Selection of latent variables based on grouped Bayes factors

Gonzalo García-Donato, María Eugenia Castellanos and Carolina Mulet
UCLM/URJC/URJC

15^{th} OBayes conference.
Athens, June 2025


Volume 34 • Number 1 • March 2025
TEST
Editors
Gonzalo García-Donato
María del Carmen Pardo
Associate Editors
Ana M. Aguilera
José M. Angulo
Eustasio del Barrio
Enrique del Castillo
Jacobo de Uña Álvarez
Manuel Febrero
Alba María Franco
Luis Ángel García Escudero
Virgilio Gómez Rubio
Wenceslao González Manteiga
Alfonso Gordaliza
Zijian Guo
María Dolores Jiménez Gamero
Ivan Kojadinovic
Rosa E. Lillo
Jesús López Fidalgo
Manuel Molina
Domingo Morale
Pepa Ramírez Cobo
Juan Eloy Ruiz Castro
María Dolores Ruiz Medina
Stefan Sperlich
Lola Ugarte Martínez
Carlos Velasco
Cen Wu
Kostas Zografos
An Official Journal of the Spanish Society of Statistics and Operations Research
SEIO.ES
Abstracted/Indexed in:
ANVUR, BFI List, CLOCKSS, CNKI, CNPIEC, Current Index to Statistics, Dimensions, EBSCO Discovery Service, EconLit, Google Scholar, ICYT, Japanese Science and Technology Agency (JST), Journal Citation Reports/Science Edition, Mathematical Reviews, Naver, OCLC WorldCat Discovery Service, Portico, ProQuest ABI/INFORM, ProQuest Advanced Technologies & Aerospace Database, ProQuest-ExLibris Primo, ProQuest-ExLibris Summon, Research Papers in Economics (RePEc), SCImago, SCOPUS, Science Citation Index Expanded (SCIE), TD Net Discovery Service, UGC-CARE List (India), zbMATH

Selection of latent variables based on grouped Bayes factors

1. Variable selection based on Bayes factors

2. Latent variables

3. The default approach

4. Grouped Bayes factors

5. Results

6. Future research

1. Variable selection based on Bayes factors

2. Latent variables

3. The default approach

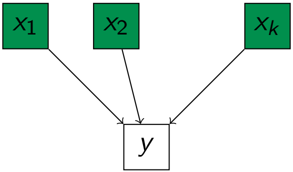
4. Grouped Bayes factors

5. Results

6. Future research

Variable selection and model uncertainty

• In variable selection we must decide which of the individual variables $x_1, \ldots, x_k$ are relevant to explain y.



• Variable selection can be embedded in a **Model Uncertainty** framework under which, each combination of variables defines a different model (a total of 2^k).

• The competing models can be compactly expressed using $\boldsymbol{\gamma} \in \{0,1\}^k$ and

$$M_\gamma : f_\gamma(\boldsymbol{y} \mid \gamma_1\beta_1\boldsymbol{x}_1, \ldots, \gamma_k\beta_k\boldsymbol{x}_k, \boldsymbol{\nu}).$$

• We denote $|\boldsymbol{\gamma}| = \sum_{i=1}^k \gamma_i$.

Posterior probabilities: Bayes factors, priors and summaries

- Posterior probabilities are obtained as

$$P(M_\gamma \mid \boldsymbol{y}) \propto B_{\boldsymbol{\gamma}}\, P(M_\gamma),\ \gamma \in \{0,1\}^k$$

- $B_{\boldsymbol{\gamma}}$ is the Bayes factor of M_γ to M_0. It is univocally defined by f_γ and the priors π_γ (Robust, hyper-g, independent, Held's priors, ...).

