end{cases} \tag{2}$$

describes a particular model $\mathcal{M}_i$ from the model space $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$, where $M = 2^p$. Given data $\mathbf{y}$ and $\mathbf{X}$, the posterior model probability of a model $\mathcal{M}_i$ is

$$p(\boldsymbol{\gamma}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\gamma}_i)p(\boldsymbol{\gamma}_i)}{\sum_{j=1}^{R} p(\mathbf{y}|\boldsymbol{\gamma}_j)p(\boldsymbol{\gamma}_j)}, \tag{3}$$

where $p(\mathbf{y}|\boldsymbol{\gamma}_i)$ is the marginal likelihood, and $p(\boldsymbol{\gamma}_i)$ is the prior model probability. Given the posterior model probabilities, we can make inference on the quantities of interest by Bayesian model averaging (BMA). In our context, we are interested in the relevance of indicators included in $\mathbf{X}$. To assess this relevance, we rely on inclusion probabilities defined for variable x_i as

$$p(x_i|\mathbf{y}) = \sum_{j=1}^{M} \mathbb{I}\{x_i \in \mathcal{M}_j\} p(\boldsymbol{\gamma}_j|\mathbf{y}), \tag{4}$$

using the posterior model probabilities defined above and the variable-specific indicator function $\mathbb{I}\{x_i \in \mathcal{M}_j\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function that equals one if the set defined by the condition inside curly brackets is non-empty, or zero otherwise. Of course, we can rewrite the indicator function as $\mathbb{I}\{x_i \in \mathcal{M}_j\} = \mathbb{I}\{\gamma_i = 1\}$. The statistic (4) is equal to the sum of the posterior weights of all models that contain variable x_i. It can thus be regarded as the posterior probability of a variable being in the forecast model. Jacobson and Karlsson (2004) and Eklund and Karlsson (2007) employ inclusion probabilities to assess the relevance of indicators for Swedish inflation. Scharnagl and Schumacher (2007) provide an application to Euro area inflation.

2.2 The S^3 algorithm

If p is large, the dimension M of the model space is large, and simulation methods have to be employed to search over the model space $\mathcal{M}$. The basic idea of shotgun stochastic search is that in the neighborhood of the current model, we can expect several other models with a similar fit. Therefore, the identification and evaluation of this neighborhood as a description of this specific region of model space could be fruitful. Shotgun Stochastic Search (S^3) identifies the neighborhood as each regression model that differs from the current model in one variable. It compares all models that differ from the current one in this respect, and thus "shoots" in various directions. This is done by looking at a score, which is in most cases equal to the posterior probability $p(\boldsymbol{\gamma}_i|\mathbf{y})$. From the neighborhood, a new candidate model is chosen. Evaluating all models close to the present one in parallel helps to move in the direction and the exploration of regions of model space with high posterior probabilities, i.e. searching for many good models in the neighborhood of good models.

Let us denote the set of models collected by S^3 by $\mathcal{G}$. Starting point of the algorithm is model $\boldsymbol{\gamma}^{[0]}$, and therefore $\mathcal{G} = \{\boldsymbol{\gamma}^{[0]}\}$. A constant B is chosen which denotes the maximum number of elements in $\mathcal{G}$. For $r = 1, ..., R$, the following steps are iterated:

- Step 1: Given $\boldsymbol{\gamma}^{[r]}$, construct the neighborhood

$$\text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right) = \left\{\boldsymbol{\gamma}^{+\,[r]}, \boldsymbol{\gamma}^{o\,[r]}, \boldsymbol{\gamma}^{-\,[r]}\right\} \tag{5}$$

  and compute the posterior model probability $p(\boldsymbol{\gamma}|\mathbf{y})$ for all $\boldsymbol{\gamma} \in \text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right)$. Update the model space $\mathcal{G}$ according to $\mathcal{G} \cup \text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right)$. If $|\mathcal{G}| > B$, remove $|\mathcal{G}| - B$ models with lowest scores. The neighborhood is defined as follows:

  - $\boldsymbol{\gamma}^{+[r]}$, addition: one variable is added from the set of currently excluded variables. The new model contains $k + 1$ variables.
  - $\boldsymbol{\gamma}^{o[r]}$, replacement: one of the currently included k variable is replaced by one of the currently excluded $p - k$ variables. The number of included variables does not change.
  - $\boldsymbol{\gamma}^{-[r]}$, deletion: one variable is excluded from the model. The new model contains $k - 1$ variables.

- Step 2: Sampling of single models $\boldsymbol{\gamma}_*^{+\,[r]}$, $\boldsymbol{\gamma}_*^{o\,[r]}$ and $\boldsymbol{\gamma}_*^{-\,[r]}$ from $\boldsymbol{\gamma}^{+\,[r]}$, $\boldsymbol{\gamma}^{o\,[r]}$ and $\boldsymbol{\gamma}^{-\,[r]}$ separately, with probabilities proportional to $p(\boldsymbol{\gamma}|\mathbf{y})$ and normalization within each subset.

