Model selection for sparse high-dimensional learning

Pierre-Alexandre Mattei

UNIVERSITÉ PARIS DESCARTES

École doctorale de Paris centre

Laboratoire d'accueil : MAP5, UMR CNRS 8145

Sélection de modèles parcimonieux pour l'apprentissage statistique en grande dimension

Par Pierre-Alexandre Mattei

Thèse de doctorat de Mathématiques appliquées

Dirigée par Charles Bouveyron et Pierre Latouche

Présentée et soutenue publiquement le 26/10/2017 devant le jury composé de :

Francis BACH	INRIA – École Normale Supérieure	Président
Charles BOUVEYRON	Université Paris Descartes	Directeur
Gilles CELEUX	INRIA	Examinateur
Julien CHIQUET	AgroParisTech – INRA	Invité
Julie JOSSE	École Polytechnique	Examinatrice
Pierre LATOUCHE	Université Paris 1 Panthéon-Sorbonne	Encadrant
Jean-Michel MARIN	Université de Montpellier	Rapporteur

Après avis des rapporteurs : Nial FRIEL (University College Dublin)
Jean-Michel MARIN (Université de Montpellier)

Résumé

Le déferlement numérique qui caractérise l'ère scientifique moderne a entrainé l'apparition de nouveaux types de données partageant une démesure commune : l'acquisition simultanée et rapide d'un très grand nombre de quantités observables. Qu'elles proviennent de puces ADN, de spectromètres de masse ou d'imagerie par résonance nucléaire, ces bases de données, qualifiées de *données de grande dimension*, sont désormais omniprésentes, tant dans le monde scientifique que technologique. Le traitement de ces données de grande dimension nécessite un renouvellement profond de l'arsenal statistique traditionnel, qui se trouve inadapté à ce nouveau cadre, notamment en raison du très grand nombre de variables impliquées. En effet, confrontée aux cas impliquant un plus grand nombre de variables que d'observations, une grande partie des techniques statistiques classiques est incapable de donner des résultats satisfaisants.

Dans un premier temps, nous introduisons les problèmes statistiques inhérents aux modèles de données de grande dimension. Plusieurs solutions classiques sont détaillées et nous motivons le choix de l'approche empruntée au cours de cette thèse : le paradigme bayésien de sélection de modèles. Ce dernier fait ensuite l'objet d'une revue de littérature détaillée, en insistant sur plusieurs développements récents.

Viennent ensuite trois chapitres de contributions nouvelles à la sélection de modèles en grande dimension. En premier lieu, nous présentons un nouvel algorithme pour la régression linéaire bayésienne parcimonieuse en grande dimension, dont les performances sont très bonnes, tant sur données réelles que simulées. Une nouvelle base de données de régression linéaire est également introduite : il s'agit de prédire la fréquentation du musée d'Orsay à l'aide de données vélibs. Ensuite, nous nous penchons sur le problème de la sélection de modèles pour l'analyse en composantes principales (ACP). En nous basant sur un résultat théorique nouveau, nous effectuons les premiers calculs exacts de vraisemblance marginale pour ce modèle. Cela nous permet de proposer deux nouveaux algorithmes pour l'ACP parcimonieuse, un premier, appelé GSPPCA, permettant d'effectuer de la sélection de variables, et un second, appelé NGPPCA, permettant d'estimer la dimension intrinsèque de données de grande dimension. Les performances empiriques de ces deux techniques sont extrêmement compétitives. Dans le cadre de données d'expression ADN notamment, l'approche de sélection de variables proposée permet de déceler sans supervision des ensembles de gènes particulièrement pertinents.

Mots-Clefs : Apprentissage statistique, Grande dimension, Parcimonie, Sélection de modèles, Statistique bayésienne.

Abstract

The numerical surge that characterizes the modern scientific era led to the rise of new kinds of data united in one common immoderation: the simultaneous acquisition of a large number of measurable quantities. Whether coming from DNA microarrays, mass spectrometers, or nuclear magnetic resonance, these data, usually called *high-dimensional*, are now ubiquitous in scientific and technological worlds. Processing these data calls for an important renewal of the traditional statistical toolset, unfit for such frameworks that involve a large number of variables. Indeed, when the number of variables exceeds the number of observations, most traditional statistical techniques become inefficient.

First, we give a brief overview of the statistical issues that arise with high-dimensional data. Several popular solutions are presented, and we present some arguments in favor of the method utilized and advocated in this thesis: Bayesian model uncertainty. This chosen framework is the subject of a detailed review that insists on several recent developments.

After these surveys come three original contributions to high-dimensional model selection. A new algorithm for high-dimensional sparse regression called SpinyReg is presented. It compares favorably to state-of-the-art methods on both real and synthetic data sets. A new data set for high-dimensional regression is also described: it involves predicting the number of visitors in the Orsay museum in Paris using bike-sharing data. We focus next on model selection for high-dimensional principal component analysis (PCA). Using a new theoretical result, we derive the first closed-form expression of the marginal likelihood of a PCA model. This allows us to propose two algorithms for model selection in PCA. A first one called globally sparse probabilistic PCA (GSPPCA) that allows to perform scalable variable selection, and a second one called normal-gamma probabilistic PCA (NGPPCA) that estimates the intrinsic dimensionality of a high-dimensional data set. Both methods are competitive with other popular approaches. In particular, using unlabelled DNA microarray data, GSPPCA is able to select genes that are more biologically relevant than several popular approaches.

Keywords: Bayesian statistics, High-dimensional data, Model selection, Sparsity, Statistical machine learning

Rien ne m'est sûr que la chose incertaine

François Villon, 1458

Remerciements

Fidèle au plaidoyer en faveur de la parcimonie que ce document se veut être, je me suis efforcé de concilier sur cette page concision et reconnaissance(s).

Je ne saurais convenablement exprimer, parcimonieusement ou non, l'ampleur de la gratitude que j'éprouve pour Charles et Pierre, capitaines au long cours de ces trois années de doctorat. Leurs conseils, leurs idées débordantes, leurs encouragements et leur disponibilité ont considérablement vivifié ces trois années et donné corps à ce travail.

Je suis extrêmement reconnaissant envers Nial Friel et Jean-Michel Marin, qui ont rapporté ce travail. Les discussions avec l'un dans une grande ville d'Irlande ainsi que les cours vigoureux de l'autre dans un petit village des Alpes ont façonné ma formation aux statistiques bayésiennes. Mille mercis également à Francis Bach, Gilles Celeux, Julien Chiquet et Julie Josse d'avoir accepté de faire partie de mon jury.

Je remercie chaleureusement tous ceux qui ont croisé mon chemin au MAP5, au SAMM, à l'University College Dublin ainsi qu'à l'ITU of Copenhagen : chercheurs pédagogues, doctorants amicaux ou administrateurs compréhensifs. En particulier, merci à tous ceux avec qui j'ai eu la chance de directement collaborer : Michael Fop, Jes Frellsen, Warith Harchaoui et Brendan Murphy. Leurs relectures et conseils avisés ont considérablement influencé et amélioré ce manuscrit. J'ai hâte de poursuivre certains travaux que nous n'avons pu qu'esquisser.

Mon orientation vers les mathématiques appliquées doit inévitablement à un bon nombre de professeurs de mathématiques, qui m'ont fait découvrir et constamment redécouvrir cette discipline qui marie si joliment rigueur et imagination. Toujours par souci de parcimonie, je ne citerai que Pierre Savignoni, qui m'a tout simplement appris à aimer faire des maths, mais que tous les autres - depuis l'école primaire jusqu'aux écoles d'été - soient assurés de ma gratitude.

Je remercie profondément mes amis, ceux qui ont grandi avec moi sur cette petite île montagneuse comme ceux que j'ai rencontrés ensuite, dans un foyer catholique austère seulement en apparence, sur les bancs d'écoles finalement assez normales, ou près des machines à café de divers laboratoires.

Je termine ces remerciements parcimonieux en exprimant ma reconnaissance infinie envers ma famille, en particulier mes parents, mon frère, ma tante et oncle, qui m'ont accompagné avec tant de constance et de chaleur au cours de ces trois années. Enfin, cette thèse doit beaucoup à Valérie, sa première et si patiente relectrice, qui m'aidera j'espère à écrire les prochains chapitres de ma vie.

Contents

Essential nomenclature

n	number of observations in a data set
p	number of variables in a data set
i	observation index, in $\{1, ..., n\}$
j	variable index, in $\{1, ..., p\}$
$\mathbf{X}$	data matrix, in $\mathbb{R}^{n \times p}$
$\boldsymbol{\beta}$	vector
$\|\|\boldsymbol{\beta}\|\|_2$	Euclidean norm of $\boldsymbol{\beta}$
$\|\|\boldsymbol{\beta}\|\|_0$	ℓ_0 "norm" of $\boldsymbol{\beta}$ (number of nonzero coefficients)
$\mathbf{x}_i$	observation in $\mathbb{R}^p$ (i-th row of $\mathbf{X}$)
$\mathbf{v}$	binary vector in $\{0, 1\}^p$
$\text{Supp}(\mathbf{v})$	support of a vector (set of nonzero coefficients of $\mathbf{v}$)
$\bar{\mathbf{v}}$	binary vector whose support is the complement of $\text{Supp}(\mathbf{v})$
$\mathbf{X}_\mathbf{v}$	data matrix obtained by keeping only the variables that correspond to the support of $\mathbf{v}$
$\boldsymbol{\beta}_\mathbf{v}$	vector obtained by keeping only the coefficients that correspond to the support of $\mathbf{v}$
$\mathbf{A} \odot \mathbf{B}$	Hadamard (entrywise) product between two matrices of the same dimension

$\mathcal{S}_p^+$	set of positive semidefinite matrices of size p
$\mathcal{S}_p^{++}$	set positive definite matrices of size p
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$
$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian density with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}$, evaluated at $\mathbf{x} \in \mathbb{R}^p$
$\mathcal{B}(p)$	Bernoulli distribution with parameter p

Γ	gamma function
J_ν	Bessel function of the first kind of order $\nu \in \mathbb{R}$
K_ν	modified Bessel function of the second kind of order $\nu \in \mathbb{R}$

1

High-Dimensional Statistical Machine Learning

1.1 The ubiquity of high-dimensional data

The traditional scientific paradigm that prevailed in statistics during the 20th century focused on data which involves a small number n of *observations* (patients in a medical cohort, plants in an agricultural experiment, citizens in a political poll) is large compared to the number p of *variables* (or features), which are carefully-designed measurements that are conducted for each observation (the blood pressure of a patient, the height of a plant, or the age of a voter). The last decades have witnessed a formidable development of data-acquisition technologies, leading to the availability of data of a dramatically different nature. These data, which may involve thousands or millions of variables (and potentially much less observations), are commonly referred to as *high-dimensional*. High-dimensional data have become

ubiquitous in several scientific and technological fields, resulting in a phenomenon sometimes called *big data*. We illustrate this presence with a few important examples.

- **Computer vision** was perhaps the first field to be confronted to high-dimensional data. Indeed, in an image, each pixel corresponds to one variable, which means that computer vision has routinely dealt with hundreds of variables since the 1980s (see e.g. Sirovich and Kirby, 1987). Modern high-resolution, hyperspectral, or video data may involve millions of variables.

- **Biotechnology** has also produced countless example of very high-dimensional data in the last decades. In particular, recent advances in genomics allow nowadays to determine the genomic profile of individuals at a relatively limited cost. For example, a DNA microarray can measure the expression levels of thousands of genes. There are many kinds of microarrays (Drăghici, 2012, Chapter 3), we show one example in Figure 1.1.

- **Chemometrics** witnessed the simultaneous development of mass spectrometry (MS), near-infrared (NIR) spectroscopy and nuclear magnetic resonance (NMR) spectroscopy, which produced high-dimensional spectra of various chemical samples. As an example, we show in Figure 1.2 the NIR spectra of several meat samples. Classifying such spectra is extremely useful in food authenticity applications.