Posterior probabilities: Bayes factors, priors and summaries

- Posterior probabilities are obtained as

$$P(M_\gamma \mid \boldsymbol{y}) \propto B_{\boldsymbol{\gamma}}\, P(M_\gamma),\ \gamma \in \{0,1\}^k$$

- $B_{\boldsymbol{\gamma}}$ is the Bayes factor of M_γ to M_0. It is univocally defined by f_γ and the priors π_γ (Robust, hyper-g, independent, Held's priors, ...).

Bayes factors' idiosyncrasy

Bayes factors share a common behaviour (based on asymptotics)

$$B_{\boldsymbol{\gamma}} \approx e^{\Lambda_\gamma/2} \times n^{-|\gamma|/2} \times C(\pi_\gamma, \pi_0),$$

- B_γ **rewards fit** via Λ_γ, the deviance of M_γ relative to M_0,
- B_γ **penalizes complexity** via $n^{-|\gamma|/2}$, where $|\gamma|$ is the number of ones in γ.

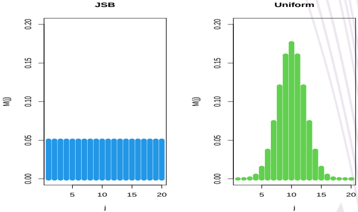
For LM and g-priors: $B_{\boldsymbol{\gamma}} = \left(\frac{1+n}{1+nQ_\gamma}\right)^{(n-k_0)/2} (1+n)^{-|\gamma|/2}$

• $P(M_\gamma)$ is the prior probability of M_γ. We normally use priors that only depend on $|\boldsymbol{\gamma}|$

$$P(M_\gamma) = \frac{\mathfrak{M}(|\boldsymbol{\gamma}|)}{\binom{k}{|\boldsymbol{\gamma}|}}, \ \mathfrak{M}(\cdot) \text{ is a p.m.f. over } \{0, 1, \ldots, k\}.$$

Alternatives:

- Jeffreys-Scott-Berger: $\mathfrak{M}(j) = \frac{1}{k+1}$,
- Uniform: $\mathfrak{M}(j) = \binom{k}{j}/2^k$.



• Regarding summaries, **posterior inclusion probabilities** are very popular

Report relevance of x_i through: $$P(\gamma_i = 1 \mid \boldsymbol{y}) = \sum_{\gamma:\gamma_i=1} P(M_\gamma \mid \boldsymbol{y}).$$

These give rise to the Median Probability Model (MPM) [Barbieri and Berger, 2004]

1. Variable selection based on Bayes factors

2. Latent variables

3. The default approach

4. Grouped Bayes factors

5. Results

6. Future research

Latent variables. Approaching the concept through examples

Latent variables. Approaching the concept through examples

- How to assess the relevance of **cultural level** in an economic study? By means of number of bookshops; number of cinemas; number of theaters; number of science museums; number of modern art museums; etc.

Latent variables. Approaching the concept through examples

- How to assess the relevance of **cultural level** in an economic study? By means of number of bookshops; number of cinemas; number of theaters; number of science museums; number of modern art museums; etc.
- How to analyse the importance of **sedentarism** in an epidemiological study?
  Through the number of hours using the mobile; number of hours with videogames; frequency of usage of mobile phone and number of hours playing sports; etc.

Latent variables. Approaching the concept through examples

- How to assess the relevance of **cultural level** in an economic study? By means of number of bookshops; number of cinemas; number of theaters; number of science museums; number of modern art museums; etc.
- How to analyse the importance of **sedentarism** in an epidemiological study?
  Through the number of hours using the mobile; number of hours with videogames; frequency of usage of mobile phone and number of hours playing sports; etc.

Latent variable Z. (A name we freely borrow from multivariate statistics)

A hypothetical (not directly observable) construct whose relevance in a study can only be assessed indirectly, by means of a set of observable **indicator variables** $z_1, \ldots, z_\ell$ collected by the researcher.

