BAYESIAN ASSESSMENT OF DIMENSIONALITY IN REDUCED RANK REGRESSION

JUKKA CORANDER AND MATTIAS VILLANI

ABSTRACT. We consider Bayesian inference about the dimensionality in the multivariate reduced rank regression framework, which encompasses several models such as MANOVA, factor analysis and cointegration models for multiple time series. The fractional Bayes approach is used to derive a closed form approximation to the posterior distribution of the dimensionality and some asymptotic properties of the approximation are proved. Finite sample properties are studied by simulation and the method is applied to growth curve data and cointegrated multivariate time series.

1. INTRODUCTION

A common situation in multivariate analysis involves exploration of the relationships between sets of variables, either by explicit parametric models or by descriptive methods such as principal components and canonical correlations. Although it was early understood that these instances may be jointly represented in terms of multivariate regression obeying a so called reduced rank structure for certain parameters (see, *e.g.*, the pioneering work by Anderson, 1951), such an approach has only recently been fully appreciated by the general statistical community.

An essential strength of the reduced rank regression (RRR) framework is its generality, as it encompasses several well-known models such that MANOVA, factor analysis, linear simultaneous equations models and many other models for multiple time series. For a thorough treatment concerning the time series models, see Ahn and Reinsel (1990), Geweke (1996), Johansen (1995), Velu, Reinsel and Wichern (1986) and for the others, see Anderson (1984, 1994), and the references therein. An excellent review of various issues may also be found in Reinsel and Velu (1998).

The typical model uncertainty in regular full rank multivariate regression is about the choice of relevant predictor variables. Several reasonable solutions are available for this latter model choice problem, see, *e.g.*, Brown et al. (1999) or George and Foster (2000). It has been more of a challenge to produce sensible

Key words and phrases. Cointegration, Fractional Bayes, Growth curves, Multivariate Regression.

We thank an associate editor and two referees for helpful comments. The first author is grateful for financial support from the Academy of Finland, grant no. 50203. The second author gratefully acknowledges financial support from the Swedish Council of Research in Humanities and Social Sciences (HSFR), grant no. F0582/1999 and the Swedish Research Council (Vetenskapsrådet) grant no. 412-2002-1007.

inferences on the dimensionality of the subspace of regression coefficients for a *fixed* set of predictor variables. Geweke (1996) proposes a computationally demanding procedure to assessing dimensionality in RRR and Kleibergen and Paap (2002) use elaborate Monte Carlo simulation schemes to obtain the posterior distribution of the dimensionality.

The methods that tend to be used in applications, however, are the readily computable one-formula solutions without subjective input from the user, such as information theoretic criteria (Akaike, 1974), approximate logarithmic Bayes factor (Schwarz, 1978) or sequential tests (Anderson, 1951; Izenman, 1980; Johansen, 1995). There have also been some proposals within a specific class of reduced rank models, see *e.g.* Chao and Phillips (1999) for a criterion tailored to cointegration models. Among these methods, only the approach of Schwarz (1978) aims at approximating the posterior distribution of the dimensionality. This is an important point as the posterior distribution forms an attractive representation of the uncertainty in the dimensionality inference. The Schwarz approximation is known to be rather rough, however, and it underestimates the underlying model dimension in many instances (Kass and Raftery, 1995).

We utilize the fractional marginal likelihood (FML) approach of O'Hagan (1995, 1997) to derive the approximate posterior distribution of the dimension of the RRR parameter structure. Our method provides an analytically tractable solution, and is applicable without subjective input from the user. Its properties are investigated both theoretically and by applying it to several real and simulated data sets.

The present paper is structured as follows. In the next section we define formally the RRR model and in Section 3 we derive the posterior distribution of the dimensionality. In Section 4 and 5 the proposed procedure is illustrated on both real and simulated data. Some concluding remarks are given in the final section.

2. Reduced rank regression

Consider the following multivariate regression

$$y_i = \Pi x_i + \Gamma z_i + \varepsilon_i, \tag{1}$$

where x_i and z_i are q_1 and q_2-dimensional vectors of predictors, respectively. Π and Γ are $p \times q_1$ and $p \times q_2$ coefficient matrices and $\varepsilon_i \overset{iid}{\sim} N_p(0, \Sigma)$, $i = 1, ..., n$, are the model errors.

The reduced rank regression, introduced by Anderson (1951), allows for the possibility that Π has less than full rank, and the rank of Π will be termed the *regression dimensionality*. The other coefficient matrix Γ, may also be rank deficient (Velu, 1991), but we shall here retain the assumption of a full rank for Γ; see also Remark 2 below, however.

A reduced rank r of Π can be modelled explicitly by writing $\Pi = \Psi\Lambda$, where Ψ and Λ are $p \times r$ and $r \times q_1$ matrices, respectively. Inserting this decomposition into (1) gives

$$y_i = \Psi \Lambda x_i + \Gamma z_i + \varepsilon_i, \tag{2}$$

where Λx_i may now be considered as a new set of r predictors with regression coefficient matrix Ψ. One would expect that Λ is related to the canonical vectors between y_i and x_i, corrected for z_i, and that Ψ is related to the corresponding (partial) canonical correlations, and this is indeed the case (Reinsel and Velu, 1998). In the sequel, we let $r_{\max} = \min(p, q_1)$ denote the maximal rank of Π.

The reduced rank regression in (2) encompasses many of the most widely used multivariate models as special cases, e.g. the common factor model, MANOVA, growth curve models and the cointegrated vector autoregressive model in multivariate time series analysis; Reinsel and Velu (1998) reviews some of the models from a reduced rank perspective.