1.2 Failures of classical approaches in high dimensions

High-dimensional data is the source of many statistical challenges. Most of them appear to be linked to the geometry of high-dimensional spaces. We provide some insight on this geometry and discuss statistical consequences.

1.2.1 Bellman's curse of dimensionality and the peculiar geometry of high-dimensional spaces

Let us begin with two very simple yet extremely counter-intuitive facts about high-dimensional Euclidean geometry (both illustrated in Figure 1.3).

- The volume of the p-dimensional unit hyperball is given by the formula

$$V_p = \frac{\pi^{p/2}}{\Gamma(p/2-1)} \xrightarrow[p\to\infty]{} 0,$$

  where Γ is the gamma function. Consequently, as p grows to infinity, the volume of the hyperball vanishes (notably compared to the volume of the hypercube that stays constant when p grows). This illustrates a potentially disastrous problem: *high-dimensional Euclidean neigborhoods are essentially empty*. This fact was called the *empty space phenomenon* by Scott and Thompson (1983).

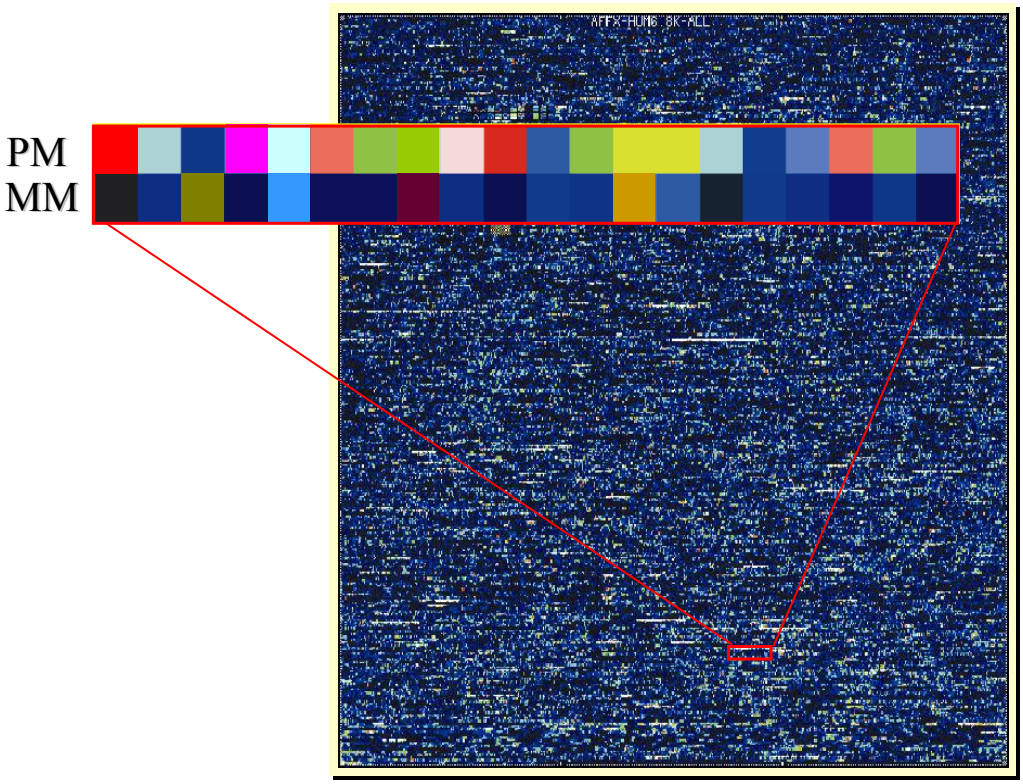


Figure 1.1 – Affymetrix microarray, reproduced with permission from Drăghici (2012, p. 113). The zoom corresponds to one gene, represented by a set of 20 *probes* (fragments of DNA printed on a solid substrate). The two rows in the zoom correspond to the two versions of the probe: the perfect match (PM, which matches the target sequence perfectly) and the mismatch (MM, which is identical to the PM except for one nucleotide change at the central base position). If the biological sample studied contains the gene of interest, the PM is expected to hybridize and the central mismatch prevents the MM to hybridize. The *expression level* of the gene is usually defined as the average difference between the PM and the MM.

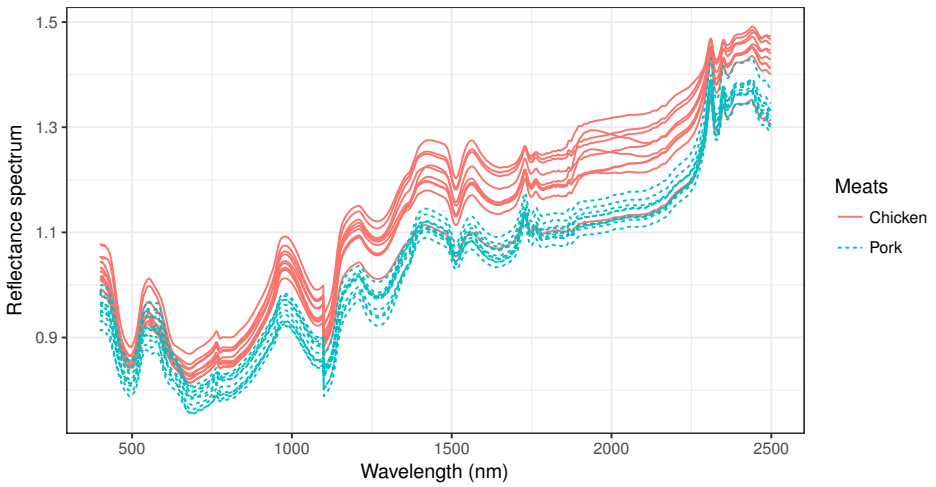


Figure 1.2 – Ten chicken and ten pork spectra from the food authenticity data of McElhinney et al. (1999). Each NIR spectrum consists in 1050 wavelengths. On the full data set, which involves five different meats, popular statistical machine learning algorithms such as support vector machines (SVMs, Cortes and Vapnik, 1995) and random forests (Breiman, 2001) perform poorly because of the dimensionality of the data (Murphy et al., 2010).

- The volume of the "1% shell" (obtained by removing a concentric ball of radius 0.99 from the unit sphere) of this hyperball is equal to $1 - 0.99^p$. Therefore, in high-dimensional settings, the volume of the hyperball will be not only extremely small, but also concentrated in the outer shell. This is an instance of a much more general probabilistic phenomenon called the *concentration of measure* (Ledoux, 2001) which roughly states that smooth real functions of high-dimensional and weakly correlated random variables are nearly constant (in the simple shell example, the high-dimensional random variable is simply a uniform variable over the hyperball and the function is the Euclidean norm, which is almost constant equal to one).

As illustrated by these simple examples, the Euclidean distance, which has been the most popular measure of error in statistics since Legendre's (1805) leasts squares, has a counter-intuitive behavior in high dimensions. This will lead classical distance-based classifiers to perform poorly in high-dimensional settings. However, the peculiar geometry of high-dimensional spaces has also several benefits – for example, the concentration of measure phenomenon allows to compute useful concentrations inequalities (Boucheron et al., 2013) and suggests that high-dimensional data has a natural tendency to lie close to low-dimensional manifolds. This ambivalent view was already famously described by Bellman, in the preface of this book *Dynamic Programming* (1957),

> *All this* [the problems related to high-dimensional geometry] *may be subsumed under the heading "the curse of dimensionality". Since this is a curse (...) there is no need to feel discouraged about the possibility of obtaining significant results despite it.*

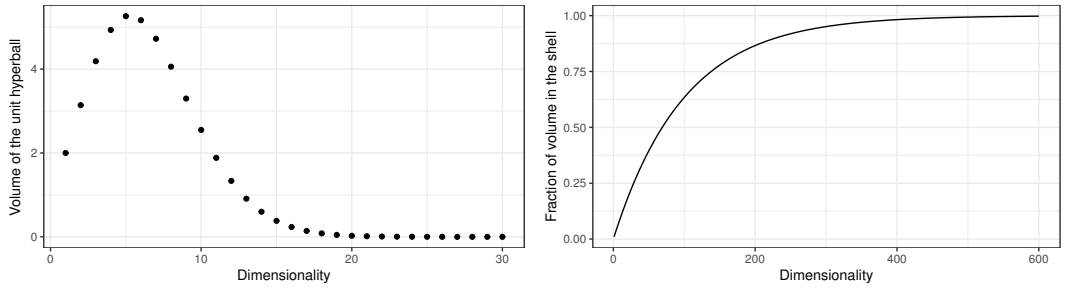


Figure 1.3 – *Left:* Volume of the unit hyperball when the dimension grows. *Right:* Fraction of hyperball volume in the outer shell.

1.2.2 Statistical failures in high-dimensions

The mildest statistical methods are affected by the curse of dimensionality. Among the victims, one of the most severe is *the empirical covariance*. When dealing with p-dimensional data, computing the empirical covariance implies estimating $p(p-1)/2$ parameters, which means that the dimensionality of the parameter space is actually much higher than the (already large) dimensionality of the data. This heuristically explains why covariance estimation is extremely challenging in high dimensions. We illustrate this obstacle in Figure 1.4 in a very simple setting that shows that, even with i.i.d. standard normal data, the empirical covariance is an unreliable estimator of the true covariance. Empirical covariances are widely used in statistics, from portfolio analysis (Markowitz, 1952) to model-based classification (Fraley and Raftery, 2002). In particular, principal component analysis (PCA, see e.g. Jolliffe and Cadima, 2016, for a review), which is arguably the most popular dimension reduction technique, depends on the use of a reliable estimate of the covariance matrix. While reducing the dimensionality of the data using PCA appears as a natural technique to tackle the curse of dimensionality, the failure of the empirical covariance will lead PCA to be potentially useless in very high-dimensional problems (Johnstone and Lu, 2009). Another standard statistical tool that suffers dramatically from the curse of dimensionality is *linear regression*. Gaussian linear regression is equivalent to solving a noisy linear system

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a vector of responses, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a matrix of predictors (often called the *design matrix*) and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter. When p remains smaller than n, the most classical estimate of $\boldsymbol{\beta}$ is given by *ordinary least-squares* (OLS)

$$\boldsymbol{\beta}_{\mathrm{LS}} = \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\mathbf{w}||_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{1.2}$$

In the very simple orthogonal setting where $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, it can be shown that the OLS estimation error is

$$\mathbb{E}[||\boldsymbol{\beta} - \boldsymbol{\beta}_{\mathrm{LS}}||_2^2] = p\sigma^2, \tag{1.3}$$

which grows linearly with the dimensionality. Things get even worse in more general settings: when p is larger than n, we are trying to solve an ill-posed linear system with *more unknowns than equations*, and the least-squares optimization problem has infinitely many solutions.