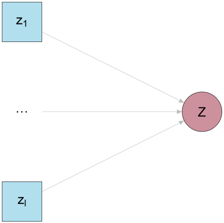
Latent variables. Approaching the concept through examples

- How to assess the relevance of **cultural level** in an economic study? By means of number of bookshops; number of cinemas; number of theaters; number of science museums; number of modern art museums; etc.
- How to analyse the importance of **sedentarism** in an epidemiological study?
  Through the number of hours using the mobile; number of hours with videogames; frequency of usage of mobile phone and number of hours playing sports; etc.

Latent variable Z. (A name we freely borrow from multivariate statistics)

A hypothetical (not directly observable) construct whose relevance in a study can only be assessed indirectly, by means of a set of observable **indicator variables** $z_1, \ldots, z_\ell$ collected by the researcher.

Latent variables

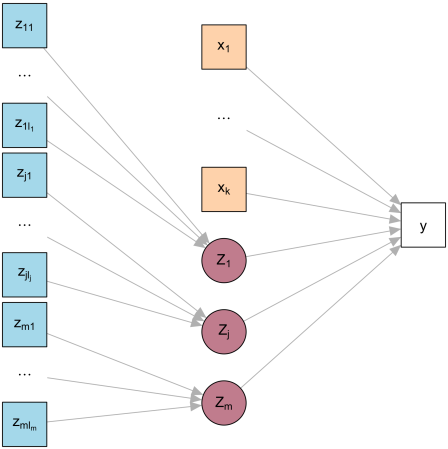


Key highlights:

- A latent variable tries to capture a concept with a name.
- The number of indicators ℓ is rather arbitrary.
- Indicators are expected to be correlated.

The goal in this research

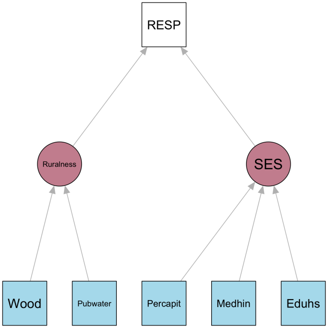
To develop methodology for latent (Z) and individual (x) variable selection from a Bayesian Model Uncertainty perspective.



Goal: which of $x_1, \dots, x_k$ and/or $Z_1, \dots, Z_m$ are relevant to explain y?

An epidemiological illustration: [Wall and Li, 2003]

In $n = 87$ Minessota counties, relation of mortality due to respiratory diseases **RESP** with **ruralness** and social economic status **SES**



Indicators:

- For **ruralness**
  - pubwater: Per cent of population with access to public water,
  - wood: Per cent of population using wood to heat the home.
- For **SES**
  - eduhs: per cent with high-school education,
  - medhhin: Median household income (in dollars),
  - percapit: Per capita income (in dollars).

1. Variable selection based on Bayes factors

2. Latent variables

3. The default approach

4. Grouped Bayes factors

5. Results

6. Future research

The default approach

- Solve the standard variable selection problem with **all** regressors (model space with $2^{k+\ell_1+\cdots+\ell_m}$).

The default approach

- Solve the standard variable selection problem with **all** regressors (model space with $2^{k+\ell_1+\cdots+\ell_m}$).
- Obtain posterior inclusion probabilities,

The default approach

- Solve the standard variable selection problem with **all** regressors (model space with $2^{k+\ell_1+\cdots+\ell_m}$).
- Obtain posterior inclusion probabilities,
- Infer about the relevance of Z_j using the maximum of inclusion probabilities of its indicators.

The default approach

- Solve the standard variable selection problem with **all** regressors (model space with $2^{k+\ell_1+\cdots+\ell_m}$).
- Obtain posterior inclusion probabilities,
- Infer about the relevance of Z_j using the maximum of inclusion probabilities of its indicators.

Latent	**SES**			**ruralness**	
Indicator	Eduhs	Medhin	Percapit	Pubwater	Wood
$\pi(\gamma_i = 1 \mid \boldsymbol{y})$	0.36	0.39	0.32	0.14	0.97

Table: Posterior Inclusion probabilities in Wall and Li example

Conclusion

ruralness is a relevant latent variable to explain **RESP** while **SES** is not.