Two different assumptions will be entertained for the asymptotic behavior of the predictors in z and x. In the first case it is assumed that $n^{-1}\sum_{i=1}^n z_i z_i'$, $n^{-1}\sum_{i=1}^n x_i x_i'$ and $n^{-1}\sum_{i=1}^n x_i z_i'$ all converge in probability to non-stochastic matrices and that

$$\begin{pmatrix} \operatorname{plim} n^{-1}\sum_{i=1}^n x_i x_i' & \operatorname{plim} n^{-1}\sum_{i=1}^n z_i x_i' \\ \operatorname{plim} n^{-1}\sum_{i=1}^n x_i z_i' & \operatorname{plim} n^{-1}\sum_{i=1}^n z_i z_i' \end{pmatrix}$$

is positive definite; such data will be referred to as *stationary* and includes all of the above mentioned models except the cointegration models. To handle the partially non-stationary behavior in cointegration models we need to make an alternative set of assumptions which assumes the same asymptotic behavior for z, but different convergence rates of x in the directions determined by are Λ and its orthogonal complement, $\Lambda_\perp$. More precisely, $n^{-1}\sum_{i=1}^n (\Lambda x_i)(\Lambda x_i)'$ is assumed to converge in probability to a non-stochastic matrix, whereas $n^{-2}\sum_{i=1}^n (\Lambda_\perp x_i)(\Lambda_\perp x_i)'$ is assumed to converge weakly to a stochastic matrix (see Johansen (1995) for more details); such data will be referred to as *cointegrated*. Our asymptotic result will be shown to hold for both stationary and non-stationary data.

As $\Psi U U^{-1} \Lambda = \Psi \Lambda$, for any $r \times r$ non-singular matrix U, the model (2) is not identified. In this paper we impose exactly identifying restrictions on Λ, see *e.g.* Johansen (1995), for a general class of restrictions. It will be evident that our proposed inference tool is invariant to the choice of restrictions.

3. Posterior distribution of the regression dimensionality

The question of dimensionality in RRR is essentially a model inference problem. Given a class of available models, the Bayesian paradigm provides in principle a solution to any model assessment problem, by quantifying the model uncertainty in terms of posterior probabilities (see Bernardo and Smith, 1994), and has been successfully applied to a wide range of problems; for a general review, see Robert and Casella (1999). More specifically, let $M_1, ..., M_h \in \mathcal{M}$ denote the models under comparison, x the available data, $L_i(x|\theta_i)$ the likelihood given M_i and $c_i \pi_i(\theta_i)$ the prior given M_i, where c_i is a normalizing constant. The posterior distribution over $\mathcal{M}$ is then given by

$$p(M_i|x) \propto m_x(M_i)p(M_i)$$

where $m_x(M_i) = c_i \int L_i(x|\theta_i)\pi_i(\theta_i)d\theta_i$ is the marginal likelihood of model M_i and $p(M_i)$ is the prior probability of the ith model.

The non-linear structure of the RRR model makes it far from straight-forward to elicit proper prior distributions on the model parameters. Even if that is managed, the integrals in the marginal likelihood have to be computed numerically. This is feasible with modern technology (see *e.g.* Kleibergen and Paap (2002)), but such computations still seem to deter practitioners from using Bayesian methods.

Here we shall aim at a closed form expression for the posterior distribution of the rank by using improper priors. It is well-known that due to fact that the normalizing constant c_i is arbitrary for improper priors, the usual improper priors cannot be directly used for model comparison, see *e.g.* O'Hagan (1995). One way to confront this problem is to use a small part of the data, usually referred to as the training sample, to convert the improper prior into a proper posterior and subsequently use this updated distribution as a prior for the remaining post-training observations. More precisely, we partition the data x into two sets x_1 and x_2, where x_1 contains the m observations used for training and x_2 contains the remaining $n - m$ post-training observations. The *partial marginal likelihood* (O'Hagan, 1995) is defined as

$$m_{x_2|x_1}(M_i) = \frac{m_x(M_i)}{m_{x_1}(M_i)} = \frac{\int L_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\int L_i(x_1|\theta_i)\pi_i(\theta_i)d\theta_i}, \tag{3}$$

where it should be noted that the arbitrary normalizing constant c_i in the prior cancels as a result of the conditioning on the training sample.

The fundamental difficulty with the partial marginal likelihood is the choice of training sample from the set of $\binom{n}{m}$ possible training samples. O'Hagan (1995) argued that $L_i(x_1|\theta_i)$ in (3) may be approximated by $L_i(x|\theta_i)^b$, where $b = m/n$ is the *training fraction*. Note how this approximation replaces the difficult problem of choosing a particular subset of the data for training with the much easier one of choosing the training fraction. This leads to the *fractional marginal likelihood* (FML)

$$m_b(M_i) = \frac{\int L_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\int L_i(x|\theta_i)^b\pi_i(\theta_i)d\theta_i}. \tag{4}$$

The training fraction is chosen to be *minimal*, *i.e.* $b = m/n$, where m is the smallest possible number of observations necessary to convert $\pi_i(\theta_i)$ into a proper distribution for the *largest* model in $\mathcal{M}$.

Even though the motivation of the FML is asymptotic, O'Hagan (1995) argued that it can be succesfully used in finite samples. This was confirmed in Villani (2001) in the context of lag length inference for multivariate autoregressive processes where the FML was shown to yield consistently more reliable results than the most commonly used lag length selection methods. For further details and properties of the fractional Bayes approach, we refer to O'Hagan (1995, 1997).