- Step 3: Sampling of a model $\boldsymbol{\gamma}^{[r+1]}$ from $\left\{\boldsymbol{\gamma}_*^{+\,[r]}, \boldsymbol{\gamma}_*^{o\,[r]}, \boldsymbol{\gamma}_*^{-\,[r]}\right\}$ with probabilities proportional to $p(\boldsymbol{\gamma}|\mathbf{y})$ and normalization within this set. With model $\boldsymbol{\gamma}^{[r+1]}$, we go to step 1.

In the end, $\mathcal{G}$ contains the B best models in terms of posterior model probability as found by S^3, not just the sequence of chosen models $\boldsymbol{\gamma}^{[0]}$, ..., $\boldsymbol{\gamma}^{[R]}$. In particular, $\mathcal{G}$ contains

the best models from the union of neighborhoods

$$\bigcup_{r=0}^{R} \text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right). \tag{6}$$

The hierarchical sampling in steps 2 and 3 takes into account that the three parts of the neighborhood have different dimensions. If $2 \leq k < p$, the sizes of the subsets in the neighborhood are $\left|\boldsymbol{\gamma}^{+[r]}\right| = p - k$, $\left|\boldsymbol{\gamma}^{o[r]}\right| = k\,(p-k)$, and $\left|\boldsymbol{\gamma}^{-[r]}\right| = k$, respectively. If $k = p$, $\boldsymbol{\gamma}^{+} = \emptyset$. By splitting the sampling into steps 2 and 3, we remove any dependence from the current size of a model and the three moves adding, deleting, and swapping become equally important a priori.

2.3 Comparison of S^3 and MC^3

The basic MC^3 approach is based on a chain with elements $r = 1, ..., R$. At each replication r, the algorithm consists of two steps, see Brown et al. (2002) and Jacobson and Karlsson (2004):

- Step 1: Given the last element of the chain $\boldsymbol{\gamma}^{[r]}$, a new candidate model $\boldsymbol{\gamma}'$ is chosen following two moves:
  - move 1: With probability p_A a variable is drawn from the set of all potential variables. If this variable is already included in the current model $\boldsymbol{\gamma}^{[r]}$ it will be dropped and if it is not it will be added.
  - move 2: With probability $1 - p_A$ a randomly chosen variable from the current model is substituted by a randomly drawn variable from the set of excluded variables.
- Step 2: The candidate model corresponding to $\boldsymbol{\gamma}'$ is accepted with probability

$$\alpha = \min\left\{1, \frac{p\left(\mathbf{y}|\boldsymbol{\gamma}'\right) p\left(\boldsymbol{\gamma}'\right)}{p\left(\mathbf{y}|\boldsymbol{\gamma}^{[r]}\right) p\left(\boldsymbol{\gamma}^{[r]}\right)}\right\}. \tag{7}$$

  Here, $\boldsymbol{\gamma}^{[r]}$ represents the current model. $p\left(\mathbf{y}|\boldsymbol{\gamma}'\right)$ is the marginal likelihood, and $p\left(\boldsymbol{\gamma}\right)$ the model prior. If the draw accepts the candidate, $\boldsymbol{\gamma}'$ becomes $\boldsymbol{\gamma}^{[r+1]}$, and we go to step 1.

In step 1, Brown et al. (2002) use $p_A = 0.5$. Thus, swapping in move 2 receives the same probability as move 1.[1] Raftery et al. (1997) only carry out move 1, without move 2, thus neglecting candidate models of constant size. Sampling the candidate model as above can also be described by application of the proposal distribution $T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[r]})$ defined as

$$T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[r]}) = \begin{cases} 0 & \text{for} \quad \boldsymbol{\gamma}' \notin \text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right) \\ \text{const} & \text{for} \quad \boldsymbol{\gamma}' \in \text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right) \end{cases}, \tag{8}$$

[1] Due to the dimensional imbalance of the two groups, the probability of selecting a model in move 1 is $\frac{1}{p}$ and $\frac{1}{k}\frac{1}{p-k}$ for move 2, where p is the number of potential explanatory variables and k is the number of variables included in the current model.

where $\text{nbd}(\boldsymbol{\gamma}^{[r]})$ can be defined as in Raftery et al. (1997) with $\text{nbd}(\boldsymbol{\gamma}^{[r]}) = \{\boldsymbol{\gamma}^{+}, \boldsymbol{\gamma}^{-}\}$, thus model sets obtained from swapping $\boldsymbol{\gamma}^{o}$ are neglected. Following the hierarchical selection with swapping by Brown et al. (2002) and Jacobson and Karlsson (2004), the probability of selecting a model in move 1 is $\frac{1}{p}$ and $\frac{1}{k}\frac{1}{p-k}$ for move 2, where p is the number of potential explanatory variables and k is the number of variables included in the current model. Thus, at each replication, MC^3 selects a candidate model without referring to its posterior model probability. Only in step 2, the decision on accepting or rejecting the candidate takes into account the posterior model probabilities.