Questions about the default approach

- Is this a sensible approach?

Questions about the default approach

- Is this a sensible approach?
- Which is the role of $\ell_1, \ldots, \ell_m$?

Questions about the default approach

- Is this a sensible approach?
- Which is the role of $\ell_1, \ldots, \ell_m$?
- What is the importance of the dependence structure?

We addressed these questions using a simulated experiment.

Simulation scheme: Data generative model

- Four latent variables, Z_1, Z_2, Z_3, Z_4 that are a linear combination of a large number ($\ell_* = 50$) of zero mean multivariate normal indicators

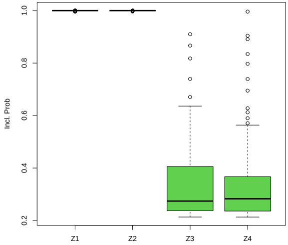
$$Z_j = \frac{1}{\sqrt{V(\sum_{h=1}^{50} z_{jh})}} \sum_{h=1}^{50} z_{jh}, \quad j = 1, 2, 3, 4,$$

- The indicators forming each latent have correlations: $\rho_1 = \rho_3 = 0.9$, $\rho_2 = \rho_4 = 0.6$.
- The data generative model is

$$y = 1 * Z_1 + 1 * Z_2 + 0 * Z_3 + 0 * Z_4 + N(0, 1) \quad (n = 50)$$

- 100 datasets were generated this way

Simulation scheme: Oracle results



	Z_1	Z_2	Z_3	Z_4
ρ_j	0.9	0.6	0.9	0.6
β_j	1	1	0	0

Table: Details of Data Generative process

Simulation scheme: emulating a real situation and performance of default approach

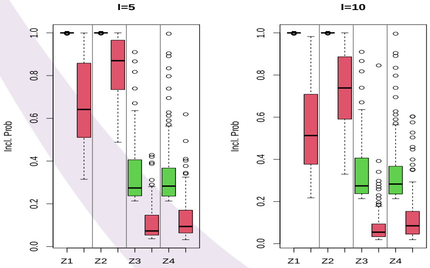
As in a real situation:

- We only have access to a certain number of indicators per each latent $\ell \in \{5, 10\}$.

Simulation scheme: emulating a real situation and performance of default approach

As in a real situation:

- We only have access to a certain number of indicators per each latent $\ell \in \{5, 10\}$.
- Perform the default approach, retaining the maximum of the posterior inclusion probabilities of the indicators.



- The default approach severely dilutes the strength of true signals. Worst when correlation is high.
- This effect is amplified when ℓ increases (which is highly unsatisfactory).
- It underestimates the probability of false signals.

Everything goes wrong with the default approach

These observations are a clear manifestation of the [Barbieri et al., 2021] [BBGR] paper:

BBGR: highly correlated variables fail to cooperate

The Median Probability Model may well not include any covariates that are highly correlated. (Inclusion probabilities are low)

1. Variable selection based on Bayes factors

2. Latent variables

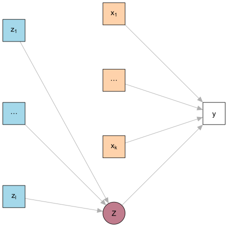
3. The default approach

4. Grouped Bayes factors

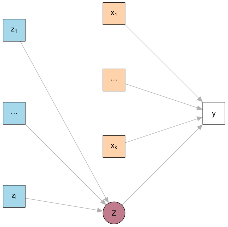
5. Results

6. Future research

The case with one latent variable



The case with one latent variable

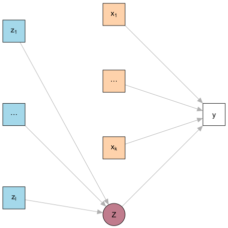


	Individual			Latent (Z)			Cardinality
y	x_1	$\cdots$	x_k	z_1	$\cdots$	z_ℓ	
	γ_1	$\cdots$	γ_k	δ_1	$\cdots$	δ_ℓ	$2^{k+\ell}$
	γ_1	$\cdots$	γ_k		τ		2^{k+1}

Where

- $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$; $\gamma_i = 1$ if x_i is active and zero otherwise.
- Similarly variable z_i is active if $\delta_i = 1$ (and zero otherwise)
- $\tau = 1$ (Z active) if $\delta_i = 1$ for any i.