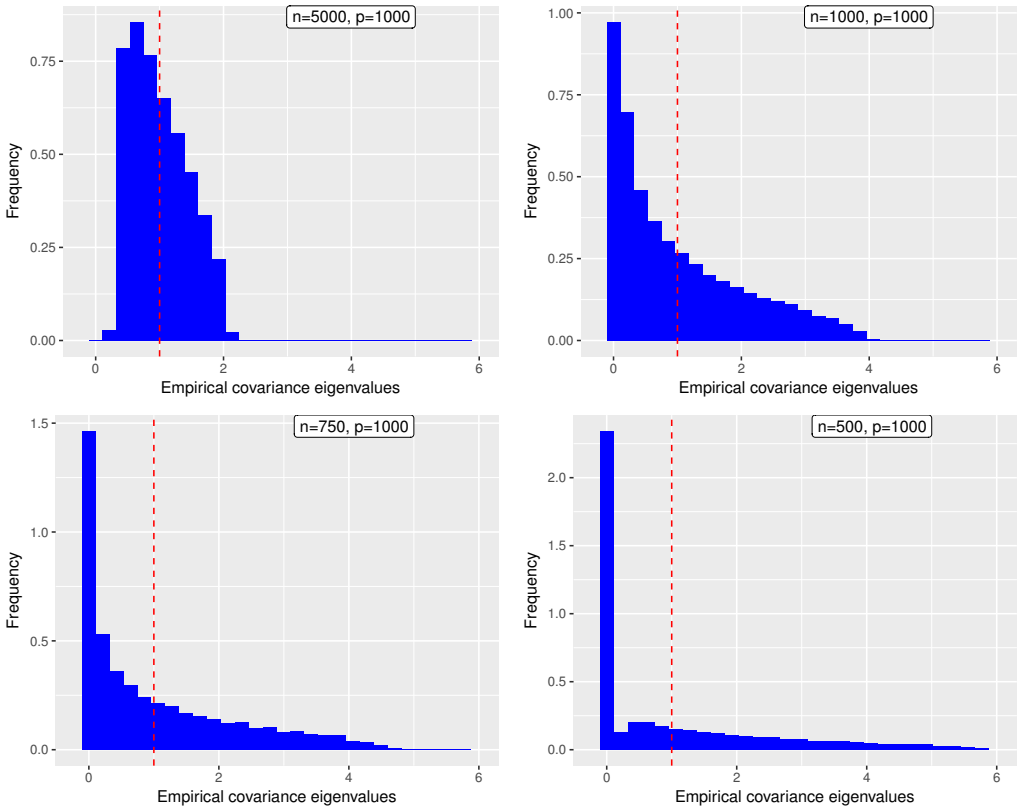


Figure 1.4 – Spectra of the empirical covariances of i.i.d. data coming from a $\mathcal{N}(0, \mathbf{I}_{1000})$ distribution. The number of observations has to be much higher than p for the spectrum to concentrate towards the theoretical value of 1. For smaller values of n, many eigenvalues are null or very small, and the rest of the spectrum is very spread out.

1.3 High-dimensional learning: sparsity and model selection

We present here the sparse approach, which has imposed itself as a very popular way to tackle the curse of dimensionality.

1.3.1 The bet on sparsity

Parametric statistical models assume that the observed data $\mathbf{X}$ comes from a density in a parametrized family $(p(\cdot|\boldsymbol{\theta}))_{\boldsymbol{\theta}\in\Theta}$. Usually, the dimension of Θ grows with the dimensionality p of the data (linearly, as in linear regression, or even quadratically, as in covariance estimation), which renders the estimation problem extremely challenging when dealing with high-dimensional data. An efficient way to tackle this issue is to constrain the parameter space. In particular, *sparsity constraints* have been extremely successful in recent years. In their simplest form, these constraints involve finding a parameter that has at most $q \ll p$ nonzero coefficients. Note however that sparsity constraints can be more subtle than simply assuming that $\boldsymbol{\theta}$ has a limited number of nonzero coefficients – some examples include locally and globally sparse PCA (Chapter 4), structured sparsity (see e.g. Jenatton et al., 2011) or low-rank matrix completion (see e.g. Candès and Tao, 2010). Under the assumption that there exists an optimal sparse parameter, most of the problems linked to the curse of dimensionality tend to vanish (Candès, 2014). This motivates to focus on sparse settings in high-dimensional scenarios, even if we don't actually believe the data-generating mechanism to be sparse. Or, in the words of Hastie et al. (2015),

> *This has been termed the "bet on sparsity" principle:* Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

Let us illustrate this bet with an example. Consider the food authenticity data set presented in Figure 1.2: we have $n = 110$ meat samples of pork and chicken, of which 1050-dimensional NIR spectra have been measured. For visualization purposes, we wish to reduce the dimensionality of the data to 2. While it is possible to use regular PCA, it leads to the destruction of most of the discriminative information between the two classes, as shown in Figure 1.5. This illustrates the failure of covariance estimation outlined in the previous section. What if we were to bet on sparsity? Rather than projecting the data onto the subspace spanned by the top two eigenvectors of the empirical covariance matrix (as in the regular PCA procedure), we could choose to project it onto a subspace spanned by 2 sparse vectors that share the same support, as we advocate in Chapter 4. For this purpose, we use the methodology called *globally sparse probabilistic PCA* which is detailed in Chapter 4. Since PCA is expected to perform well when $p \approx n$, we choose the two basis vectors to be n-sparse. As shown on Figure 1.5, this leads to a much more sensible reduction of the dimensionality, that manages to keep the discriminative information between the two meats.

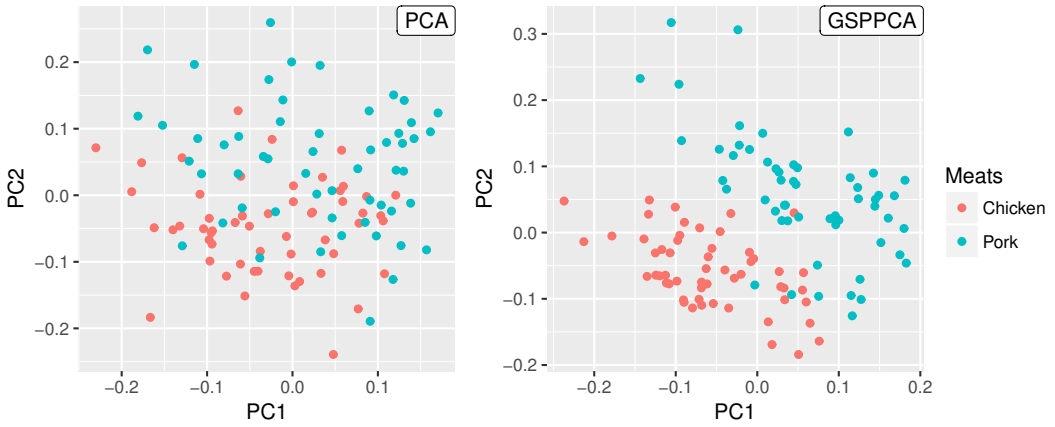


Figure 1.5 – PCA projections for the food authenticity data of McElhinney et al. (1999). *Left:* PCA using all 1050 variables; most valuable information about the nature of the meat appears to be lost. *Right:* PCA using only 110 variables (which corresponds to the number of observations) selected using globally sparse probabilistic PCA (see Chapter 4); the two classes are almost linearly separable.

1.3.2 Algorithms for sparse high-dimensional learning

A natural way to find a sparse parameter is to maximize a penalized version of the likelihood

$$\hat{\boldsymbol{\theta}} \in \text{argmax}_{\boldsymbol{\theta}\in\Theta} \log p(\mathbf{X}|\boldsymbol{\theta}) - \lambda||\boldsymbol{\theta}||_0, \tag{1.4}$$

where $||\boldsymbol{\theta}||_0 = \#\text{Supp}(\boldsymbol{\theta})$ is the ℓ_0 (pseudo)norm of the vector $\boldsymbol{\theta}$. When the dimensionality does not exceed a few dozens, this can be done exactly using combinatorial algorithms (see e.g. Mazumder and Radchenko, 2017, for recent perspectives on linear regression). Unfortunately, exact resolution is generally computationally challenging because of the discrete nature of the ℓ_0 norm – for example, in the case of linear regression, it leads to a NP-hard problem. Therefore, for higher-dimensional problems, approximate solutions are usually pursued of (1.4). Greedy techniques such as *stepwise variable selection* have been popular for several decades (Weisberg, 1980, Section 8.5). In particular, these methods have been recently successful at tackling the empirical covariance failure in model-based clustering (Fop and Murphy, 2017) and classification (Murphy et al., 2010; Maugis et al., 2011).

While techniques that directly attack the optimization problem (1.4) are usually efficient, they do not scale to very large dimensions. Consequently, an important body of work has been devoted to the study of simpler related optimization problems. In particular, following the seminal work of Tibshirani (1996) and Chen et al. (1998), much effort has been directed towards ℓ_1 relaxations of the form

$$\hat{\boldsymbol{\theta}} \in \text{argmax}_{\boldsymbol{\theta}\in\Theta} \log p(\mathbf{X}|\boldsymbol{\theta}) - \lambda||\boldsymbol{\theta}||_1. \tag{1.5}$$

Such optimization problems – often called *lasso problems*, following Tibshirani (1996) – can be solved much easier than the ℓ_0 penalized problem (1.4). Indeed, due to the convexity of the ℓ_1 norm, many computational strategies linked to convex optimization are available (Hastie et al., 2015, Chapter 5). While ℓ_1 penalization was originally heuristically motivated

as a more tractable relaxation of the ℓ_0 norm, it can be actually shown that, under suitable assumptions on the problem at hand, ℓ_1 penalization can solve, approximatively or even exactly, the original ℓ_0 penalized problem (see Candès, 2014, and Hastie et al., 2015, Chapter 11, for recent reviews). This fact comes as a surprise, as summarized by Candès and Tao (2010):

> *The surprise here is that admittedly, there is* a priori *no good reason to suspect that convex relaxation might work so well. There is* a priori *no good reason to suspect that the gap between what combinatorial and convex optimization can do is this small. In this sense, we find these findings a little unexpected.*

1.3.3 From sparsity to model selection

In spite of their many merits, both penalization procedures share several common calibration problems. In complex models, there is no good justification of the use of the ℓ_0 (or the ℓ_1) norm since the number of nonzero parameters may provide a poor complexity measure (see e.g. Gao and Jojic, 2016). Even in simple models such as Gaussian linear regression, more subtly tailored penalties are sometimes more useful – such as the elastic net of Zou and Hastie (2005) or the slope heuristics of Birgé and Massart (2007).

In the Bayesian setting, however, these calibration problems can be solved in a systemized way. Indeed, we can recast seeking sparsity as a *model selection problem*: seeing all possible supports of $\boldsymbol{\theta}$ as candidate statistical models, finding a sparse parameter is equivalent to finding an optimal model. As we explain in the next chapter, the Bayesian paradigm that we study in this thesis provides a simple and coherent framework for solving this model selection problem, and can be seen as an automatic way to design sparsity-inducing penalties.

1.4 Organization of the thesis

While this first chapter briefly introduced the challenges offered by high-dimensional data, the next chapter will be devoted to a review of the foundations and some recent advances of the Bayesian framework of model selection and uncertainty. Introduced by Harold Jeffreys and Dorothy Wrinch in the 1930s, this probabilistic paradigm will constitute our main technical tool to learn from high-dimensional data.

Then will come the original contributions of this thesis. Chapter 3 will present a new algorithm for high-dimensional linear regression named SpinyReg. Thorough empirical examination shows that this algorithm leads to very competitive predictive and interpretative performance. In particular, we introduce a new data set that uses social transportation to predict a touristic index: the number of visitors of the Orsay museum in Paris. In Chapter 4, we introduce a framework for variable selection in high-dimensional principal component analysis (PCA). From a theoretical perspective, we derive the first closed-form expression of the marginal likelihood of a PCA model. This allows us to design an highly scalable algorithm for unsupervised variable selection – with $O(np)$ complexity. On several real and synthetic data sets, this algorithm, called GSPPCA for *globally sparse probabilistic PCA*, vastly outperforms traditional sparse PCA approaches. In particular, GSPPCA is applied

to a DNA microarray data set from which it is able to select much more relevant genes than his competitors. The main theoretical contribution behind GSPPCA, detailed and extended in Appendix A, led us to develop consequently an algorithm that estimates the intrinsic dimension of a high-dimensional data set. This framework, presented in Chapter 5, uses a new prior structure for the Bayesian PCA model to perform exact model selection. When the number of observations is very small, our approach proves to be extremely useful. Chapter 6 is devoted to a brief overview of these contributions and to perspectives for ongoing and future work.