There are at least three major differences between MC^3 and S^3:

1. The new candidate model within MC^3 is chosen according to the two-step procedure just described. Within each move, the probability for each model defining the relevant part of the neighborhood is the same. In S^3, the choice of the candidate model depends on the model posterior probabilities in the whole neighborhood. Neglecting the dimensional imbalance in the subsets of the neighborhood, Hans et al. (2007) define the proposal distribution of a candidate model in S^3 according to

$$T(\boldsymbol{\gamma}'; \boldsymbol{\gamma}^{[r]}) = \frac{p\left(\mathbf{y}|\boldsymbol{\gamma}'\right) p\left(\boldsymbol{\gamma}'\right) \times 1(\boldsymbol{\gamma}' \in \text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right))}{\sum_{\boldsymbol{\gamma} \in \text{nbd}\left(\boldsymbol{\gamma}^{[r]}\right)} p\left(\mathbf{y}|\boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right)}. \tag{9}$$

   Thus, the move to a new candidate model is highly dependent on its posterior probability. Therefore, S^3 concentrates on specific regions of the model space, namely those with high posterior mass, whereas MC^3 neglects the posterior information for selecting a candidate model as in (8).

2. The chain in MC^3 includes all accepted candidate models. The chain in S^3 includes all neighborhoods of all candidate models, dependent on B and the posterior mass already in the chain.

3. The dimensional differences in the neighborhood are taken into account in different ways. However, the neighborhood in MC^3 can be defined exactly as in S^3. Thus, we can remove the dependence on the dimensional discrepancies between deleting, adding, and swapping completely.

All in all, we expect the difference in the proposal distribution (difference 1) as most relevant. Whereas the other two differences can be accounted for easily, the differences in how a candidate is selected can affect the chain heavily according to the results in Hans et al. (2007). We will now compare the two approaches in a Monte Carlo analysis.

3 Monte Carlo simulations

To assess the computational effectiveness of S^3, we carry out a Monte Carlo simulation exercise, where S^3 is compared to MC^3 as proposed in Raftery et al. (1997) and Brown et al. (2002).[2] The design of the Monte Carlo analysis follows Fernandez, Ley and Steel

[2] As in Brown et al. (2002) and Jacobson and Karlsson (2004), we use $p_A = 0.5$ in step 1 of MC^3, see section 2.3.

(2001), Eklund and Karlsson (2007) and Hans et al. (2007) in terms of data generating process (DGP). Thus, we discuss the performance of S^3 in a standard framework that has been extensively employed in the economics literature. In an additional step, we extend that framework to a large regressor case, that might be relevant in the present context of finding predictors for inflation from a large set of indicators.

3.1 DGP

This data generating process is used by Fernandez, Ley and Steel (2001) and Eklund and Karlsson (2007). A $(T \times 15)$-dimensional matrix of 15 predictors $\mathbf{X}$ is generated with sample size $T = 100$. The first ten variables $\mathbf{x}_1$, ..., $\mathbf{x}_{10}$ are iid $N(0, 1)$ and the other five variables are constructed according to

$$(\mathbf{x}_{11}, ..., \mathbf{x}_{15}) = (\mathbf{x}_1, ..., \mathbf{x}_5) \begin{pmatrix} 0.3 & 0.5 & 0.7 & 0.9 & 1.1 \end{pmatrix}' \boldsymbol{\tau} + \mathbf{e} \tag{10}$$

with $\boldsymbol{\tau} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$. $\mathbf{e}$ is $(T \times 5)$-dimensional vector of shocks and iid $N(0, 1)$. This produces a correlation between the first five and the last five predictors. The theoretical correlation coefficient increases from 0.153 ($\mathbf{x}_1$) to 0.561 ($\mathbf{x}_5$). The theoretical value of the correlation between the last five regressors is 0.740. The endogenous variable is generated according to

$$y_t = 4 + 2x_{1,t} - x_{5,t} + 1.5x_{7,t} + x_{11,t} + 0.5x_{13,t} + \sigma\varepsilon_t \tag{11}$$

where the disturbances ε_t are iid $N(0, 1)$ and $\sigma = 2.5$. Of course, the DGP is in line with the general model (1), based on the true coefficient

$$\boldsymbol{\beta} = \begin{pmatrix} 2 & 0 & 0 & 0 & -1 & 0 & 1.5 & 0 & 0 & 0 & 1 & 0 & 0.5 & 0 & 0 \end{pmatrix}'.$$

Due to the high collinearity implied by the DGP, we denote this DGP in the results below as 'multicollinearity'. As an alternative, the same DGP without multicollinearity is employed, where (10) simplifies to

$$(\mathbf{x}_{11}, ..., \mathbf{x}_{15}) = \mathbf{e}. \tag{12}$$

We denote this DGP below as 'basic'. Thus, our experiment contains two DGPs with $p = 15$ variables. The analysis is based on 1000 Monte Carlo replications of the DGP. For each replication the number of MC^3 iterations or evaluated models in S^3 is 150000. Note that by relating the number of evaluated models in S^3 to the number of iterations in MC^3, where one model is evaluated only, we try to make the computational burden comparable. The number of burn-in iterations (or evaluated models) is 5000. In each replication, the algorithms start using a randomly drawn vector of five variables.