Recall: each combination $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ defines a model and we have a Bayes factor $B_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\boldsymbol{y})$ for it.



$\mathcal{B}_{\gamma,\tau}(\boldsymbol{y})$ Grouped Bayes factor of Latent active vs. the null

$$\mathcal{B}_{\gamma,1}(\boldsymbol{y}) = \sum_{\boldsymbol{\delta}\in\{0,1\}^{\ell}-(0,\ldots,0)} \pi_Z(\boldsymbol{\delta}) B_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\boldsymbol{y}),$$

$$\mathcal{B}_{\gamma,0}(\boldsymbol{y}) = B_{\gamma,0}(\boldsymbol{y})$$

Notice $\mathcal{B}_{\boldsymbol{\gamma},\tau}(\boldsymbol{y})$

- is an actual Bayes factor.
- is a weighted average of $2^{\ell} - 1$ single Bayes factors.

The question is how to specify the "prior" $\pi_Z(\boldsymbol{\delta})$.

About $\pi_Z(\boldsymbol{\delta})$. Default objective candidates

$$\pi_Z(\boldsymbol{\delta}) = \frac{\mathfrak{M}_Z(|\boldsymbol{\delta}|)}{\binom{\ell}{|\boldsymbol{\delta}|}}$$

Mimicking standard proposals in VS, alternatives are

- Uniform: $\mathfrak{M}_Z(j) = \frac{\binom{\ell}{j}}{2^\ell - 1},\ j = 1, \ldots, \ell.$
- JSB: $\mathfrak{M}_Z(j) = \frac{1}{\ell},\ j = 1, \ldots, \ell.$

The second alternative is the solution in [García-Donato and Paulo, 2022] proposed to handle qualitative variables (factors).

About $\pi_Z(\boldsymbol{\delta})$. Default objective candidates

$$\pi_Z(\boldsymbol{\delta}) = \frac{\mathfrak{M}_Z(|\boldsymbol{\delta}|)}{\binom{\ell}{|\boldsymbol{\delta}|}}$$

Mimicking standard proposals in VS, alternatives are

- Uniform: $\mathfrak{M}_Z(j) = \frac{\binom{\ell}{j}}{2^\ell - 1}$, $j = 1, \ldots, \ell$.
- JSB: $\mathfrak{M}_Z(j) = \frac{1}{\ell}$, $j = 1, \ldots, \ell$.

The second alternative is the solution in [García-Donato and Paulo, 2022] proposed to handle qualitative variables (factors).
We thought this was the solution, but realized of a

strange phenomenon:

About $\pi_Z(\boldsymbol{\delta})$. Default objective candidates

$$\pi_Z(\boldsymbol{\delta}) = \frac{\mathfrak{M}_Z(|\boldsymbol{\delta}|)}{\binom{\ell}{|\boldsymbol{\delta}|}}$$

Mimicking standard proposals in VS, alternatives are

- Uniform: $\mathfrak{M}_Z(j) = \frac{\binom{\ell}{j}}{2^\ell - 1}$, $j = 1, \ldots, \ell$.
- JSB: $\mathfrak{M}_Z(j) = \frac{1}{\ell}$, $j = 1, \ldots, \ell$.