Conditional on Λ, the RRR model is a regular full rank multivariate regression. This proves to be a good starting point as the FML of a full rank multivariate regression, stated in the next lemma, is a simple expression. In the following lemma, let

$$S_{xz} = n^{-1} \sum_{i=1}^{n} x_i z_i'.$$

Lemma 1. *The fractional marginal likelihood of the full rank multivariate regression in (1), given the prior* $\pi(\cdot) \propto |\Sigma|^{-(p+1)/2}$*, is*

$$m_b(q_1, q_2) = \left|S_{yy} - S_{yz} S_{zz}^{-1} S_{zy}\right|^{-(n-m)/2} \frac{\Gamma_p[n - (q_1 + q_2)]}{\Gamma_p[m - (q_1 + q_2)]} \prod_{i=1}^{r_{\max}} (1 - \rho_i^2)^{-(n-m)/2} \quad (5)$$

where $\Gamma_p(a) = \Pi_{i=1}^{p} \Gamma[(a - i + 1)/2]$*, for integer* $a \geq p$*,* ρ_i *is the ith largest partial canonical correlation between* y *and* x*, corrected for* z*, and* $m = p + q_1 + q_2$ *is the size of the minimal training sample. A constant which does not depend on the predictors has been discarded in (5).*

Proof. A trivial extension of Theorem 3.1 in Villani (2001) gives

$$m_b(q_1, q_2) \propto \frac{\Gamma_p[n - (q_1 + q_2)]}{\Gamma_p[m - (q_1 + q_2)]} |\hat{\Sigma}|^{-(n-m)/2},$$

where $\hat{\Sigma} = S_{yy} - S_{y(x,z)} S_{(x,z)(x,z)}^{-1} S_{(x,z)y}$ is the ML estimate of Σ in model (1). The lemma now follows from the following well-known result (Johansen, 1995, p. 91-92)

$$|\hat{\Sigma}| = \left|S_{yy} - S_{yz} S_{zz}^{-1} S_{zy}\right| \prod_{i=1}^{r_{\max}} (1 - \rho_i^2).$$

□

If Λ in (2) was known, the model would be an ordinary multivariate regression with predictors Λx_i and z_i, and, consequently, Lemma 1 could be applied directly to compute the FML. With Λ unknown, a naive approach would be to replace Λ by an estimate $\hat{\Lambda}$, and use Lemma 1 with $q_1 = r$. This entirely ignores the uncertainty in Λ is therefore easily seen to favor smaller regression dimensions.

The integrals in (4) with respect to Λ seem intractable for any imaginable prior. Therefore, we suggest an approximate FML of rank r, using (5) with a correction for the degrees of freedom lost in the estimation of Λ, motivated as follows. Notice that $q_1 + q_2$ in (5) equals the number of parameters in each response equation. The number of parameters in each equation of the reduced rank regression depends on the arbitrarily chosen identifying restrictions on Λ and cannot thus be directly used as a correction term. However, as the number of unrestricted parameters in Λ is $(q_1 - r)r$, the average number of parameters in each equation of (2) is $c_r = r + (q_1 - r)(r/p) + q_2$, which is well defined and a natural modification of $q_1 + q_2$ in (5). Note that $c_0 = q_2$ and $c_{r_{\max}} = q_1 + q_2$, which is sensible as the models with $r = 0$ and $r = r_{\max}$ are full rank multivariate regressions with q_2 and

$q_1 + q_2$ predictors, respectively, and Lemma 1 holds exactly. As the first factor of $|\hat{\Sigma}|$ is independent of r, we discard it in the following definition.

Definition 1. The approximate fractional marginal likelihood of rank r is

$$m_b(r) = \frac{\Gamma_p(n - c_r)}{\Gamma_p(m - c_r)} \prod_{j=1}^{r} (1 - \rho_j^2)^{-(n-m)/2}, \tag{6}$$

where ρ_j is the jth largest partial canonical correlation between y_i and x_i corrected for the predictors in z, $c_r = r + (q_1 - r)(r/p) + q_2$, $m = p + q_1 + q_2$ and a constant which does not depend on r has been discarded.

Remark 1. The approximate FML of the rank generalizes immediately to a joint assessment criterion for r and the set of relevant predictors in x and z. If z varies across the models under consideration, the factor $|S_{yy} - S_{yz}S_{zz}^{-1}S_{zy}|$ in Lemma 1 must be retained as it depends on z. The aspects of joint assessment will be illustrated in the numerical examples.

Remark 2. Several generalizations of the RRR model in (2) have been presented in the literature, see Reinsel and Velu (1998) for the relevant references. Velu (1991), for example, allows also Γ to be rank deficient. The way we have defined the approximate FML in Definition 1 makes it straightforward to cover such extensions of the model. The fact that dimensionality inference based on likelihood ratio tests are intractable in many of the extended models makes this point especially important.

As noted above, the approximate FML equals the exact FML for $r = 0$ and $r = r_{\max}$. Further justification of our approximation are given by the following asymptotic properties, established in the next lemma and theorem.

Lemma 2.

$$\log m_b(r) = -\frac{n}{2} \sum_{j=1}^{r} \log(1 - \rho_j^2) - \frac{c_r p \log n}{2} + O_p(1).$$

Proof: As $\sum_{j=1}^{r} \log(1 - \rho_j^2)$ is $O_p(1)$ under both sets of assumptions on the regressors (Anderson, 1984, p.499 for the stationary case and Johansen, 1995, p. 158 for the non-stationary case), the result is a trivial modification of Theorem 3.2 in Villani (2001).

Lemma 2 shows that the approximate FML is asymptotically equivalent to another well known approximation of the marginal likelihood initially derived by Schwarz (1978), usually termed the SBC or the BIC criterion. The SBC criterion has a general tendency to support too small models in finite samples (Kass and Raftery, 1995) and the simulation results in Section 5 suggest that our approximation improves on SBC.

Kim (1998) shows, in a rather general setting, that a first order approximation of the marginal likelihood has the form of a penalized likelihood where the

parameters are penalized according to the convergence rates of their corresponding ML estimators, which for the parameters in the cointegration vectors is n^{-1} instead of the usual $n^{-1/2}$ (see e.g. Johansen, 1995, Theorem 13.3). Kim has termed this approximation the generalized SBC criterion. By simply redefining c_r to $r + 2(q_1 - r)(r/p) + q_2$, the approximate FML in Definition 1 would be asymptotically equal to Kim's generalized SBC criterion. However, the likelihood function is of the same form whether the data generating process is stationary of not (Sims and Uhlig, 1991). Modifing the approximate FML along the lines suggested by Kim (1998) would therefore violate the celebrated likelihood principle, which follows directly from Bayes theorem, but may also be given a non-Bayesian justification (Birnbaum, 1962). This is a strong argument for keeping c_r in its original form and we have chosen to do so in the empirical illustrations in the next section.