In addition to the DGP above, which is standard in the economics literature on Bayesian techniques cited above, we also consider a higher-dimensional model that contains in addition five variables, that are iid $N(0, 1)$ and do not help to explain y_t. Thus, this DGP provides $p = 20$ potential explanatory variables that are noisier and have over-

all less information content. In this case, it is more difficult for the two algorithms to find the good indicators for y_t. This goes more in the direction of finding regressors in large datasets as in Stock and Watson (1999), Eklund and Karlsson (2007) and De Mol et al. (2008). In large datasets as employed in the empirical application below, we often observe that certain indicators turn out to have little or no information content at all, for example longer lags for certain indicators. As the model space increases considerably by this extension, we increase also the number of evaluated models to 1000000.

3.2 Priors and posterior distributions

We now specify the prior distributions and define the marginal likelihood for a model $\mathcal{M}_i$ based on regressors $\mathbf{X}_i$ of size $(T \times k_i)$. The priors employed below are standard in the literature on Bayesian model averaging, for example Koop (2003). We employ the same prior distributions for MC^3 and S^3.

Concerning the prior distribution of the coefficients $\boldsymbol{\beta}$, we use a normal-gamma natural conjugate prior

$$\boldsymbol{\beta}_i | h \sim \text{N}\left(\mathbf{0}_{k_i}, h^{-1}\underline{\mathbf{V}}_i\right), \tag{13}$$

and the variance is specified by using the g-prior

$$\underline{\mathbf{V}}_i = \left(g_i \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \tag{14}$$

and

$$h = \sigma^2 \tag{15}$$

where σ^2 is the variance of the error term, see (11) above. Concerning the hyperparameter g_i, we follow Fernandez, Ley and Steel (2001), and use

$$g_i = \begin{cases} \frac{1}{p^2} & \text{if} \quad T \leq p^2 \\ \frac{1}{T} & \text{if} \quad T > p^2 \end{cases}, \tag{16}$$

where p is again the number of potential explanatory variables. Using the g-prior, the marginal likelihood of model $\mathcal{M}_i$ is

$$p\left(\mathbf{y}|\boldsymbol{\gamma}_i\right) \propto \left(\frac{g_i}{g_i+1}\right)^{\frac{k_i}{2}} \left[\frac{1}{g_i+1}\mathbf{y}'\mathbf{P}_{\mathbf{X}_i}\mathbf{y} + \frac{g_i}{g_i+1}\left(\mathbf{y}-\overline{y}\right)'\left(\mathbf{y}-\overline{y}\right)\right]^{-\frac{T-1}{2}} \tag{17}$$

with $\mathbf{P}_{\mathbf{X}_i} = \mathbf{I}_T - \mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'$. $\overline{y}$ is the in-sample mean of $\mathbf{y}$. The model prior is equal to

$$p\left(\boldsymbol{\gamma}_i\right) = \pi^{k_i}\left(1-\pi\right)^{p-k_i} \tag{18}$$

where π is a hyperparameter representing the probability that a variable is in the model. This induces a binomial prior distribution over model size

$$\Pr\left(|\boldsymbol{\gamma}| = k\right) = \left(\begin{array}{c} p \\ k \end{array}\right)\pi^k\left(1-\pi\right)^{p-k} \tag{19}$$

A priori, the expected model size equals $p\pi$. In our exercise, it is assumed that the model size k is on average 5, in line with DGP (11). The corresponding hyperparameter $\pi = 0.33$ in (18) is specified appropriately to ensure the prior model size.

Given $p(\boldsymbol{\gamma}_i)$ and $p(\mathbf{y}|\boldsymbol{\gamma}_i)$, we can calculate the posterior probability of a model $\mathcal{M}_i$ according to $p(\boldsymbol{\gamma}_i|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\gamma}_i)\, p(\boldsymbol{\gamma}_i)$ up to a normalizing constant which is equal for all models.

3.3 Criteria for evaluation of performance

In a first step, the complete set of all potential models is generated. This is possible as the number of all potential models in the DGP equals $M = 2^{15} = 32768$ which is rather small. All models are estimated and evaluated. In particular, for all models $\forall i = 1, \ldots, M$, the posterior probability $p(\boldsymbol{\gamma}_i|\mathbf{y})$ is calculated. This set is the basis for evaluating the performance of MC^3 and S^3. The performance can be evaluated on a variety of measures. As the number of iterations is very different for both algorithms due to the inclusion of complete neighborhoods by the S^3 algorithm, the evaluation is done in the context of the number of models evaluated rather than by just comparing the number of iterations.

Relative posterior density To investigate how quickly the two algorithms accumulate posterior mass, we count the number of model evaluations until a certain fraction of posterior mass of the true distribution is reached. Differences may occur, for example, if an algorithm has some tendency of visit models with low posterior probability too often. The calculations are based on the analytical posteriors as defined above.

Hans et al. (2007) compare MC^3 relative to S^3 only in an empirical exercise, as they are not able to evaluate all possible models due to the huge number of regressors. The DGP chosen here from economic applications, however, allows for a systematic Monte Carlo investigation, as the model space is small enough. All models can be evaluated, and it is possible to relate the posterior mass accumulated by both algorithms to the "true" mass.