2

Bayesian Model Uncertainty

2.1 Introduction: collecting data, fitting many models

Today, the conventional statistical process embodied by Fisher's (1938) famous exhortation

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*

has been largely reversed. Indeed, as we illustrated in Chapter 1, modern scientific research often involves the simultaneous measurement of a large number of (potentially irrelevant) variables before statistical practice is actually set in motion. Rather than falsifying or corroborating predetermined hypotheses, researchers mine these high-dimensional data using a very large number of statistical models. This new scientific method – caricatured by the motto *collect data first, ask questions later* – was powered by the recent rise of automatic statistical software, illustrated for example by the growing popularity of Stan (Carpenter et al., 2016), PyMC3 (Salvatier et al., 2016), or Edward (Tran et al., 2016).

In this new context, it appears of paramount importance to be able to compare the relevance and the performance of these many models, and to identify the best ones. *Bayesian model uncertainty* provides a systematized way of answering these questions. This approach, whose history is briefly summarized in next subsection, has witnessed a remarkable evolution in the last decades, that has brought about several new theoretical and methodological advances. The foundations of Bayesian model uncertainty, as well as some of these recent developments are the focus of this chapter. In particular, we insist on links with penalized model selection and learning theory, predictive out-of-sample performance in the non-asymptotic regime, and extensions to wider classes of probabilistic frameworks including unidentifiable, likelihood-free, and high-dimensional models.

2.1.1 A brief history of Bayesian model uncertainty

Bayesian model uncertainty is essentially founded on the idea of spreading prior beliefs between competing models, implying that the marginal distribution of the data follows a mixture of all model-specific marginals. This paradigm was initially developed by Sir Harold Jeffreys and Dorothy Wrinch in a series of papers (Wrinch and Jeffreys, 1919, 1921, 1923), culminating with Jeffreys's book *Theory of Probability* (1939). For a recent perspective on the place of Bayesian model uncertainty in *Theory of Probability*, see Robert et al. (2009). It is worth mentioning that Jeffreys considered it an essential piece of his scientific work. Indeed, in a 1983 interview with Dennis Lindley quoted by Etz and Wagenmakers (2017), Jeffreys stated that he thought that his most important contribution to probability and statistics was "the idea of a significance test (...) putting half the probability into a constant being 0, and distributing the other half over a range of possible values".

Independently, similar ideas were developed by J. B. S. Haldane (1932), as recently exhibited by Etz and Wagenmakers (2017), and also by Alan Turing, who designed related tests to decrypt Enigma codes during World War II, as testified by Turing's main statistical collaborator I. J. Good (1979).

This scientific paradigm gained considerable popularity in the beginning of the 1990s, in particular with David MacKay's (1991) thesis which had a significant impact on the then-burgeoning machine learning community, and with the review paper of Robert Kass and Adrian Raftery (1995), which quickly disseminated Jeffreys's ideas to the whole scientific community.

2.2 The foundations of Bayesian model uncertainty

In this section, we present the Bayesian framework of model uncertainty, essentially founded by Jeffreys in his book *Theory of Probability* (1939). We start with some data $\mathcal{D}$ living in a probability space and with a family $\mathcal{M}_1, ..., \mathcal{M}_{d_{\max}}$ of candidate statistical models. Unless specified otherwise, these models correspond to parametric families (indexed by $\Theta_1, ..., \Theta_{d_{\max}}$) of probability measures over the data space, which are absolutely continuous with respect to a reference measure (usually the Lebesgue or the counting measure).

2.2.1 Handling model uncertainty with Bayes theorem

The Bayesian framework may be summarized in a single sentence: *model unknown quantities as random variables to assess their uncertain nature*. Under model uncertainty, there are two different kinds of unknowns: models and their parameters. We assume therefore that priors $p(\mathcal{M}_d)$ and $p(\theta|\mathcal{M}_d)$ are specified. As in Draper (1995), we may summarize this framework by contemplating what we will call the *expanded model*

$$\mathcal{M}_{\text{Exp}} : \begin{cases} \mathcal{M}_d \sim p(\cdot) \\ \theta \sim p(\cdot|\mathcal{M}_d) \\ \mathcal{D} \sim p(\cdot|\theta, \mathcal{M}_d). \end{cases} \tag{2.1}$$

A way of interpreting this three-stage hierarchical model is to see the global prior distribution (over both models and parameters) as a way of sampling distributions $p(\cdot|\theta, \mathcal{M}_d)$ over the data space. In this very general framework, it is actually not necessary to assume that the model family is finite, or even countable. Draper (1995) advocates for example the use of a continuous model family to gain flexibility, and shows several applications (see also Gelman et al., 2013, Chapter 7). Note that the resulting marginal distribution of the data is a mixture model (an infinite one in the case of an infinite model family).

Now that we have specified the probabilistic architecture, model uncertainty will be tackled automatically by the Bayesian machinery. Indeed, from Bayes's theorem, we obtain posterior probabilities of models as, for all $d \in \{1, ..., d_{\max}\}$,

$$p(\mathcal{M}_d|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_d)p(\mathcal{M}_d),$$

where

$$p(\mathcal{D}|\mathcal{M}_d) = \int_{\Theta_d} p(\mathcal{D}|\theta, \mathcal{M}_d)p(\theta|\mathcal{M}_d)d\theta \tag{2.2}$$

is the *marginal likelihood* of model $\mathcal{M}_d$ – also known as *evidence* (see e.g. MacKay, 2003) or *type II likelihood* (see e.g. Berger, 1985). This quantity, which may be interpreted as the prior mean of the likelihood function, will play a central role in Bayesian model uncertainty. Besides computing posterior model probabilities, that have an intuitive interpretation for assessing model uncertainty within the family at hand (see Section 2.2.2), it is also useful to conduct pairwise model comparisons between two models, say $\mathcal{M}_d$ and $\mathcal{M}_{d'}$. This can be done using the *posterior odds* against $\mathcal{M}_{d'}$

$$\frac{p(\mathcal{M}_d|\mathcal{D})}{p(\mathcal{M}_{d'}|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{M}_d)}{p(\mathcal{D}|\mathcal{M}_{d'})} \frac{p(\mathcal{M}_d)}{p(\mathcal{M}_{d'})}. \tag{2.3}$$

Posterior odds involve two terms: the prior odds $p(\mathcal{M}_d)/p(\mathcal{M}_{d'})$ which only depend on the prior distribution over the family of models, and the ratio of marginal likelihoods,

$$\mathrm{BF}_{d/d'} = \frac{p(\mathcal{D}|\mathcal{M}_d)}{p(\mathcal{D}|\mathcal{M}_{d'})}, \tag{2.4}$$

called the *Bayes factor* – a term partly coined by Alan Turing during World War II, who called it the "factor" (Good, 1979). The main appeal of the Bayes factor is that, regardless of prior probabilities of models, it provides a good summary of the relative support for $\mathcal{M}_d$ against $\mathcal{M}_d$ provided by the data. Although extremely convenient, this simple interpretation has been subject to much debate (see Section 2.2.2).

Now that we have a posterior distribution over the family of models, how can we make use of this knowledge of model uncertainty to take decisions?

The first answer is *Bayesian model selection*: settling for the model with the largest posterior probability, leading to the choice

$$d^* = \underset{d\in\{1,\dots,d_{\max}\}}{\arg\max}\ p(\mathcal{M}_d|\mathcal{D}). \tag{2.5}$$

This offers a systematized way of choosing a single model within the family, and can be seen as an instance of hypothesis testing. It is worth mentioning that Bayesian model selection was originally described by Jeffreys as an alternative to classical hypothesis tests. For perspectives on the links between the different approaches of testing, see Berger (2003).

However, when no model truly stands out, it is often better to combine all models (or some of the bests) to grasp more fully the complexity of the data. There comes the second important Bayesian approach of model uncertainty: *Bayesian model averaging* (BMA). BMA allows to borrow strength from all models to conduct better predictions. Specifically, assume that we are interested in a quantity Δ, that has the same meaning in all models. This quantity can be a value that we wish to predict (like the temperature of the Pacific ocean using several forecasting models, as in Raftery et al., 2005), or a parameter that appears in all models (like the coefficient of a linear regression model). For a more detailed discussion on what it means to have "the same meaning in all models", see the discussion of Draper (1999) and the rejoinder of the excellent BMA tutorial of Hoeting et al. (1999). The BMA posterior distribution of Δ will be its posterior distribution within $\mathcal{M}_{\mathrm{Exp}}$,

$$p(\Delta|\mathcal{D}) = \sum_{d=1}^{d_{\max}} p(\Delta|\mathcal{M}_d, \mathcal{D})p(\mathcal{M}_d|\mathcal{D}), \tag{2.6}$$

which corresponds to a mixture of all model-specific posteriors. Taking the mean of the BMA posterior gives a natural point estimate for predicting the value of Δ,

$$\hat{\Delta} = \mathbb{E}_\Delta(\Delta|\mathcal{D}) = \sum_{d=1}^{d_{\max}} \mathbb{E}_\Delta(\Delta|\mathcal{M}_d, \mathcal{D})p(\mathcal{M}_d|\mathcal{D}). \tag{2.7}$$

Sometimes, the average may not be conducted over all models, but solely over a smaller subfamily, as in Madigan and Raftery's (1994) Occam's window.

These two popular techniques, which will constitute our main focus, can be embedded within a larger decision theoretic framework. In this context, Bayesian model selection corresponds to the 0-1 loss and BMA corresponds to the squared loss (see e.g. Bernardo and Smith, 1994, Section 6.1 or Clyde and George, 2004, Section 6).

2.2.2 Interpretations of Bayesian model uncertainty

2.2.2.a Interpretation of posterior model probabilities

Contrarily to other techniques that tackle the model uncertainty problem, the Bayesian approach produces easily interpretable results. Indeed, posterior model probabilities are readily understandable, even by non-statisticians, because of their intuitive nature. But what is their precise meaning? Formally, for each $d \in \{1, ..., d_{\max}\}$, the quantity $p(\mathcal{M}_d|\mathcal{D})$ is the probability that $\mathcal{M}_d$ is true given the data, given that we accept the prior distributions over models and their parameters, and given that one of the models at hand is actually true. There are several points of this statement that need further description. First, the controversial question of the relevance of the chosen priors raises many concerns, as described in Section 2.2.3.b. Second, the assumption that one of the models is actually true is often problematic. Indeed, in most applied cases, it appears overoptimistic to assume that the true data-generating model is contained within the tested family. In particular, in problems coming from social sciences or psychology, it seems clear that the true data-generating mechanism is likely to be beyond the reach of scientists (see e.g. Gelman and Shalizi, 2013). However, reasoning from the perspective that one of the models is true may remain scientifically valid on several grounds. Indeed, most scientific inference is made conditionally on models (models that are usually known to be false) in order to actually conduct science – a striking example is that the notoriously false Newtonian mechanics still flourish today, because it provides a scientifically convenient framework. Rather than conditioning on a single model, Bayesian model uncertainty conditions on a set of models. This is perhaps as wrong as conditioning on a single model, but it certainly is more useful. Conditioning a scientific process on a wrong hypothesis is indeed acceptable as long as this hypothesis is powerful or useful. Or, as famously explained by Box (1979),

> *The law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules. For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".*

Therefore, the (slightly less formal) interpretation of posterior model probabilities that we will adopt is that $p(\mathcal{M}_d|\mathcal{D})$ is *the probability that $\mathcal{M}_d$ is the most useful model within the family at hand.* Actually, a formalization of this interpretation (which gives a precise predictive sense to the "usefulness" of a model) in the case where the true model is out of hand – this situation is often referred to as the $\mathcal{M}$-open scenario, as defined by Bernardo

and Smith (1994, Section 6.1.2) – was provided by Dawid's (1984) *prequential* (predictive sequential) analysis. For discussions on prequential analysis and related predictive interpretations, see also Kass and Raftery (1995, Section 3.2), Gneiting and Raftery (2007, Section 7.1) and Vehtari and Ojanen (2012, Section 5.6). Similarly, Germain et al. (2016) gave a new and theoretically grounded predictive foundation of Bayesian model uncertainty which gives more support to the interpretation that we advocate here. We present a detailed overview of this approach in Section 2.2.4.