It may, however, be of interest to explore certain properties of a Bayesian *procedure* in terms of repeated sampling, keeping in mind that such considerations should *not* enter the inferences for a specific data set; this is particularly true for the type of minimally subjective Bayesian analysis presented here. One such property is that of consistency, the ability to get arbitrarily close to the truth as the sample size grows large.

Theorem 1. *The posterior mode estimator of the regression dimensionality based on the approximate FML in Definition 1 is (weakly) consistent both for stationary and cointegrated data.*

Proof: Let r_0 denote the true dimensionality and $\hat{r}$ the posterior mode estimator of the dimensionality. To show that $\hat{r} \xrightarrow{p} r_0$ we need to show that for all $r \neq r_0$

$$\Pr[m_b(r_0)/m_b(r) > 1] \to 1 \quad \text{as} \quad n \to \infty,$$

which is clearly satisfied if

$$\frac{m_b(r_0)}{m_b(r)} \xrightarrow{p} \infty \quad \text{as } n \to \infty$$

for all $r \neq r_0$. Consider first the case where $r < r_0$. By Lemma 2 we have

$$\log \frac{m_b(r_0)}{m_b(r)} = -\frac{n}{2} \sum_{j=r+1}^{r_0} \log(1 - \rho_j^2) - \frac{(c_{r_0} - c_r) p \log n}{2} + O_p(1). \tag{7}$$

The first term of (7) is $O_p(n)$ and strictly positive for all n as ρ_j^2 $(j = 1, ..., r_0)$ have strictly positive limits (Anderson, 1984, p.499 for the stationary case and Johansen, 1995, p. 158 for the cointegrated case). Thus, the first term of (7) dominates the penalty term of order $\log n$, and $m_b(r_0)/m_b(r) \xrightarrow{p} \infty$ as $n \to \infty$. Now assume that $r > r_0$. We then have

$$\log \frac{m_b(r_0)}{m_b(r)} = \frac{n}{2} \sum_{j=r_0+1}^{r} \log(1 - \rho_j^2) + \frac{(c_r - c_{r_0}) p \log n}{2} + O_p(1).$$

As the $r_{\max} - r_0$ smallest canonical correlations ρ_j^2 $(j = r_0 + 1, ..., r_{\max})$ converge to zero at the rate n^{-1} (Anderson, 1984, p.499 for the stationary case and Johansen, 1995, p. 159 for the cointegrated case), the first term is $O_p(1)$. Thus, the strictly positive penalty term $\frac{(c_r - c_{r_0})p \log n}{2}$ dominates and we again have that $m_b(r_0)/m_b(r) \xrightarrow{p} \infty$ as $n \to \infty$. ■

Remark 1. *The asymptotics of the FML in Lemma 2 and Theorem 1 is the same for stationary and cointegrated data, despite the substantially different asymptotical behavior of sample product moment matrices for these two types of data. The reason is that the FML in Definition 1 only depends on the data through the sample partial canonical correlations, ρ_j $(j = 1, ..., r)$ and that the order of magnitude of the ρ_j are the same for stationary and cointegrated data (see the proof of Theorem 1).*

4. Empirical illustrations

Example 1. *Assessing dimensionality in growth curve analysis.*

Growth curve analysis, or GMANOVA, was introduced by Potthoff and Roy (1964) as a framework for analyzing repeated measurements on a response variable over time. The basic model is of the form

$$y_i = \Psi \Lambda x_i + \varepsilon_i, \tag{8}$$

where x_i is a set of time-invariant predictor variables. The model in (8) is exactly the RRR model without z-predictors. In the original setting of Porthoff and Roy (1964), Ψ was assumed to be a known matrix, completely specified in terms of a parametric function of time, e.g. a polynomial of degree $r - 1$, which determines the general shape of the mean response profiles over time. Here, Λx_i are the coefficients of the parametric function for the ith individual. A natural extension of the model is to let Ψ be any unknown matrix and to infer the dimensionality of Ψ from data as in the RRR model (Reinsel and Velu, 2003). As Reinsel and Velu point out, the columns of Ψ then represent unknown basis functions for the mean response profiles over time.

To illustrate dimensionality determination in the context of growth curve analysis, we use a bioassay data set from Volund (1980), subsequently analyzed in Reinsel and Velu (2003). The response comprises measurements of blood sugar concentration on $n = 36$ rabbits at $1, 2, 3, 4$ and 5 hours after the administration of an insulin dose; the response is thus 5-dimensional. The 36 rabbits were divided in four balanced groups according to a 2×2 design with factors *insulin type* and *dose level*. Four predictors were used: Indicators for insulin type and dose level, initial blood sugar concentration and an interaction variable between dose level and initial concentration. A constant term was also added to the model.

In this and all other examples we use a uniform prior for the regression dimensionality. The other candidate default methods are those widely used in the statistical literature, AIC (Akaike, 1974), SBC (Schwarz, 1978) and the sequential likelihood ratio test (Anderson, 1951).