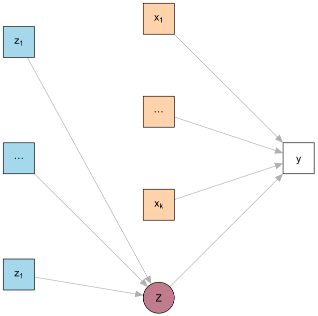
The second alternative is the solution in [García-Donato and Paulo, 2022] proposed to handle qualitative variables (factors).
We thought this was the solution, but realized of a

strange phenomenon:

For moderate to highly correlated indicators $\mathcal{B}_{\boldsymbol{\gamma},\tau}(\boldsymbol{y})$ (highly) underestimates the effect of relevant latent variables.

The strange phenomenon under the lens of [Barbieri et al., 2021]

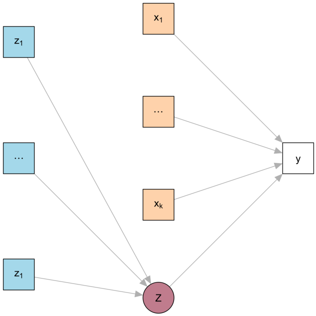
Suppose $z_1, \ldots, z_\ell$ are copies of each other. Then just one (say z_1) encapsulates the behaviour of the latent variable



- We expected: $\mathcal{B}_{\boldsymbol{\gamma},1}(\boldsymbol{y}) \approx B_{\boldsymbol{\gamma},1,0,\ldots,0}(\boldsymbol{y})$,

The strange phenomenon under the lens of [Barbieri et al., 2021]

Suppose $z_1, \dots, z_\ell$ are copies of each other. Then just one (say z_1) encapsulates the behaviour of the latent variable



- We expected: $\mathcal{B}_{\boldsymbol{\gamma},1}(\boldsymbol{y}) \approx B_{\boldsymbol{\gamma},1,0,\dots,0}(\boldsymbol{y})$,
- but instead we got: $\mathcal{B}_{\boldsymbol{\gamma},1}(\boldsymbol{y}) \approx B_{\boldsymbol{\gamma},1,0,\dots,0}(\boldsymbol{y}) \times H(n,\ell)$, where $H(n,\ell) = \sum_{j=1}^{\ell} \frac{\mathfrak{M}_Z(j)}{(1+n)^{(j-1)/2}}$

Table: Values of $H(\ell, n = 30)$.

ℓ	5	10	20
JSB	0.24	0.12	0.06
Uniform	0.23	0.02	10^{-4}.

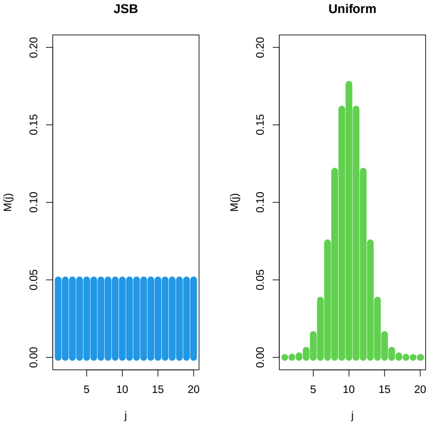
The strange phenomenon

Key message (general for any Bayes factors):

All models that have any indicator active provide a similar fit. But the great majority of those are much more complex than we would need. These models are penalized by complexity due to the idiosincrasy of Bayes factors, dimishing the weighted mean defining the grouped Bayes factor.

The strange phenomenon

The different behaviour of the priors can be easily seen by a simple look at their form:



Notice: models that emulate how Z behaves (in terms of fit) have j small.

Our proposal for $\mathfrak{M}_Z(j)$

Goal:

Define $\mathfrak{M}_Z(j)$ to assign more mass to small j if the indicators $\{z_1, \ldots, z_\ell\}$ are highly correlated.

Our proposal for $\mathfrak{M}_Z(j)$

Goal:

Define $\mathfrak{M}_Z(j)$ to assign more mass to small j if the indicators $\{z_1, \ldots, z_\ell\}$ are highly correlated.