This comparison is done in two steps: First, MC^3 and S^3 are applied to a draw of data from the DGP for a large number of model evaluations, in our example 150000. Second, we search in the chains from MC^3 and S^3, after which number of model evaluations a prespecified fraction of total posterior probability from the true distribution of models is accumulated. In particular, we present results for the relative probability ratios $[0.30, 0.50, 0.80, 0.90, 0.93, 0.95, 0.97]$. To get an impression on the time necessary to obtain the ratios above, we also search for the time (in seconds) elapsed and report that.

Finding the "true" vector We also check whether the true vector is part of the chains. Searching for the iteration number or number of models evaluated when the true vector is found gives an indication how fast the algorithms enter the regions of high posterior probability of the model space. Again, we also evaluate the time necessary for finding the true model.

3.4 Monte Carlo results

Below in table 1, we present results for the standard DGP with $p = 15$ variables with multicollinearity and without multicollinearity, denoted as basic. Based on the basic DGP

Table 1: Number of model evaluations need or time elapsed to reach relative posterior mass or find best model, $p = 15$ variables

A. Model evaluations

posterior mass	0.30	0.50	0.80	0.90	0.93	0.95	0.97	best model
basic								
MC^3	8858	14124	26420	34508	38017	41737	49417	16403
S^3	30	73	183	913	2246	5908	23304	83
multicollinearity								
MC^3	7901	14129	25915	37086	41131	45468	52288	17941
S^3	59	165	1313	3983	8960	21105	71524	12920

B. Time

posterior mass	0.30	0.50	0.80	0.90	0.93	0.95	0.97	best model
basic								
MC^3	3.32	4.61	7.60	9.56	10.42	11.33	13.19	5.16
S^3	1.30	1.31	1.34	1.51	1.84	2.73	6.98	1.31
multicollinearity								
MC^3	3.03	4.51	7.31	9.96	10.92	11.95	13.57	5.41
S^3	1.27	1.30	1.58	2.22	3.41	6.33	18.44	4.32

Note: In panel A, the table contains the number of model evaluations needed to reach a selected posterior mass, or, alternatively, to find the true model. In panel B, the entries are the time elapsed to reach posterior mass or find the true model. Details on the DGP and the simulation design can be found in reported in Section 3.1, 3.2, and 3.3.

without multicollinearity, S^3 is much faster in accumulating posterior mass than MC^3 up to 97% of the posterior mass. With multicollinearity, S^3 is only faster until 95% of the mass is accumulated. Thus, S^3 is slower in accumulating the total mass, as it visits low probability models extremely seldom, in particular when the regressors are correlated. However, until 95% of mass, S^3 is always faster. Also, the true models are found earlier than for MC^3. In the search, S^3 is extremely faster than MC^3 up to 80% mass, and slows down a little bit. Interestingly, although S^3 aims at searching over the regions of high posterior probability only, it still has the ability to scan the overall distribution of models.

In table 2, we present results for the DGP with $p = 20$ variables. Compared to the case with $p = 15$ variables, the time elapsed as well as the model evaluations needed increases considerably, indicating that the five additional regressors make it much more difficult for both MC^3 and S^3 to find the regions of high posterior mass in the model space. However, S^3 does now more clearly outperform MC^3 and is quicker in all examples shown. This also

Table 2: Number of model evaluations need or time elapsed to reach relative posterior mass or find best model, $p = 20$ variables

A. Model evaluations

posterior mass	0.30	0.50	0.80	0.90	0.93	0.95	0.97	best model
basic								
MC^3	182442	277736	458765	577900	628310	667377	714622	316753
S^3	48	98	213	954	2566	7798	43550	664
multicollinearity								
MC^3	183431	295591	513946	635608	679124	713715	752143	323811
S^3	57	113	601	3422	9745	23728	114474	63645

B. Time

posterior mass	0.30	0.50	0.80	0.90	0.93	0.95	0.97	best model
basic								
MC^3	57.91	81.79	127.03	156.78	169.36	179.09	190.83	91.51
S^3	11.60	11.62	11.64	11.81	12.18	13.37	21.48	11.74
multicollinearity								
MC^3	59.14	87.70	143.23	174.13	185.13	193.87	203.53	94.78
S^3	11.15	11.16	11.27	11.90	13.32	16.41	36.46	25.00

Note: In panel A, the table contains the number of model evaluations needed to reach a selected posterior mass, or, alternatively, to find the true model. In panel B, the entries are the time elapsed to reach posterior mass or find the true model. Details on the DGP and the simulation design can be found in reported in Section 3.1, 3.2, and 3.3.

holds for the multicollinear data. Thus, we can conclude that S^3 can cope with noisy data better than MC^3. Interestingly, Hans et al. (2007) argue that due to the concentration on regions of high posterior mass, S^3 might not be able to approximate the full distribution of models, as it neglects models with low posterior probability. However, our results show that more than 90% of posterior mass can be discovered quicker than MC^3. Thus, the S^3 algorithm seems to approximate the overall distribution quite well.