The posterior distribution of the rank is $p(r|y,x) = .000, .117, .808, .073, .002$ and .001, which clearly indicates that the most likely value of r on the basis of data is 2, although there is a non-negligible degree of uncertainty related to the choice of the rank. As a comparison, the p-values for asymptotic likelihood ratio tests of $H_0 : r = 0, 1, ..., r_{\max} - 1$ against $H_1 : r = p$ are $.000, .022, 0.473, 0.919$ and 0.680, respectively. Thus, 1% and 5% significance levels lead to choices $r = 1$ and $r = 2$, respectively. The other default criteria choose $r = 1$ (SBC) and $r = 2$ (AIC). Using SBC to approximate the posterior distribution gives $p_{SBC}(r|y,x) = .000$, .587, .408, .005, .000 and .000. As expected, the SBC approximation puts more probability mass on smaller r.

Reinsel and Velu (2003) also raise the question of whether any of the predictor variables are redundant for the response. In particular, there was a priori reasons for believing in the absence of an insulin type effect. Let A_j denote the event that the jth predictor in x is redundant. The answer to whether or not a variable is redundant is given by the posterior probability of this event

$$p(A_j|y,x) = \sum_{r=1}^{r_{\max}} p(A_j|r,y,x)p(r|y,x),$$

where $p(A_j|r,y,x)$ is obtained from the posterior distribution over the models with different subsets of predictors conditional on rank r and $p(r|y,x)$ is the posterior distribution of the rank stated above. In the blood sugar data, $p(A_j|y,x) = .000$, .444, .023, .003, .152 for $j = 1, 2, ..., 5$ (first predictor is the constant). Thus, the predictors *insulin type effect* and *interaction between dose level and initial concentration* have non-negligible redundancy probability, whereas *dose level* and *initial concentration* seem to be important predictors, which is in accordance with the results of Reinsel and Velu (2003).

Example 2. *Assessing dimensionality for cointegrated time series.*

Next we consider the case of the cointegrated vector autoregressive (VAR) process

$$\Delta w_t = \Psi \Lambda w_{t-1} + \sum_{i=1}^{k} \Gamma_i \Delta w_{t-i} + \Phi d_t + \varepsilon_t, \qquad t = 1, ..., n, \tag{9}$$

where w_t is an observation on a p-dimensional process at time t, Δ is the time difference operator, d_t is a vector of deterministic trends and $\varepsilon_t \overset{iid}{\sim} N(0, \Sigma)$. The r rows of Λ are the cointegration vectors which determine r linear stationary combinations between the otherwise non-stationary components of w_t. Here, Ψ is a $p \times r$ matrix of adjustment coefficients. The model in (9) can be put in RRR form by defining $y_t = \Delta w_t$, $x_t = w_{t-1}$, $z_t = (w'_{t-1}, ..., w'_{t-k}, d'_t)'$ and $\Gamma = (\Gamma_1, ..., \Gamma_k, \Phi)$.

It has been customary to consider five different forms for the deterministic trends in the cointegrated VAR model, depending on the desired trending behavior of both the original process and the cointegration relations, Λw_{t-1}. The trend models are:

- Model 1: The level data w_t have no deterministic trends and the cointegrating relations Λw_{t-1} do not have intercepts
- Model 2: The level data w_t have no deterministic trends and the cointegrating relations Λw_{t-1} have intercepts
- Model 3: The level data w_t have linear trends and the cointegrating relations Λw_{t-1} have intercepts
- Model 4: The level data w_t have linear trends and the cointegrating relations Λw_{t-1} have linear trends
- Model 5: The level data w_t have quadratic trends and the cointegrating relations Λw_{t-1} have linear trends

The different trend types are obtained by adding either 1 or t, or both, to w_{t-1} and d_t, see Johansen (1995) for a detailed description.

The dimensionality determination in cointegrated processes thus involves a wider variety of quantities: the number of lags in the process ($k = 0, 1, ...$) the number of cointegration relationships ($r = 0, 1, ..., r_{\max}$) and the choice of trend type ($s = 1, 2, ..., 5$).

As an illustration we analyze a small monetary system of Denmark, described in Johansen (1995), which consists of log real money (M2), log real income, the bank deposit rate and a bond rate. Data are in quarterly observations over the period 1974:1 to 1987:3. To keep the range of models at a reasonable level, we restrict the number of lags to be no larger than four. If this upper limit should prove to be too small, the posterior mass would be concentrated on the largest values, and the model class could thereby be extended after the initial analysis.

Note that when either $r = 0$ or $r = r_{\max}$, some of the (k, r, s)-combinations define the same model. To list the equivalent models, we make use of the symbol $\overset{M}{=}$ as an equivalence relation over the set of (k, r, s)-combinations. The following equivalences then holds for all k:

$$(k, r_{\max}, 2) \overset{M}{=} (k, r_{\max}, 3), \quad (k, r_{\max}, 4) \overset{M}{=} (k, r_{\max}, 5),$$
$$(k, 0, s = 1) \overset{M}{=} (k, 0, s = 2) \text{ and } (k, 0, s = 3) \overset{M}{=} (k, 0, s = 4).$$

Removing all redundant models leaves 105 models to be considered and the models with largest posterior probabilities are listed in Table 1.

TABLE 1

THE MOST PROBABLE (k, r, s) COMBINATIONS IN THE DANISH MONETARY DATA.

k	1	1	1	1	1	1	1	1	0	1
r	3	3	2	4	2	1	2	3	4	3
s	3	2	2	2	3	3	4	4	4	5
$p(k, r, s\|y, x, z)$	.106	.093	.091	.060	.059	.055	.054	.049	.040	.038

First, we notice that the posterior of the number of lags in the process is concentrated on $k = 1$, suggesting that there is no need to enlarge the initial class of models in this respect. It is also clear that the uncertainty about cointegration rank is prominent. For comparison, the asymptotic criteria choose the following models: SBC ($k = 0$, $r = 1$, $s = 2$), AIC ($k = 1$, $r = 3$, $s = 3$), which further

illustrates the uncertainty related to the choice of a model. Guided by graphical inspection of the time series, sequential likelihood testing and residual diagnostics, Johansen (1995) decided to use $(k = 1, r = 1, s = 2)$; the so called trace test is on the bordeline of rejecting $r = 0$ in favor of $r = 1$ at the 10% level, whereas the maximal eigenvalue test rejects $r = 0$ in favor of $r = 1$ at the 5% level but is not able to reject $r = 1$ in favor of $r = 2$ at any conventional significance level.