Suppose ρ_Z is a [0,1] measure of the correlation among $\{z_1, \ldots, z_\ell\}$. We have worked with

$$j - 1 \sim \text{Binom}(\ell - 1, p), \quad p \sim \text{Beta}(1, (1 - \rho_Z)^{-1}).$$

but also with something more simple inspired by Principal Components (PC)

$$\mathfrak{M}_Z(j) \propto \text{Proportion of Variance explained using the first } j \text{ PCs}$$

Extreme cases:

- $\mathfrak{M}_Z(j) = 1_{\{1\}}(j)$ if the indicators are copies (and $H(n, \ell) = 1$)
- $\mathfrak{M}_Z(j) = \frac{1}{\ell}$ (JSB) if the indicators are orthogonal.

Are these proposals really objective?

- These possibilities depend on the data, but only through the matrix of regressors (the block corresponding to the indicators).

Are these proposals really objective?

• These possibilities depend on the data, but only through the matrix of regressors (the block corresponding to the indicators).

• Recall that g-priors (and essentially all conventional priors) have a similar dependence on data.

The general case

	Individual covs			Z_1			$\cdots$	Z_m		
y	x_1	$\cdots$	x_k	z_1^1	$\cdots$	$z_{\ell_1}^1$	$\cdots$	z_1^m	$\cdots$	$z_{\ell_m}^m$
	γ_1	$\cdots$	γ_k	δ_{11}	$\cdots$	$\delta_{1\ell_1}$	$\cdots$	δ_{m1}	$\cdots$	$\delta_{m\ell_m}$
				τ_1			$\cdots$	τ_m		

Grouped Bayes factor: (assuming wlog $\tau_1 = 1, \ldots, \tau_r = 1$ and the rest zero)

$$\mathcal{B}_{\gamma,\tau}(\boldsymbol{y}) = \sum_{\delta_1 \in \{0,1\}^{\ell_1} - 0} \cdots \sum_{\delta_r \in \{0,1\}^{\ell_r} - 0} \pi_{(Z_1,\ldots,Z_r)}(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_r)\, B_{\gamma,\delta_1,\ldots,\delta_r,0,\ldots,0}(\boldsymbol{y}),$$

and

$$\pi_{(Z_1,\ldots,Z_r)}(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_r) = \prod_{j=1}^{r} \pi_{Z_j}(\boldsymbol{\delta}_j),$$

and as before, specify $\pi_{Z_j}(\boldsymbol{\delta}_j)$ using the corresponding correlation structure on the indicators defining ech Z_j.
To complete the prior, use over the 2^{m+k} Grouped Bayes factors, the JSB prior

Computation/searching

- It can be easily seen that we are implicitly defining a prior over the complete model space $2^{k+\ell_1+\cdots+\ell_m}$.
- Hence we can apply standard samplings algorithms to obtain the posterior distribution.

Computation/searching

- It can be easily seen that we are implicitly defining a prior over the complete model space $2^{k+\ell_1+\cdots+\ell_m}$.
- Hence we can apply standard samplings algorithms to obtain the posterior distribution.

We are big fans of one of the simplest Gibbs sampling algorithm

- It is highly reliable,
- completely automatic,
- very easy to implement,
- extremely fast implementations are possible for sparse settings,

More details in [Garcia-Donato and Martinez-Beneito, 2013, Garcia-Donato and Castellanos, 2024]

1. Variable selection based on Bayes factors

2. Latent variables

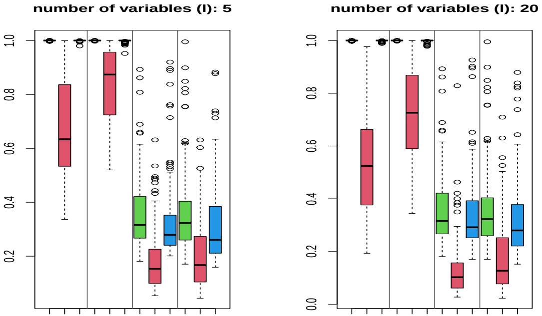
3. The default approach

4. Grouped Bayes factors

5. Results

6. Future research

Results (in the previous simulation scheme)



green:Oracle; Red: default; Blue: Ours (based on PCs).