4 Empirical illustration

Below, we analyze the relative performance of MC^3 and S^3 on a large macroeconomic dataset from Stock and Watson (2002) that has been used in many applications, see for example De Mol et al. (2008) as a recent example. The data includes real variables (sectoral industrial production, employment and hours worked), nominal variables (consumer and producer price indices, wages, money aggregates), asset prices (stock prices and exchange rates), the yield curve and surveys, for a total of 131 variables. The sample has a monthly frequency and ranges from $1959M01$ to $2003M12$. The series are transformed to obtain stationarity. In general, for real variables, such as employment, industrial production, and sales, we take the monthly growth rate. We take first differences for series already expressed in rates: unemployment rate, capacity utilization, interest rate and some surveys. Prices and wages are transformed to first differences of annual inflation following De Mol et al. (2008). The variable we forecast is $y_{t+h}^h = \pi_{t+h} - \pi_t$ and annual inflation $\pi_t = 100 \times \ln(P_t/P_{t-12})$ with monthly CPI denoted P_t (series mnemonic PUNEW). Thus, the equation (1) contains y_{t+h}^h on the left-hand side. The forecasts for the level of inflation are recovered as $\pi_{T+h|T} = y_{T+h|T}^h + \pi_T$. We consider two estimation periods ending in $1970M01$ and $2002M12$ with a window of 10 years, i.e. parameters are estimated at each time using the most recent 10 years of data. The forecast horizon is $h = 12$ as in De Mol et al. (2008). On the right-hand side of equation (1), we consider not only t-dated predictors, but also up to 6 lags of them. Thus, our model is in line with the specifications by Stock and Watson (2002). Note that this choice implies that we have $p = 131 \times (6 + 1) = 917$ potential predictors on the right-hand side. Thus, the model space includes 2^{917} variables, which is very large compared to other applications on inflation forecasting, but not in the literature the S^3 algorithm is taken from, see Hans et al. (2007). Note that the 131 predictor variables include also inflation; thus, autoregressive terms are allowed to matter. Differently from the marginal likelihood employed in the Monte Carlo exercise above, we follow here Eklund and Karlsson (2007) and choose the predictive likelihood in order to compute model weights. In both samples, the training sample contains 60% of the initial observations, whereas the evaluation or hold-out sample contains the final 40% of observations, for details on the choice of these parameter settings, see Eklund and Karlsson (2007). In general, the predictive measures turn out to be superior to the marginal likelihood when structural instabilities are present. Based on the data used here, De Mol et al. (2008) indeed report certain instabilities with respect to the selection of predictors for inflation over time, see also Banerjee and Marcellino (2006). Following these results, we also use predictive posterior weights as in Eklund and

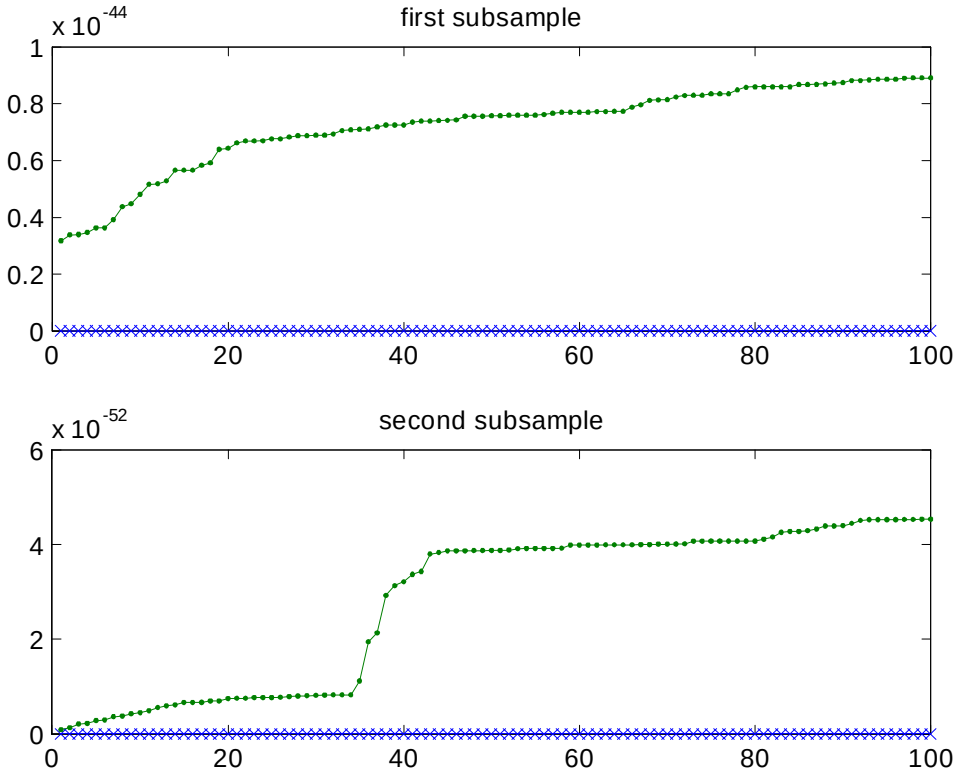
Karlsson (2007).