There may be reasons for preferring a certain number of cointegrating relations, perhaps suggested by economic theory. Such considerations are usually dealt with in an ad hoc fashion, sometimes to the extent that the estimated number of relations is nearly ignored. In a Bayesian analysis such prior information is of course easily incorporated via the prior probabilities of the models, $p(M_i)$.

5. Simulation results

A small scale simulation study was conducted to learn more about the properties of the approximate FML in Definition 1. To be able to compare our results with methods used in typical applications, we confine our investigation to point estimates of the dimensionality. The posterior mode estimator is chosen as the Bayesian estimator.

First, we consider models where $p = q_1$, $q_2 = 0$ and $r = 0$ or 1, given as

$$y_i = \Pi x_i + \varepsilon_i,$$

where $x_i \overset{iid}{\sim} N_{q_1}(0, I_{q_1})$. The reduced rank of Π is made explicit by the parameterization

$$\Pi = u\lambda v',$$

where u and v are p-dimensional vectors of unit length and $\lambda \geq 0$ is the single non-zero singular value of Π, *i.e.* $\text{rank}(\Pi) \leq 1$. This parametric setup enables us to systematically investigate the behavior of model assessment criteria as a function of the informativeness of the sample data, which increases for both λ and n. To make the results less dependent of specific choices of parameter values, we let u and v be uniformly distributed on the unit p-sphere, and let the error covariance Σ have an outspread Wishart distribution $W_p(I_p, p)$.

Figure 1 below shows the relative frequencies of choosing $r = 1$ as a function of λ in different settings with $p = 2, 4$ and $n = 25, 50, 100$, over 10.000 simulated configurations. A striking feature of this figure is the bad performance of AIC, especially when data are relatively informative. The posterior mode estimator compares favorably to the other estimators in general and to SBC in particular. This latter point is important as SBC is the only competing approximation of the posterior distribution. The largest difference between posterior mode estimator and other methods appears for $p = 4$ and $n = 25$. The sample size $n = 25$ should in no way be regarded as unimportant from an applied viewpoint, it is probably a fair representation of the uncertainty in many real world applications where the sample size is substantially larger but the data are much less tidy than those resulting from our generating model.

INSERT FIGURE HERE

FIGURE 1. Relative frequency of choice $r = 1$ in 10.000 simulations as a function of λ for three samples sizes, $n = 25, 50, 100$. Graphs in the first row correspond to the model with $p = 2$ and graphs in the second row to $p = 4$. FML $(- \bullet -)$, SBC $(-\Box-)$, LRT $\alpha = .01$ (—), LRT $\alpha = .05$ $(\cdots)$ and AIC (- - -).

INSERT FIGURE HERE

FIGURE 2. Relative frequency of choice $r = 2$ in 5.000 simulations as a function of the sample size. FML $(- \bullet -)$, SBC $(-\Box-)$, LRT $\alpha = .01$ (—), LRT $\alpha = .05$ $(\cdots)$ and AIC (- - -).

To illustrate the rank assessment in more detail for a specific parameter setting, as a function of n, we consider an RRR model with $p = 3, q_1 = 4, q_2 = 0$ and $r = 2$, specified by

$$y_i = \begin{pmatrix} 1 & 0 \\ .5 & 1 \\ .5 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 & -1 \\ 0 & 1 & 0 & -1 \end{pmatrix} + \varepsilon_i$$

where the entries of x_i are independent standard normal variates and the error terms follow the $N(0, \Sigma)$ distribution with

$$\Sigma = \begin{pmatrix} 3 & \cdot & \cdot \\ 2 & 3 & \cdot \\ .5 & 1 & 1 \end{pmatrix}$$

Here we adopt a similar strategy as O'Hagan (1995, p. 108). Data sets of size 200 observations were generated from the model and the rank was determined using $m = 20, ..., 200$ first observations. The rank assessment process was replicated 5.000 times, to estimate the (relative frequency) probabilities of different rank choices. The relative frequency of choosing the true rank for $m = 20, ..., 70$ is shown in Figure 2. This range of sample sizes captures the essential differences between the methods, while keeping details in the plot distinguishable.

We see here that, for small sample sizes, the sequential likelihood ratio (LR) tests do not choose the right dimensionality very often and as n increases, their error probabilities approach the chosen significance level α. AIC yields slightly higher probabilities of choosing the true rank than the posterior mode estimator for the smallest sample sizes, however, it overestimates r with relatively high probability as n grows. For instance, AIC gives probability .155 to the choice $r = 3$, even for $n = 200$ (not shown in the figure). The relative frequency of $r = 2$ for the posterior mode estimator and SBC was very close to one for $n = 200$, thus confirming their consistency.

6. Concluding remarks

We have investigated Bayesian dimensionality assessment in reduced rank regression in the default sense, where a priori information about parameters is not easily expressible in a probabilistic form. For instance, it may be difficult to elicit the priors due to the intrinsic complexity of the RRR model parameters. The proposed approximate Bayes approach seems to perform reasonably well in a wide variety of situations, and our results suggest that the solutions provided by some of the earlier established asymptotic criteria should be considered with cautiousness.

We want to stress the importance of a proper quantification of the uncertainty in the rank inference. A Bayesian approach, approximate or not, delivers, through the posterior distribution over the regression dimensionality, the relevant information to be used in averaging further analyses, *e.g.* prediction exercises or parametric hypothesis testing, over the model space.

References

[1] Ahn, S. K., and G. C. Reinsel (1990), Estimation for partially nonstationary multivariate autoregressive models, *Journal of the American Statistical Association* **85**, 813-23.