- The approach based on grouped Bayes factors reconstructs the Oracle (both with true and spurious latent variables).
- Slightly better as the correlation is higher and the number of indicators increases.

Results in the illustrative example

(In gray what we obtained with the default)

	SES			ruralness	
Indicator	Eduhs	Medhin	Percapit	Pubwater	Wood
$P(=1 \mid \boldsymbol{y})$		.92		1	
$P(\gamma_i = 1 \mid \boldsymbol{y})$	0.36	0.39	0.32	0.14	0.97

Table: Posterior Inclusion probabilities.

New conclusion

ruralness and **SES** both are relevant.

Related Literature

There is a rich and recent literature on Bayesian approaches to variable selection with *groups* of variables; a concept which is obviously connected with ours. Nevertheless, researchers have mainly focused on situations with

- dummies,
- basis expansions.

In these situations, the variables forming a group are *structurally* tied (not *conceptually* tied), defining an observable entity (a qualitative variable or a functional) (as opposed to a *theoretical* variable).

Related Literature (cont')

The authors have extended well-established methods and ideas to these scenarios.

- Agarwal et al (24): group informed g-prior,
- Regularization with groups (like group LASSO, etc).
- Grouped spike and slab,
- Group sparsity.

1. Variable selection based on Bayes factors

2. Latent variables

3. The default approach

4. Grouped Bayes factors

5. Results

6. Future research

Short term

- Other examples, other responses, other Bayes factors;
- Very large model spaces (pathway genes);
- Interpretation of the individual inclusion probabilities?
- Connections with other informed prior distribution (dilution priors)

Short term

- Other examples, other responses, other Bayes factors;
- Very large model spaces (pathway genes);
- Interpretation of the individual inclusion probabilities?
- Connections with other informed prior distribution (dilution priors)

Medium term

- Our approach opens the possibility of studying more complex structures and a promising path to Bayesian confirmatory Factor analysis (BCFA).

Short term

- Other examples, other responses, other Bayes factors;
- Very large model spaces (pathway genes);
- Interpretation of the individual inclusion probabilities?
- Connections with other informed prior distribution (dilution priors)

Medium term

- Our approach opens the possibility of studying more complex structures and a promising path to Bayesian confirmatory Factor analysis (BCFA).

Science fiction

- A succesful implementation of the above BCFA would open the possibility of Bayesian methods to exploratory Factor analysis in which the latent variables are to be "discovered".

Thanks!

Funded by

- Grant PID2022-138201NB-100 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.
- Fondo Europeo de Desarrollo Regional SBPLY/21/180501/000241

References I

Barbieri, M. M. and Berger, J. O. (2004).
Optimal predictive model selection.
Annals of Statistics, 32:870–897.

Barbieri, M. M., Berger, J. O., George, E. I., and Ročková, V. (2021).
The Median Probability Model and Correlated Variables.
Bayesian Analysis, 16(4):1085–1112.

Garcia-Donato, G. and Castellanos, M. (2024).
Searching in ultra high dimensional sparse model spaces: New performance tests and boosting gibbs sampling algorithms.
Technical report, Preprint available in Research Square.

Garcia-Donato, G. and Martinez-Beneito, M. A. (2013).
On Sampling strategies in Bayesian variable selection problems with large model spaces.
Journal of the American Statistical Association, 108(501):340–352.

References II

García-Donato, G. and Paulo, R. (2022).
Variable selection in the presence of factors: A model selection perspective.
Journal of the American Statistical Association, 117(540):1847–1857.

Wall, M. M. and Li, R. (2003).
Comparison of multiple regression to two latent variable techniques for estimation and prediction.
Statistics in Medicine, 22:3671–3685.