Given the data and models, we carry out two empirical exercises: First, we compare the relative performance of MC^3 and S^3 for the two sample sizes. We limit ourselves two this rather small number of periods in order to let the exercise remain computationally feasible. Second, we employ S^3 to make an attempt to find good predictors of US inflation in the data.

4.1 Relative performance with respect to US inflation

In this exercise, for a given number of 5000000 model evaluations, we report the posterior mass accumulated by the two algorithms. Thus, we can evaluate to what extent the results from the Monte Carlo exercise can be confirmed by empirics in the present context of forecasting US inflation. The results can be found in figure 1. The horizontal axis

Figure 1: Posterior predictive mass accumulated



Note: The figure shows posterior predictive mass accumulated by MC^3 (line with crosses) and S^3 (circle line) applied to US data dependent on the models evaluated in the respective chains. On the horizontal axis, the model evaluations are rescaled so that 100 corresponds to the maximum number of model evaluations (5000000). The two subsamples end in 1970M01 and 2002M12, respectively, and contain a window of 10 years of data.

displays the model evaluations, rescaled so 100 corresponds to the maximum number of model evaluations (5000000). The results show that S^3 is capable of accumulating posterior predictive mass much quicker than MC^3. Even if the number of models evaluated

increases, MC^3 cannot catch up to S^3.[3] Thus, S^3 seems to be a more useful algorithm in the present context with a very large dimension of the data.

4.2 What are the good predictors for US inflation?

Based on S^3 only, we compute inclusion probabilities that help to distinguish good from bad predictors of inflation. De Mol et al. (2008) carry out a similar exercise based on least-angle regressions and find a temporal instability of the selection of variables. Banerjee and Marcellino (2006) evaluate ADL models with single indicators for US inflation with respect to their out-of sample performance and find that the best predictors vary over time. Complementing the work by De Mol et al. (2008), we provide inclusion probabilities for predictors estimated by BMA, see (4) above. The simulation-based techniques employed here allow us to make probability statements on the relevance of indicators for US inflation. In table 3, we provide the inclusion probabilities for different variables that perform best. In each case, the inclusion probabilities are defined as the sum of posterior predictive model weights of those models that include a particular variable or its lags. The tables show

Table 3: Inclusion probabilities of variables

1st sample		2nd sample	
Building permits, total (HSBR)	1.00	Employees, mining (CES006)	1.00
Housing starts, west (HSWST)	0.92	C&I loans outstanding (FCLNQ)	0.84
CPI, commodities (PUC)	0.14	Personal income (a0m052)	0.40
Consumer expectations (HHSNTN)	0.12	Interest rate, U.S.Treasury, 1-yr. (FYGT1)	0.40
PCE deflator (GMDCN)	0.08	Personal income less transfers (A0M051)	0.37
PCE deflator, non-durables (GMDC)	0.08	Employees, non-durables (CES033)	0.33
Building permits, west (HSBWST)	0.08	Employees, manufacturing (CES015)	0.27
M2, real (FM2DQ)	0.07	Commercial paper minus Fed funds rate (scp90)	0.25
CPI, all items (PUNEW)	0.07	Persons unemployed 27 weeks + (LHU27)	0.12
CPI, all items less shelter (PUXHS)	0.05	Interest rate, U.S.Treasury bills, 6-mo. (FYGM6)	0.10

Note: The entries represent inclusion probabilites as the sum of posterior predictive model weights that include a particular variable or its lags. The two subsamples end in 1970M01 and 2002M12, respectively.

that indeed the relevant indicators change over time. In particular, we find no indicator in the top 10 of the second sample, that is also member of the top 10 in the first sample. Interestingly, lags of inflation (PUNEW) matter in the first part of the sample, whereas they do not rank among the top 10 in the second sample, perhaps reflecting the decline in inflation persistence in the Great Moderation. Overall, our findings of instability are in line with De Mol et al. (2008). They suggest that due to the collinearity and instability in the data, different indicators are selected over time and estimation can be very sensitive to minor perturbations of the data and model specifications. The authors indicate that representatives of clusters or groups of variables might change over time. We follow this conjecture and investigate the role of groups of variables based on an identification of representatives of particular groups of indicators, following Scharnagl and Schumacher

[3] Indeed, MC^3 also accumulates posterior mass, but to a small extent only, so it does not show up in the figure due to the huge difference to S^3.

(2007). The group inclusion probabilities are defined as the sum of posterior predictive model weights that include at least one representative variable or its lags of a particular group. In the dataset used here, the groups are defined following the classification by Stock and Watson (2002). In particular, we distinguish the groups: employment and hours, exchange rates, housing starts and sales, interest rates and spreads, inventories and orders, money and credit, price indexes and wages, real output and income, sales and stock prices. Given this classification, we investigate whether at least some groups can be identified that matter for inflation in both periods of time. Results are presented in table 4. In the first subsample, the groups housing starts and sales (1.00), price indexes and

Table 4: Group inclusion probabilities

	1st sample	2nd sample
employment and hours	0.19	1.00
exchange rates	0.04	0.03
housing starts and sales	1.00	0.05
interest rates and spreads	0.09	0.81
inventories and orders	0.05	0.04
misc	0.12	0.05
money and credit	0.22	0.85
price indexes and wages	0.58	0.35
real output and income	0.16	0.80
retail, manufacturing and trade sales	0.02	0.01
stock prices	0.07	0.03