2.2.2.b Interpretation of Bayes factors

A key asset of Bayes factors is that, contrarily to posterior model probabilities, they do not depend on prior model probabilities (which are often arbitrary). However, this independence comes at the price of a more controversial interpretability. Since the Bayes factor is equal to the ratio of posterior odds to prior odds, it appears natural to consider it as the quantification of the evidence provided by the data in favor of a model – or, as Good (1952) called it, the *weight of evidence*. This interpretation, which dates back to Jeffreys and Wrinch, was advocated notably by Berger (1985, Section 4.3.3) and Kass and Raftery (1995). This interpretation was criticized by Lavine and Schervish (1999), who showed that, rather than seeing a Bayes factor as a measure of support, it was more sensible to interpret it as a measure of the change of support brought about by the data. In their words,

> *What the Bayes factor actually measures is the change in the odds in favor of the hypothesis when going from the prior to the posterior.*

In a similar fashion, Lindley (1997) warned against the use of Bayes factors, and suggested to rather use posterior odds.

2.2.3 Specifying prior distributions

2.2.3.a Model prior probabilities: non-informative priors and the simplicity postulate

When there is little prior information about the plausibility of different models, it is reasonable to follow Keynes's (1921, Chapter 4) principle of indifference and to choose the uniform prior over models $p(\mathcal{M}_d) \propto 1$. In this setting, using posterior model probabilities will be equivalent to using Bayes factors. However, it is often appropriate to seek more sensible priors that translate some form of prior knowledge. For example, in variable selection problems that involve a very large number of variables (e.g. 10.000 genes), it appears reasonable to give a higher prior probabilities to models that involve only a moderate amount of variables (e.g. preferring a priori a model that involve 100 genes over one that involves 10.000). For examples of similar approaches, see Narisetty and He (2014) or Yang et al. (2016). Actually, this rationale was already advocated by Jeffreys (1961, p. 46):

> *All we have to say is that the simpler laws have the greater prior probabilities. This is what Wrinch and I called the* simplicity postulate.

This simplicity postulate is linked to a philosophic principle known as *Occam's razor*, named after the 14th century philosopher and theologian William of Occam. Occam's razor essen-

tially states that, in the absence of strong evidence against it, a simpler hypothesis should be preferred to a more complex one. Actually, Bayesian model uncertainty involves *two* Occam's razors. The first one is precisely the simplicity postulate, and the second one is the fact that, when two models explain the data equally well, the simplest one has a larger marginal likelihood (see Section 2.2.5).

2.2.3.b Parameter prior probabilities and the Jeffreys-Lindley paradox

In Bayesian parameter inference, the influence of the prior distribution tends to disappear in the long run (when the number of observations tends to infinity). A formalization of this argument is the celebrated Bernstein-von Mises theorem (see e.g. Van der Vaart, 2000, Chapter 10). This phenomenon is less present when tackling model uncertainty, and poorly chosen prior distributions may lead to disastrous results even in the asymptotic regime. A famous instance of this problem is the *Jeffreys-Lindley paradox*, which essentially states that using improper or very diffuse prior distributions for parameters will lead to selecting the simplest model, regardless of the data. Popularized by Lindley (1957), this paradox had already been pointed out by Jeffreys (1939). It is also known as the *Bartlett paradox* because of Bartlett's (1957) early insight on it. For more details, see Spanos (2013) or Robert (2014) on the epistemological side, and Robert (1993) or Villa and Walker (2017) on the technical side. The main concern about this paradox is that diffuse priors are often chosen as default priors because of their objective nature. Thus, some particular care has to be taken when specifying priors in the presence of model uncertainty. While the use of improper or diffuse priors is generally proscribed for model selection purposes, several approaches have been proposed to bypass this problem. A first simple instance where improper priors may be acceptable is the case where a parameter appears in all models, like the intercept or the noise variance of a linear regression model (see e.g. Marin and Robert, 2014, pp. 44, 82). Another option is to use some form of resampling. First, perform Bayesian inference on a subset of the data using a (potentially improper) prior distribution, then use the obtained posterior as a prior for the rest of the data. This idea is the foundation of the *fractional Bayes factors* of O'Hagan (1995) and the *intrinsic Bayes factors* of Berger and Pericchi (1996). These techniques share the usual drawbacks of subsampling methods: they are computationally intensive and are inadequate when the number of observations is small.

Since using improper or diffuse priors is difficult in model uncertainty contexts, it appears necessary to use methods that allow to choose proper priors. Several approaches exist (Bayarri and Berger, 2013, Section 18.6), but we chose to focus mainly in this thesis on the *empirical Bayes* technique. For each model, empirical Bayes consider a parametric family of priors $(p(\theta|\mathcal{M}_d, \eta))_{\eta \in E_d}$ and treat η as a frequentist parameter to be estimated by the data. Usually, η is estimated by *maximum marginal likelihood*

$$\hat{\eta} \in \underset{\eta \in E_d}{\arg\max} \log p(\mathcal{D}|\mathcal{M}_d, \eta), \tag{2.8}$$

but other estimation procedures (like the method of moments) can be used. The prior $p(\theta|\mathcal{M}_d, \hat{\eta})$ is eventually chosen. While choosing such a data-dependent prior might be disconcerting, it can be seen as an approximation to a fully Bayesian approach that would

use a prior distribution for η (MacKay, 1994, Section 6.3). Moreover, it leads to very good empirical and theoretical performances in several contexts, such as linear regression (Cui and George, 2008,Liang et al., 2008, and Chapter 3 of this thesis) or principal component analysis (Chapter 4). In a sense, the empirical Bayes maximization problem is equivalent to performing continuous model selection by contemplating E_d as the model space. It is sometimes possible to avoid performing maximum marginal likelihood for each model by averaging over all models: this technique is referred to as glocal empirical Bayes (Liang et al., 2008).

EXAMPLE: THE JEFFREYS-LINDLEY PARADOX FOR PREDICTING OZONE CONCENTRATION Considering the Ozone data set of Chambers et al. (1983), we wish to predict daily ozone concentration in New York city using three explanatory variables: wind speed, maximum temperature, and solar radiation. For this purpose, we use linear regression with Zellner's (1986) g prior. As usual in a variable selection framework, we index the model space using a binary vector $\mathbf{v} \in \{0,1\}^3$ which indicates which variables are deemed relevant in model $\mathcal{M}_{\mathbf{v}}$. We denote by $\mathbf{Y}$ the vector of observed concentrations, and by $\mathbf{X}$ the matrix of explanatory variables. There are $\#\{0,1\}^3 = 8$ models, defined by

$$\mathcal{M}_{\mathbf{v}} : \mathbf{Y} = \mu \mathbf{1}_p + \mathbf{X}_{\mathbf{v}} \boldsymbol{\beta}_{\mathbf{v}} + \boldsymbol{\varepsilon}, \tag{2.9}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \phi^{-1}\mathbf{I}_n)$. As in Liang et al. (2008), we consider the following prior distributions:

$$p(\mathcal{M}_{\mathbf{v}}) \propto 1,\ p(\mu, \phi | \mathcal{M}_{\mathbf{v}}) \propto \frac{1}{\phi},\ \text{and}\ \boldsymbol{\beta}_{\mathbf{v}} | \phi, \mathcal{M}_{\mathbf{v}} \sim \mathcal{N}\left(0, \frac{g}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}\right). \tag{2.10}$$

The prior distribution of μ and ϕ is improper, but this is acceptable because both parameters appear in all models (which is not the case for $\boldsymbol{\beta}_{\mathbf{v}}$). When g becomes large, the prior distribution of $\boldsymbol{\beta}_{\mathbf{v}}$ becomes very flat, and the Jeffreys-Lindley paradox comes into play. To get a grasp of this phenomenon, we may look at the *posterior inclusion probabilities* for all three variables, defined as the posterior probability that the corresponding coefficient is nonzero (Figure 2.1). When g is very large, Bayesian model uncertainty suggests that all three variables are useless. Three popular ways of automatically choosing g are also displayed. As we can see, using any of these reasonable choices allows to get very far from the Jeffreys-Lindley regime.

While we separated here for clarity the problems of finding priors over model space and parameters, it is worth mentioning that several interesting works considered the *simultaneous* specification of these priors (Dawid, 2011; Dellaportas et al., 2012). Let us finish this subsection by quoting Jordan (2011), then president of the International Society for Bayesian Analysis, summarizing a survey he conducted across several senior statisticians regarding important open problems in Bayesian statistics,

> *Many people feel that prior specification for model selection is still wide open.*

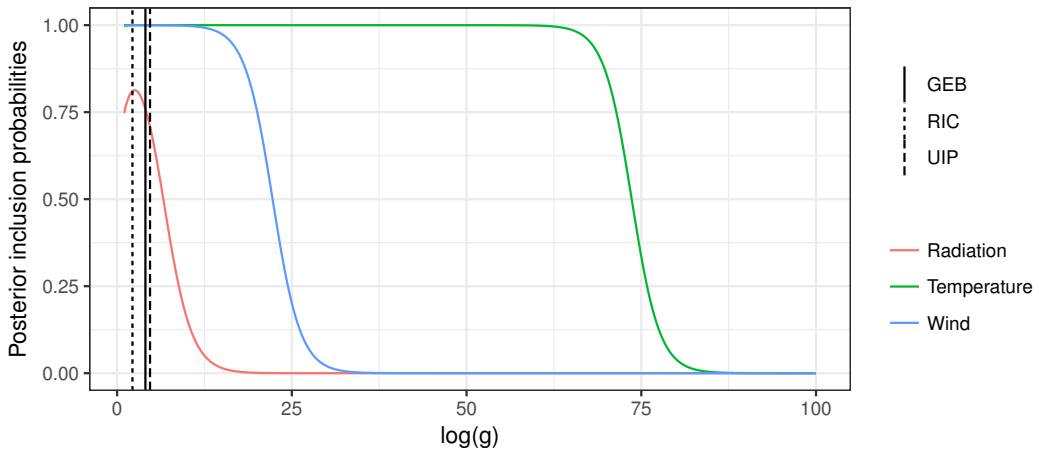


Figure 2.1 – The Jeffreys-Lindley paradox for linear regression with g-priors for the Ozone data set. As g becomes very large, the prior distribution of the regression vector becomes less and less informative, leading to the progressive dismissal of all three explanatory variables. Three proposals of automatic determination of g are also displayed: global empirical Bayes (GEB), risk information criterion (RIC) and unit information prior (UIP, see e.g. Liang et al., 2008, for more details on these three techniques).

2.2.4 Theoretical guarantees of Bayesian model uncertainty

Theoretical guarantees of model selection schemes can fall within several frameworks. The two main criteria at play are usually *the modeling assumptions* (is there a "true model" or not?) and *the nature of the guarantees* (asymptotic or not? predictive of explanatory?).

2.2.4.a Is finding the ``true model'' desirable ?

In most practical cases, it appears unrealistic to assume that one model within the available family did actually generate the data. However, this assumption is commonly made when statisticians assess the performance of a model selection scheme. A reason for this is that, in the (overly optimistic) framework where there actually were a true model, we would want a good model selection technique to find it. Or, to quote Liang et al. (2008),

> *While agreeing that no model is ever completely true, many (ourselves included) do feel it is useful to study the behavior of procedures under the assumption of a true model.*

This "good behavior in the best case scenario" framework is sometimes considered pointless as this "best case scenario" is too unrealistic (see e.g. Gelman and Shalizi, 2013, or Spiegelhalter et al., 2014). Although we believe that the true model assumption can be of interest, we will not focus on theoretical results that rely on it in this section.

Note however that, sometimes, the true model assumption can be completely valid. This is for example the case in physics, for example if one wants to choose between Newtonian gravitation or Einstein's general relativity (Jefferys and Berger, 1992).