[2] Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**, 716-727.

[3] Anderson, T. W. (1951), Estimating linear restriction on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statistics* **22**, 327-51.

[4] Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), Wiley, New York.

[5] Anderson, T. W. (1994), Inference in linear models, in: T. W. Anderson, K. T. Fang and L. Olkin (eds.), *Multivariate Analysis and Its Applications*, IMS Lecture Notes - Monograph Series 24, 1-20.

[6] Bernardo, J. M. and A. F. M. Smith (1994), *Bayesian theory*, Wiley, New York.

[7] Birnbaum, A. (1962), On the foundations of statistical inference (with discussion), *Journal of the American Statistical Association* **57**, 269-306.

[8] Brown, P. J., T. Fearn and M. Vannucci (1999), The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach, *Biometrika* **86**, 635-648.

[9] Chao, J. C. and P. C. B. Phillips (1999), Model selection in partially nonstationary vector autoregressive processes with reduced rank structure, *Journal of Econometrics* **91**, 227-271.

[10] George, E. I. and D. P. Foster (2000), Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731-747.

[11] Geweke, J. (1996), Bayesian reduced rank regression in econometrics, *Journal of Econometrics* **75**, 121-146.

[12] Izenman, A. J. (1980), Assesing dimensionality in multivariate regression, in: P. R. Krishnaiah (ed.), *Handbook of Statistics* 1, North-Holland, Amsterdam, 571-591.

[13] Johansen , S. (1995), *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, New York.

[14] Kass, R. E. and A. E. Raftery (1995), Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

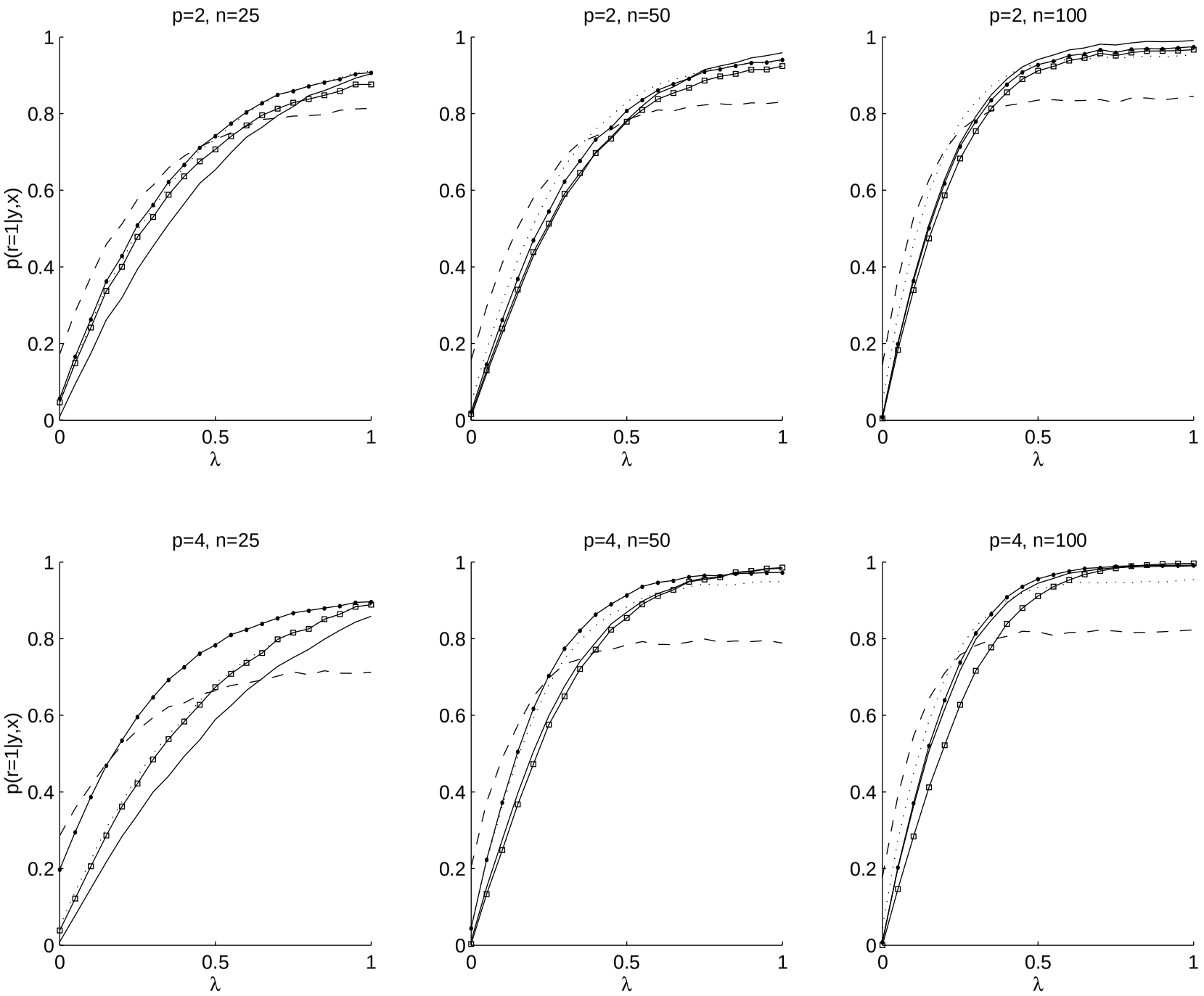
[15] Kim, J. Y. (1998), Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models, *Econometrica* **66**, 359-380.