Note: The entries represent group inclusion probabilites defined as the sum of posterior predictive model weights that include at least one variable or its lags from a particular group. The two subsamples end in 1970M01 and 2002M12, respectively.

wages (0.58), money and credit (0.21), employment and hours (0.19), and real output and income (0.16) have the highest group inclusion probabilities. In the second subsample, the groups employment and hours (1.00), money and credit (0.85), interest rates and spreads (0.81), real output and income (0.80), and price indexes and wages (0.35) matter most. Thus, also group inclusion probabilities indicate a considerable degree of instability of the relevance of indicators. Although there are some groups of predictors that seem to have information content in both subsamples, their groups inclusion probabilities differ in the subsamples quite substantially. These groups are employment and hours, money and credit, real output and income, and price indexes. Apart from these groups, there seems to be little information content in the remaining ones.

5 Conclusions

The present paper considers Shotgun Stochastic Search (S^3) as a competitive algorithm to search for predictors of inflation, when the number of potential predictors is large. In a Monte Carlo exercise and an empirical application for US inflation with about 131 predictors and their lags, S^3 outperforms standard MC^3 in terms of a quicker accumulation of posterior predictive mass. Thus, BMA with selecting candidate models according to

their relative posterior weights seems to be superior to randomly sampling of candidate models. Thus, S^3 provides us with an interesting way to the Bayesian estimation of the widely-used autoregressive distributed lag model by Stock and Watson (1999, 2002) with sampling techniques when the set of potential predictors is large.

Of course, the present investigation relied only on one particular set of Monte Carlo simulations and one application to a particular dataset, although both are standard in the literature. Depending on the particular problem, it should be checked whether S^3 is appropriate. A drawback could be that this algorithm concentrates too much on the neighborhood and, perhaps, local areas of high mass. Thus, it cannot be ruled out that other regions are left out. In these environments, it could be useful to consider to jump between traditional MC^3 and S^3 within one chain. Another more general drawback of S^3 is perhaps that despite its computational gains, the simulation-based BMA approach still requires a lot of computing time. This makes it difficult to estimate a model recursively as in Stock and Watson (1999, 2002), when the number of recursions is large. However, in case we have a moderate number of recursions and the main purpose of the analysis is to make probability statements, say, on the inclusion of variables from a large set, the BMA approach based on S^3 is a reasonable choice.

References

Banerjee, A., M. Marcellino (2006), Are there any reliable leading indicators for US inflation and GDP growth?, International Journal of Forecasting 22, 137-151.

Brown, P.J., M. Vannucci, T. Fearn (2002), Bayes model averaging with selection of regressors, Journal of the Royal Statistical Society B 64, 519-536.

De Mol, C., D. Giannone, L. Reichlin (2008), Forecasting Using a Large Number of Predictors: Is Bayesian Regression a Valid Alternative to Principal Components?, Journal of Econometrics 146, 318-328.

Doppelhofer, G., M. Weeks (2009), Jointness of Growth Determinants, Journal of Applied Econometrics 24, 209-244.

Eklund, J., S. Karlsson (2007), Forecast Combination and Model Averaging Using Predictive Measures, Econometric Reviews 26, 329-363.

Fernandez, C., E. Ley, M. Steel (2001), Benchmark Priors for Bayesian Model Averaging, Journal of Econometrics 100, 381-427.

George, E.I., R.E. McCulloch (1997), Approaches for Bayesian variable selection, Statistica Sinica 7, 339 - 373.

Hans, C. (2005), Regression model search and uncertainty with many predictors, PhD thesis, Duke University.

Hans, C., A. Dobra, M. West (2007), Shotgun stochastic search for "large p" regression, Journal of the American Statistical Association 102, 507-516.

Hoeting, J., D. Madigan, A. Raftery, C. Volinsky (1999), Bayesian Model Averaging, Statistical Science 14, 382-401.

Jacobson, T., S. Karlsson (2004), Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach, Journal of Forecasting 23, 479-496.

Koop, G. (2003), Bayesian Econometrics, Wiley & Sons.

Ley, E., M. Steel (2007), Jointness in Bayesian Variable Selection with Applications to Growth Regression, Journal of Macroeconomics 29, 476-493.

Raftery, A.E., D. Madigan, J.A. Hoeting (1997), Bayesian Model Averaging for Linear Regression Models, Journal of the American Statistical Association 92, 179-191.

Scharnagl, M., C. Schumacher (2007), Reconsidering the role of monetary indicators for euro area inflation from a Bayesian perspective, Deutsche Bundesbank Discussion Paper, Series 1: Economic Studies, No. 09/07.

Stock, J.H., M.W. Watson (1999), Forecasting Inflation, Journal of Monetary Economics 44, 293-335.

Stock, J.H., M.W. Watson (2002), Macroeconomic Forecasting Using Diffusion Indexes, Journal of Business & Economic Statistics 20, 147-162.