2.2.4.b Asymptotics

Traditionally, theoretical model selection guarantees aim at ensuring that, in the long run (when n goes to infinity), the studied technique gives a high probability to the best model. If there is no true model, the closest model in the Kullback-Leibler sense if often considered. This property is usually called *model selection consistency*. For recent perspectives on the subject, we defer the reader to Liang et al. (2008) regarding linear regression, Chatterjee et al. (2017) for non independent data, and Walker (2004), Dawid (2011) and Chib and Kuffner (2016) for broad reviews. An interestingly growing point of view is the high-dimensional scenario where it is assumed that both the number of variables and the number of observations grow to infinity (see e.g. Moreno et al., 2015, Barber et al., 2016).

2.2.4.c Out-of-sample performance

In this review, while acknowledging the usefulness of both asymptotics and the true model assumption, we wish to focus on the recent findings of Germain et al. (2016), which insure that Bayesian model selection allows to find optimal models from a predictive perspective, *even in non-asymptotic settings and when the true model is not in the family.*

Germain et al. (2016) established an important bridge between the (essentially frequentist) PAC-Bayesian theory and Bayesian model selection. The PAC-Bayesian theory, introduced by Shawe-Taylor and Williamson (1997) and McAllester (1998) and championed by Catoni (2007), aims at finding non-asymptotic *probably approximately correct* (PAC) bounds on the generalization error of a machine learning algorithm. As we will see, the PAC-Bayesian machinery also allows to find bounds on the *predictive log-likelihood* (that is, the likelihood of new, unseen data) of a Bayesian model.

In the following, we consider a supervised setting where we are dealing with n i.i.d. copies $\mathcal{D} = (X, Y) = (x_i, y_i)_{i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ of a random variable $(x, y) \sim p_{\text{data}}$. Since it is not assumed that the data-generating model lies within the family, prior model probabilities are assumed to be chosen as scores of prior usefulness of the models, and posterior model probabilities cannot eventually be seen as probabilities that the models are true (see Section 2.2.2). The predictive log-likelihood function, defined for a given model $\mathcal{M}_d$ as, for all $\theta \in \Theta_d$,

$$\mathcal{L}(\theta|\mathcal{M}_d) = \mathbb{E}_{x,y}[\log p(y|x, \theta, \mathcal{M}_d)], \tag{2.11}$$

will be the quantity of interest, as it allows to asses the out-of-sample performance of a model. We will also look at the BMA predictive log-likelihood, defined as, for all $(\theta_1, ..., \theta_{d_{\max}}) \in \Theta_1 \times ... \times \Theta_{d_{\max}}$,

$$\mathcal{L}_{\text{BMA}}(\theta_1, ..., \theta_{d_{\max}}) = \sum_{d=1}^{d_{\max}} p(\mathcal{M}_d|X, Y)\mathbb{E}_{x,y}[\log p(y|x, \theta_d, \mathcal{M}_d)]. \tag{2.12}$$

Although we will not assume that p_{data} lies within the model family, we need to make assumptions on this distribution in order to bound the predictive likelihood. Following, Germain et al. (2016), we will rely on the following *sub-gamma assumption* stated below. For more details on sub-gamma random variables, see e.g. Boucheron et al. (2013, Section 2.4).

Definition 2.1 (Sub-gamma assumption)**.** *A Bayesian model $(p(\cdot|\theta)_{\theta\in\Theta}, \pi)$ of some data $\mathcal{D}$ coming from a distribution p_{data} satisfies the sub-gamma assumption with variance factor $s^2 > 0$ and scale parameter $c > 0$ if the random variable $\log p(\mathcal{D}|\theta) - \mathbb{E}_{\mathcal{D}} \log p(\mathcal{D}|\theta)$ is a sub-gamma random variable, that is that its moment generating function is upper bounded by the one of a Gamma random variable with shape parameter s^2/c^2 and scale parameter c.*

The validity of the sub-gamma assumption, deeply linked to the theory of concentration inequalities, depends on the true distribution of the data. However, it is not necessary to assume that this true distribution belongs to the model family. Knowing which models satisfy this assumption is of paramount importance, and should be the subject of future work. Germain et al. (2016) showed that the linear regression model, for example, satisfies the sub-gamma assumption.

We can now provide out-of-sample guarantees for Bayesian inference under model uncertainty.

Theorem 2.1 (Germain et al., 2016, Corollary 6)**.** *If the expanded model satisfies the sub-gamma assumption with variance $s^2 > 0$ and scale $c > 0$, we have, with probability at least $1 - 2\delta$ over the data-generating distribution,*

$$\mathbb{E}_{\theta_1,\dots,\theta_{d_{\max}}}[\mathcal{L}_{\text{BMA}}(\theta_1, ..., \theta_{d_{\max}})|\mathcal{D}] \geq \frac{1}{n} \log \left(\sum_{d=1}^{d_{\max}} p(Y|X, \mathcal{M}_d) p(\mathcal{M}_d) \delta \right) - \frac{s^2}{2(1-c)}, \quad (2.13)$$

and, for each $d \in \{1, ..., d_{\max}\}$,

$$\mathbb{E}_{\theta}[\mathcal{L}(\theta|\mathcal{M}_d)|\mathcal{D}] \geq \frac{1}{n} \log \left(p(Y|X, \mathcal{M}_d) p(\mathcal{M}_d) \delta \right) - \frac{s^2}{2(1-c)}. \quad (2.14)$$

This theorem has two important non-asymptotic implications.

- Among the family at hand, *the model with the largest marginal likelihood is the one endowed with the strongest PAC guarantees.* This gives strong theoretical support for the predictive empirical successes of Bayesian model selection, especially in small-sample scenarios (MacKay, 1992b; Murphy et al., 2010; Celeux et al., 2012).
- Since the bound obtained using the BMA posterior is tighter, *BMA has stronger PAC guarantees than the best model of the family.* Again, this explains the well-established empirical result that BMA outperforms model selection from a predictive perspective (Hoeting et al., 1999; Raftery et al., 1996, 2005; Piironen and Vehtari, 2016). Note that several results on the superiority of BMA have been presented in the past, but usually relied on the fact that the quantity of interest was actually distributed according to the BMA posterior (Raftery and Zheng, 2003).

The BMA bound offers guarantees regarding the model averaged log-likelihood. However, in a forecasting context, it is often seen as more relevant to look at the logarithm of the BMA posterior of the response $p(y|x, \mathcal{D})$, as defined in (2.6). Indeed, this criterion corresponds to the logarithmic score of Good (1952), a strictly proper scoring rule widely used to assess the

quality of probabilistic forecasts (Gneiting and Raftery, 2007). Using Jensen's inequality, this quantity can be bounded directly using (2.13):

$$\begin{aligned}
\mathbb{E}_{x,y}[\log p(y|x,\mathcal{D})] &= \mathbb{E}_{x,y}\left[\log\left(\sum_{d=1}^{d_{\max}} p(\mathcal{M}_d|X,Y)\mathbb{E}_{\theta_d}[p(y|x,\theta_d,\mathcal{M}_d)|\mathcal{D}]\right)\right] \\
&\geq \mathbb{E}_{x,y}\left[\sum_{d=1}^{d_{\max}} p(\mathcal{M}_d|X,Y)\log\mathbb{E}_{\theta_d}[p(y|x,\theta_d,\mathcal{M}_d)|\mathcal{D}]\right] \\
&\geq \mathbb{E}_{x,y}\left[\sum_{d=1}^{d_{\max}} p(\mathcal{M}_d|X,Y)\mathbb{E}_{\theta_d}[\log p(y|x,\theta_d,\mathcal{M}_d)]\right] \\
&\geq \mathbb{E}_{\theta_1,\ldots,\theta_{d_{\max}}}[\mathcal{L}_{\text{BMA}}(\theta_1,\ldots,\theta_{d_{\max}})|\mathcal{D}].
\end{aligned}$$

This gives a new interpretation to the results of Germain et al. (2016). If we compare all models and BMA using the logarithmic scoring rule, then BMA predictions have stronger guarantees than the model with the largest posterior probability, which has itself stronger guarantees than all other models within the family. A related result was obtained by Madigan and Raftery (1994) under the strong assumption that $y|x$ exactly follows the BMA posterior.

WHAT ABOUT POINT ESTIMATION ? This PAC theorem gives guarantees on the posterior expectation of the predictive log-likelihood. However, it is often of interest to have guarantees about point estimates of θ. For each model $d \in \{1,\ldots,d_{\max}\}$, let us consider the posterior mean $\hat{\theta}_d = \mathbb{E}_\theta[\theta|\mathcal{D},\mathcal{M}_d]$, a popular Bayesian point estimate, notably because of its decision theoretic optimality under the squared loss (Berger, 1985, Section 2.2.4). If we assume that the log-likelihood function is a concave function of θ, Jensen's inequality implies that

$$\mathbb{E}_{x,y}[\log p(y|x,\hat{\theta}_d,\mathcal{M}_d)] \geq \mathbb{E}_{\theta,x,y}[\log p(y|x,\theta,\mathcal{M}_d)|\mathcal{D}], \tag{2.15}$$

which means that the predictive likelihood evaluated at $\hat{\theta}_d$ will inherit the good PAC properties of the posterior predictive likelihood bounded in (2.14). Similarly, BMA forecasts obtained with point estimates satisfy

$$\mathbb{E}_{x,y}\left[\log\left(\sum_{d=1}^{d_{\max}} p(y|x,\hat{\theta}_d,\mathcal{M}_d)p(\mathcal{M}_d|\mathcal{D})\right)\right] \geq \mathcal{L}_{\text{BMA}}(\hat{\theta}_1,\ldots,\hat{\theta}_{d_{\max}}) \tag{2.16}$$

$$\geq \mathbb{E}_{\theta_1,\ldots,\theta_{d_{\max}}}[\mathcal{L}_{\text{BMA}}(\theta_1,\ldots,\theta_{d_{\max}})|\mathcal{D}]. \tag{2.17}$$

This theorem offers some strong theoretical insight on why Bayesian model uncertainty works well in a predictive setting. However, one could argue that its merit is merely conceptual. Indeed, the fact that the bounds depend on the data-generating distribution makes them very hard to compute in practice. A good sanity check that was conducted by Germain et al. (2016) is to assess the tightness of the bound for some known model. Specifically, they considered the linear regression model (which is sub-gamma for some known scale and variance parameters) and observed that the bound was indeed tight.

Actually, Theorem 2.1 also has some important practical applications. Indeed, although the bounds themselves depend on unknown sub-gamma parameters, *the differences between*

Table 2.1 – Estimating the BMA gain for linear regression. Estimate obtained from the PAC bounds versus actual out-of-sample MSE gain obtained with model averaging. Results are averaged over 500 random replications with balanced training/test splits.

	Prostate	US crime	Housing	Ozone	Auto
$-(2\hat{\sigma}^2/n)\max_d \log p(Y\|X,\mathcal{M}_d)$	3.39×10^{-2}	1.01×10^{4}	3.46×10^{-3}	7.47	7.69×10^{-2}
Actual out-of-sample MSE gain	3.23×10^{-2}	2.89×10^{4}	4.69×10^{-3}	9.40	4.92×10^{-2}

bounds associated with different models can be computed according to their posterior odds. Indeed, the difference between bounds associated with models $\mathcal{M}_d$ and $\mathcal{M}_{d'}$ is exactly

$$\frac{1}{n}\log\left(\frac{p(\mathcal{M}_d|\mathcal{D})}{p(\mathcal{M}_{d'}|\mathcal{D})}\right).$$

In case of a uniform prior probabilities over models, the difference is $n^{-1}\log \mathrm{BF}_{d/d'}$. This gives a new, predictive, interpretation of the Bayes factors as a measure of evidence in favor of a model. If all bounds are tight, this also gives a good estimate of the generalization gain proposed by a certain model.