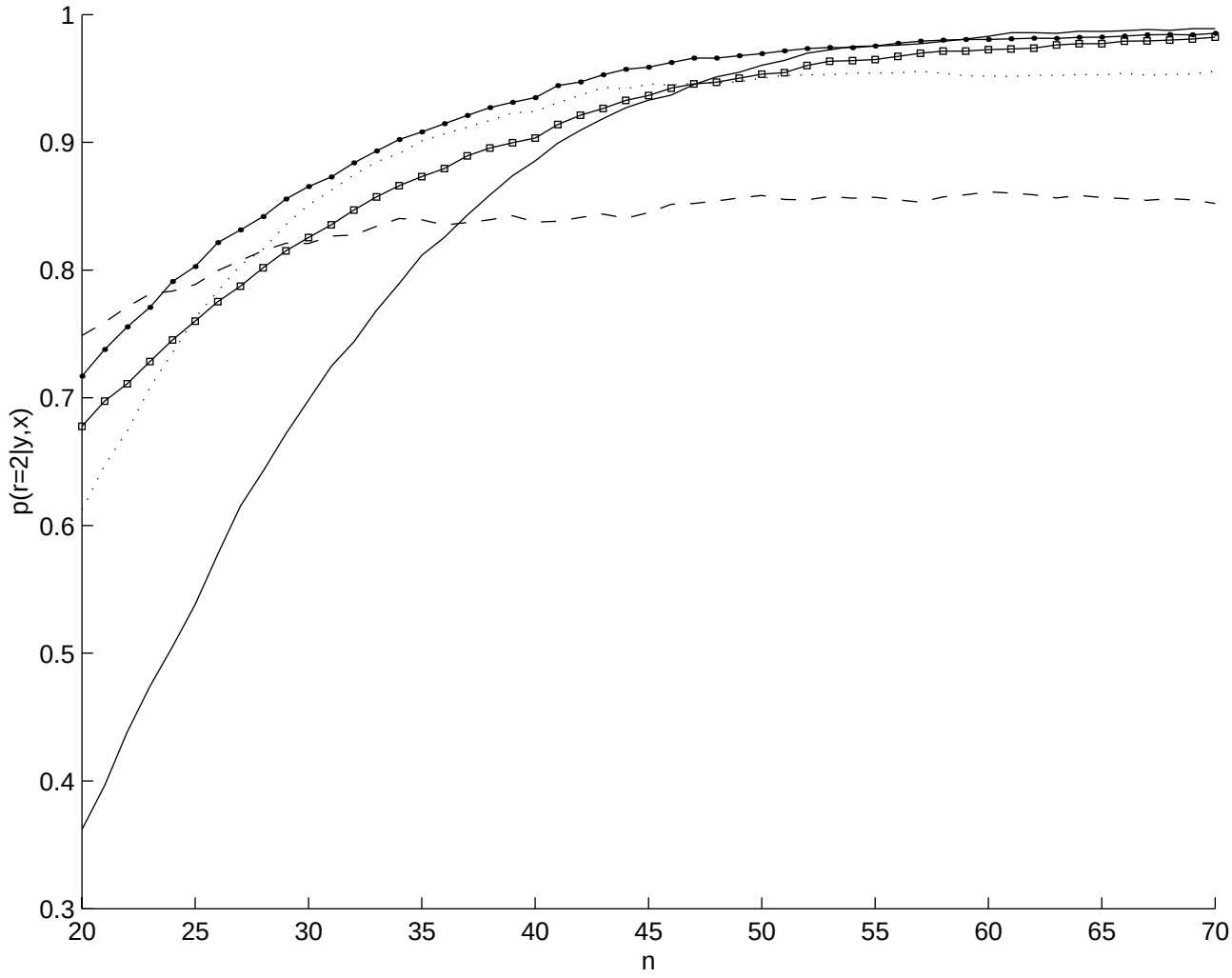
[16] Kleibergen, F. and R. Paap (2002), Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration, *Journal of Econometrics* **111**, 223-249.

[17] O'Hagan, A. (1995), Fractional Bayes factors for model comparisons. *Journal of the Royal Statistical Society* **B 57**, 99-138.
[18] O'Hagan, A. (1997), Properties of intrinsic and fractional Bayes factors, *Test* **6**, 101-118.
[19] Potthoff, R. F. and S. N. Roy (1964), A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, **51**, 313-26.
[20] Reinsel, G. C. and R. P. Velu (1998), *Multivariate reduced-rank regression: Theory and applications*. Springer-Verlag, New York.
[21] Reinsel, G. C. and R. P. Velu (2003), Reduced-rank growth curve models, *Journal of Statistical Planning and Inference* **114**, 107-129.
[22] Robert, C. P. and G. Casella (1999), *Monte Carlo statistical methods*, Springer-Verlag, New York.
[23] Schwarz, G. (1978), Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.
[24] Sims, C. A. and H. Uhlig (1991), Understanding unit rooters: a helicopter tour. *Econometrica* **59**, 1591-1599.
[25] Velu, R. P. (1991), Reduced rank models with two sets of regressors, *Applied Statistics* **40**, 159-170.
[26] Velu, R. P., G. C. Reinsel and D. W. Wichern (1986), Reduced rank models for multiple time series, *Biometrika* **73**, 105-118.
[27] Villani, M. (2001), Fractional Bayesian Lag Length Inference in Multivariate Autoregressive Processes, *Journal of Time Series Analysis* **22**, 67-86.
[28] Volund, A. (1980), Multivariate bioassay, *Biometrics* **36**, 225-236.

Rolf Nevanlinna Institute, P.O. Box 4, Fin-00014, Finland
E-mail address: jukka.corander@rni.helsinki.fi

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden and Research Department, Sveriges Riksbank, SE-103 37 Stockholm, Sweden
E-mail address: mattias.villani@riksbank.se


p=2, n=25
p=2, n=50
p=2, n=100
p=4, n=25
p=4, n=50
p=4, n=100
p(r=1|y,x)
λ
0
0.2
0.4
0.6
0.8
1
0.5


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
p(r=2|y,x)
20
25
30
35
40
45
50
55
60
65
70
n