An important consequence of this is that it provides a way to quantify the benefits of BMA over model selection. In the discussion on the BMA tutorial of Hoeting et al. (1999), Draper (1999) asked "what characteristics of a statistical example predict when BMA will lead to large gains?". While suggesting to perform BMA when the ratio n/p is small, Draper (1999) insisted on the need of more refined simple rules that will quantify the relevance of performing BMA over model selection. Such a rule can be derived using the PAC bounds. Indeed, the difference between the PAC bound of the BMA posterior and the one of the model with the largest marginal likelihood is exactly $-(1/n)\max_d \log p(Y|X,\mathcal{M}_d)$ which means that the benefits of averaging will be less important when the posterior probability of the best model is close to one. While this consideration is unsurprising, another more important consequence is that $-(1/n)\max_d \log p(Y|X,\mathcal{M}_d)$ can be seen as a good estimate of the gain of performing model averaging.

EXAMPLE: HOW USEFUL IS AVERAGING FOR LINEAR REGRESSION ? Consider the Gaussian linear regression model. The usual performance criterion is the mean squared prediction error (MSE), of which the likelihood is the simple affine transformation $\log(2\pi\sigma) - 1/(2\sigma^2)\mathrm{MSE}$. According to the PAC bounds, a rough estimate of the *out-of-sample* mean squared error difference between the predictions of the highest probability model and the model averaged ones can be given by

$$\mathrm{MSE}(\text{model selection}) - \mathrm{MSE}(\mathrm{BMA}) \approx -(2\hat{\sigma}^2/n)\max_{d\in\{1,\ldots,d_{\max}\}} \log p(Y|X,\mathcal{M}_d)$$

where $\hat{\sigma}$ is an estimate of the residual standard error. We assess the accuracy of this estimate using five standard linear regression data sets (Table 2.1) and the hyper-g-n priors of Liang et al. (2008). Interestingly, this rough estimate consistently gives a pretty good idea of the gain of performing BMA, and can be seen as a good indicator of whether or not BMA can be useful.

2.2.5 Links with penalized model selection

Both Bayesian and penalty-based approaches build on the likelihood function to perform model selection: while the former *integrates it*, the latter *maximizes it and adds a penalty*. It appears natural to seek foundational connections between these two likelihood treatments.

KULLBACK-LEIBLER PENALIZATION A simple penalized view of Bayesian model selection can be derived as follows. For some model $\mathcal{M}_d$ associated to a parameter space Θ_d, let us rewrite the log-marginal likelihood as

$$\begin{aligned}
\log p(\mathcal{D}|\mathcal{M}_d) &= \frac{p(\mathcal{D}|\mathcal{M}_d)}{p(\mathcal{D}|\mathcal{M}_d)} \log p(\mathcal{D}|\mathcal{M}_d) && (2.18)\\
&= \int_{\Theta_d} \frac{p(\mathcal{D}|\theta,\mathcal{M}_d)p(\theta|\mathcal{M}_d)}{p(\mathcal{D}|\mathcal{M}_d)} \log p(\mathcal{D}|\mathcal{M}_d) d\theta && (2.19)\\
&= \int_{\Theta_d} p(\theta|\mathcal{D},\mathcal{M}_d) \left(\log p(\theta|\mathcal{D},\mathcal{M}_d) + \log \frac{p(\theta|\mathcal{M}_d)}{p(\theta|\mathcal{D},\mathcal{M}_d)} \right) d\theta && (2.20)\\
&= \mathbb{E}_\theta[\log p(\mathcal{D}|\theta,\mathcal{M}_d)|\mathcal{D}] - \mathrm{KL}(p(\cdot|\mathcal{M}_d,\mathcal{D})||p(\cdot|\mathcal{M}_d)). && (2.21)
\end{aligned}$$

This means that maximizing the marginal likelihood can be seen as maximizing a penalized version of the posterior mean of the log-likelihood. The penalty term is simply the Kullback-Leibler divergence between the prior and the posterior, and will arguably penalize complex models in a finer way than penalties based on the number of parameters (Seeger, 2003; Zhang, 2006). Interestingly, this decomposition shows that choosing a too noninformative prior distribution (such as a Gaussian with very large variance) will lead to an explosion of the Kullback-Leibler term, and to overpenalizing the likelihood, thus choosing a perhaps too simple model. This gives an interpretation of the Jeffreys-Lindley paradox described in Section 2.2.3.b as an overpenalization phenomenon. Similar model selection schemes based on penalized versions of the posterior mean of the likelihood $\mathbb{E}_\theta[\log p(\mathcal{D}|\theta,\mathcal{M}_d)|\mathcal{D}]$ have been used in the past. Under the general setting

$$\mathrm{score}(\mathcal{D},\mathcal{M}_d) = \mathbb{E}_\theta[\log p(\mathcal{D}|\theta,\mathcal{M}_d)|\mathcal{D}] - \mathrm{pen}(\mathcal{D},\mathcal{M}_d), \tag{2.22}$$

we have the following correspondances:

- $\mathrm{pen}_{\mathrm{PBF}}(\mathcal{D},\mathcal{M}) = 0$ corresponds to the *posterior Bayes factors* of Aitkin (1991).
- $\mathrm{pen}_{\mathrm{A\&T}}(\mathcal{D},\mathcal{M}) = np/2$ corresponds to an estimator of the posterior predictive likelihood proposed by Ando and Tsay (2010). Note that Ando and Tsay (2010) also proposed a refined criterion that falls within the general setup of (2.22), but whose formula is much more complex.
- $\mathrm{pen}_{\mathrm{DIC}}(\mathcal{D},\mathcal{M}) = \log p(\mathcal{D}|\hat{\theta},\mathcal{M}_d)/2$, where $\hat{\theta}$ is the posterior mean estimate, is equivalent to the *deviance information criterion* (DIC) of Spiegelhalter et al. (2002).
- $\mathrm{pen}_{\mathrm{WAIC}_1}(\mathcal{D},\mathcal{M}_d) = \log(\mathbb{E}_\theta[p(\mathcal{D}|\theta,\mathcal{M}_d)|\mathcal{D}])/2$ and

  $$\mathrm{pen}_{\mathrm{WAIC}_2}(\mathcal{D},\mathcal{M}_d) = 2(\log(\mathbb{E}_\theta[p(\mathcal{D}|\theta,\mathcal{M}_d)|\mathcal{D}]) - \mathbb{E}_\theta[\log p(\mathcal{D}|\theta,\mathcal{M}_d)|\mathcal{D}]))$$

  are equivalent to two versions of the *widely applicable information criterion* (WAIC) of Watanabe (2009, Section 8.3).

- the *Bayesian predictive information criterion* BPIC of Ando (2007) uses a complex penalty $\text{pen}_{\text{BPIC}}(\mathcal{D}, \mathcal{M})$.

Several of these frameworks were specifically designed to estimate the posterior mean of the predictive log-likelihood function, which is exactly the quantity bounded by the PAC theorem of Germain et al. (2016). Even though pen_{BPIC} and $\text{pen}_{\text{WAIC}_2}$ lead to asymptotically unbiased estimates of this quantity, the Kullback-Leibler penalty automatically entangled with Bayesian model selection is, to the best of our knowledge, the only framework that provides strong guarantees on small-sample behavior. For more insight on the merits of these various penalization schemes, and their links with cross-validation, see Plummer (2008).

REMARK: WHY IS IT NECESSARY TO PENALIZE THE POSTERIOR MEAN OF THE LIKELIHOOD ? If we want to maximize the posterior predictive log likelihood, it seems natural to maximize the posterior mean of the log likelihood, which can be seen as an *empirical* estimate of our target. Similarly to the theory of empirical risk minimization, it is customary to add a penalty to this empirical estimate to avoid overfitting. From a Bayesian point of view, this necessity can be interpreted as follows. When we compute the posterior mean

$$\mathbb{E}_\theta[\log p(\mathcal{D}|\theta, \mathcal{M}_d)|\mathcal{D}], \tag{2.23}$$

we use the same data *twice* (to find the posterior distribution and to compute the likelihood inside the expectation), which is not consistent with the Bayesian approach. Aitkin (1991), who suggested an unpenalized use of the posterior mean of the likelihood, was criticized by several of his discussants because of this double use of the same data. As explained by Plummer (2008), the penalty is what "must be paid for using the data (...) twice".

MACKAY'S OCCAM RAZOR INTERPRETATION In his thesis and subsequent work, MacKay (1991, 1992a, 2003), inspired by Gull (1988), drew interesting connections between penalized maximum likelihood methods and Bayesian model uncertainty. The first step is to look at a Laplace approximation of the marginal likelihood. For i.i.d. data, we have, under some (unfortunately not so mild) regularity conditions that we discuss in Section (2.3.3.b)

$$\log p(\mathcal{D}|\mathcal{M}) = \underbrace{\log p(\mathcal{D}|\hat{\theta}, \mathcal{M})}_{\text{Maximized likelihood}} \underbrace{+ \log p(\hat{\theta}|\mathcal{M}) + \frac{p}{2}\log 2\pi - \frac{1}{2}\log\det A + O_p\left(\frac{1}{n}\right)}_{\text{Occam factor}} \tag{2.24}$$

where $\hat{\theta}$ is either the maximum a posteriori or the maximum-likelihood estimator of θ (in the first case the $p \times p$ matrix A is the Hessian of the log posterior, in the latter it is the observed information matrix). This means that, in the long run, *Bayesian model selection is approximatively equivalent to a form of automatically penalized maximum likelihood.* This automatically designed penalty was called the *Occam factor* by Gull (1988). It essentially depends on the prior distribution and on the "complexity" of the model. In some simple scenarios like linear regression, the Occam factor can directly be linked to the number of parameters (see Chapter 3) – this builds a direct bridge with ℓ_0 penalization. However, it is not always the case and the Occam factor penalty provides a more sensible regularization

than those based on the number of parameters (Rasmussen and Ghahramani, 2001). For a deeper interpretation of the Occam factor penalty, see MacKay (2003, p. 349). Mackay's other important insight is a graphical interpretation of this Occam razor effect. Assume for simplicity that there are only two models, one simple ($\mathcal{M}_1$) and one more complex ($\mathcal{M}_2$). The key idea is to plot the marginal distributions of the data $p(\mathcal{D}|\mathcal{M}_d)$ (seen as functions of $\mathcal{D}$) using an idealized unidimensional $\mathcal{D}$-axis where "simple" data sets are located near the center of the plot (Figure 2.2). On the one hand, the complex model will be able to provide good fits to a larger range of data sets, and the corresponding marginal distribution $p(\mathcal{D}|\mathcal{M}_d)$ will consequently be flatter. On the other hand, the simpler model will concentrate its mass around a limited number of data sets, leading to a more peaky marginal distribution. If the data comes from the $\mathcal{C}_1$ region of MacKay's plot, then the simpler model will have a larger evidence, even though it might not fit the data as well. This illustrates the automatic "Occam's razor effect" of Bayesian model uncertainty. As described in Section 2.2.3.b, another Occam's razor effect can be added by following the simplicity postulate and giving more prior probability to simpler models.

EXAMPLE: MACKAY'S PLOT FOR A SINGLE GAUSSIAN OBSERVATION We propose to plot a simple instance of MacKay's plot (Figure 2.2). Consider the case where the data consists in a single Gaussian observation $x \sim \mathcal{N}(\theta^*, 1)$ with unit variance. We wish to know whether $\theta^* = 0$. The two models are

$$\mathcal{M}_1 : x \sim \mathcal{N}(0, 1),$$

and

$$\mathcal{M}_2 : x|\theta \sim \mathcal{N}(\theta, 1),\ \theta \sim \mathcal{N}(0, s^2),$$

leading to the marginal distributions

$$x|\mathcal{M}_1 \sim \mathcal{N}(0, 1) \text{ and } x|\mathcal{M}_2 \sim \mathcal{N}(0, 1 + s^2).$$

The more complex model $\mathcal{M}_2$ will *always* provide a better fit to the data. But if x is small enough, i.e. in the region

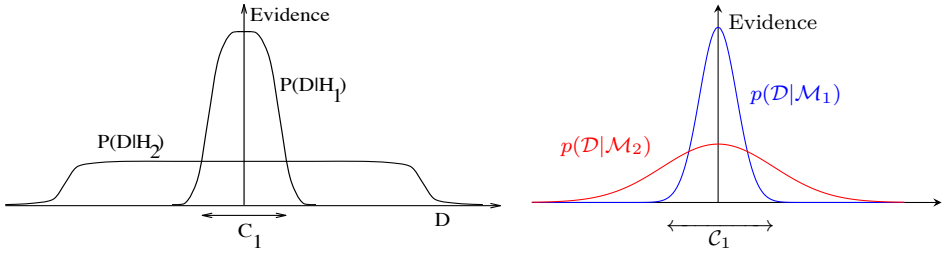
$$\mathcal{C}_1 = \left[-\left(1 - \frac{1}{1+s^2}\right) \log(1+s^2), \left(1 - \frac{1}{1+s^2}\right) \log(1+s^2) \right], \tag{2.25}$$

then the simpler zero-mean model will have a larger marginal likelihood. This illustrates the Occam's razor effect. Note that, when s goes to infinity, $\mathcal{C}_1$ becomes infinitely wide, which means that $p(\mathcal{D}|\mathcal{M}_1)$ is everywhere above $p(\mathcal{D}|\mathcal{M}_2)$. In this limiting case, the simpler model will always be preferred: once again, this is an instance of the Jeffreys-Lindley paradox (see Section 2.2.3.b). Another concrete example of MacKay's plot was given (in a discrete setting) by Murray and Ghahramani (2005).

2.3 The practice of Bayesian model uncertainty

In this section, we review the computational strategies that allow to set Bayesian model uncertainty in motion.

Figure 2.2 – MacKay's Occam razor plot. *Left:* Mackay's idealized plot, reproduced from MacKay (2003, p. 344). *Right:* MacKay's plot for a single Gaussian observation.



2.3.1 Computing marginal likelihoods

As explained in the previous section, the posterior probabilities of a model $\mathcal{M}_d$ can be computed using its marginal likelihood

$$p(\mathcal{D}|\mathcal{M}_d) = \int_{\Theta_d} p(\mathcal{D}|\theta, \mathcal{M}_d) p(\theta|\mathcal{M}_d) d\theta. \tag{2.26}$$

This quantity is therefore of paramount importance to account for model uncertainty. Unfortunately, as a potentially high-dimensional integral, it is often very difficult to compute exactly. Several approximation schemes have been developed accordingly. However, closed-form calculation of the marginal likelihood is sometimes feasible. While classical examples include multivariate Gaussian data (see e.g. Murphy, 2007) or linear regression (see e.g. Bishop, 2006, Section 3.5.1, or Marin and Robert, 2014, Section 3.4.3), more complex models have also been tackled recently, such as factor analysis (Ando, 2009), mixtures of independence models (Lin et al., 2009), two-sample nonparametric tests (Holmes et al., 2015) and principal component analysis (Chapters 4 and 5).

2.3.2 Markov chain Monte Carlo methods

Markov chain Monte Carlo methods (MCMC), the Swiss Army knife of modern Bayesian analysis, has been extensively apply to the calculation of marginal likelihoods, posterior odds, or Bayes factors. There approaches fall into two categories.

- *Within-model mehods* directly attack the marginal likelihoods of models using MCMC. These techniques include notably importance sampling and its variants (e.g. Neal, 2001), nested sampling (Skilling, 2006), power posteriors (Friel and Pettitt, 2008), and schemes based on the harmonic mean identity (e.g. Weinberg, 2012). Recent reviews devoted to this line of work are given by Robert and Wraith (2009), Marin and Robert (2010), and Friel and Wyse (2012).

- *Transdimensional methods* pioneered by Carlin and Chib (1995) and by Green's (1995) reversible jump framework, aim at obtaining samples from the posterior distribution over both models and parameters. Good reviews are provided by Sisson (2005) and Hastie and Green (2012). See also Hee et al. (2016) for recent perspectives.

For insightful comparisons between these two approaches, see Chen et al. (2000), Han and Carlin (2001), and Clyde and George (2004, Section 5). A important issue with both approaches is their limited availability in high-dimensional settings. Indeed, in these cases, the parameter space is too vast to be visited properly and MCMC integration becomes more challenging.

2.3.3 A little help from asymptotics

Computing marginal likelihoods, either exactly or using MCMC, is challenging. However, large-sample theory can also provide an interesting guide to build marginal likelihood approximations.

2.3.3.a The Laplace approximation and BIC

Recall the Laplace approximation of the marginal likelihood of Section 2.2.5:

$$\log p(\mathcal{D}|\mathcal{M}_d) = \log p(\mathcal{D}|\hat{\theta}, \mathcal{M}_d) + \log p(\hat{\theta}|\mathcal{M}_d) + \frac{\dim \Theta_d}{2} \log 2\pi - \frac{1}{2} \log \det A + O_p\left(\frac{1}{n}\right), \tag{2.27}$$

where $\hat{\theta}$ is either the maximum a posteriori or the maximum-likelihood estimator of θ (in the first case the $\dim \Theta_d \times \dim \Theta_d$ matrix A is the Hessian of the log posterior, in the latter it is the observed information matrix, evaluated at $\hat{\theta}$). When this approximation is valid, a $O_p(1/n)$ approximation of the marginal likelihood can be computed using simply the maximized likelihood and the observed information matrix. Actually, this rationale leads to even simpler approximations. Indeed, approximating the observed information matrix by n times the Fisher information matrix and dropping all the terms that are $O_p(1)$, we end up with

$$\log p(\mathcal{D}|\mathcal{M}_d) = \log p(\mathcal{D}|\hat{\theta}, \mathcal{M}_d) - \frac{\dim \Theta_d}{2} \log n + O_p(1). \tag{2.28}$$

The crude marginal likelihood proxy $\log p(\mathcal{D}|\hat{\theta}, \mathcal{M}_d) - (\dim \Theta_d/2) \log n$ involved in equation (2.28) was first derived by Schwarz (1978) and extended by Haughton (1988), who also proved that it produces a consistent model selection procedure. From this approximation, an information criterion similar to AIC can be derived, leading to the popular *Bayesian information criterion* (BIC)

$$\mathrm{BIC}(\mathcal{D}, \mathcal{M}_d) = -2 \log p(\mathcal{D}|\hat{\theta}, \mathcal{M}_d) + \dim \Theta_d \log n. \tag{2.29}$$

The BIC has the practical advantage that its off-the-shelf expression does not involve the prior distribution whatsoever, at the price of producing a rough $O_p(1)$ approximation of the marginal likelihood. However, the BIC actually corresponds to an implicit prior distribution. Indeed, assuming that the prior distribution of θ is a specific data-dependent prior, it is possible to show that Schwarz's proxy actually provides $O_p(1/\sqrt{n})$ approximation of the marginal likelihood (Kass and Wasserman, 1995; Raftery, 1995). This prior distribution, called the *unit information prior* (UIP), can be interpreted as a weakly informative prior based on an imaginary sample of one observation. For discussions on the merits and dangers of using the UIP or the BIC, see Weakliem (1999), Raftery (1999), and Kuha (2004).

2.3.3.b Towards singular asymptotics

We remained voluntarily laconic regarding the regularity conditions for the Laplace approximation (2.27) to be valid. Several of them are of importance. For thorough theoretical treatments of these conditions, see Haughton (1988) and Kass et al. (1990). We choose here to give details on the conditions that are the most often violated in practice.

First, it is assumed that $\hat{\theta}$ is an interior point of Θ_d. This can be an important issue in many cases (consider for instance a scale parameter, or g in a g-prior). Several solutions have been proposed to efficiently tackle this issue (Erkanli, 1994; Hsiao, 1997; Pauler et al., 1999).

Moreover, it is assumed that the Fisher information matrix in invertible. This condition is unfortunately violated in non-identifiable models, which are becoming ubiquitous in statistical inference. Such models, often called *singular models*, include mixture models, factor analysis, probabilistic principal component analysis, hidden Markov models, deep neural networks or reduced-rank regression. In these cases, the Laplace approximation is invalid and more refined asymptotic theory has to be invoked. As first exhibited by Watanabe (1999), algebraic geometry proves extremely useful in this context. Specifically, for a wide variety of singular models, a BIC-like approximation was derived by Watanabe (2009, Theorem 6.7),

$$\log p(\mathcal{D}|\mathcal{M}_d) = \log p(\mathcal{D}|\hat{\theta}, \mathcal{M}_d) - \lambda_d \log n + (m_d - 1) \log\log n + O_p(1), \tag{2.30}$$

where λ_d is a positive rational number called the *learning coefficient* (also known as the *real log canonical threshold* in the algebraic geometry literature) and m_d is a natural number called the *multiplicity* of λ_d. For regular models, the learning coefficient is simply equal to $\dim(\Theta_d)/2$ and its multiplicity is one, which means that Watanabe's result reduces to the BIC approximation. However, for singular models, the couple (λ_d, m_d) is an often difficult to compute quantity that depends on the true data generating distribution. A major caveat is therefore that, for (2.30) to be used, the true model (which is precisely what we are looking for) has to be known beforehand. This would seems to lead to some inextricable circular reasoning problem. To tackle this issue Watanabe (2013) proposed to combine his BIC-like approximation with thermodynamic integration (see also Friel et al., 2017). A fully deterministic solution was also provided by Drton and Plummer (2017) who got around the circular reasoning problem by averaging over different learning coefficients. They defined a *singular Bayesian information criterion* (sBIC) as the solution of a well-posed fixed-point problem (Drton and Plummer, 2017, Definition 1). This new criterion has several merits. First, it is a deterministic and computationally cheap $O_p(1)$ approximation of the marginal likelihood that reduces to the BIC when the model is regular, and is still valid when the model is singular. For these reasons, it can be considered a valid generalization of the BIC.

REMARK: THE SINGULAR OCCAM FACTOR AND THE PREDICTIVE POWER OF SINGULAR MODELS Generalizing MacKay's Occam factor rationale described in (2.31) to singular models leads to the following asymptotic decomposition of the marginal likelihood:

$$\log p(\mathcal{D}|\mathcal{M}_d) = \underbrace{\log p(\mathcal{D}|\hat{\theta}, \mathcal{M}_d)}_{\text{Maximized likelihood}} + \underbrace{(-\lambda_d) \log n + (m_d - 1) \log\log n + O_p(1)}_{\text{Occam factor}}, \tag{2.31}$$

which involves the penalty

$$\text{pen}_{\text{Occam}}(\mathcal{M}_d) = \lambda_d \log n - (m_d - 1) \log \log n.$$

Under relatively mild conditions (see Watanabe, 2009, Theorem 7.2), it can be shown that the learning coefficient will be a rational number in $[0, \dim(\Theta_d)/2]$, with multiplicity in $\{1, ..., \dim \Theta_{d_{\max}}$. Therefore, for singular models, the automatic penalty entangled with Bayesian model selection will be smaller than the regular BIC penalty

$$\text{pen}_{\text{BIC}}(\mathcal{M}_d) = \frac{\dim \Theta_d}{2} \log n.$$

This fact has two interpretations:

- For complex models, the number of parameters gives poor insight on the behavior of Bayesian Occam's razor. A phenomenon studied notably by Rasmussen and Ghahramani (2001).
- Singular models will benefit from their smaller penalties to have potentially larger marginal likelihoods than regular models. Following Germain et al. (2016), let us consider the marginal likelihood as an indicator of predictive performance. In this framework, singular models that fit the data well may therefore have stronger generalization power than regular models in the asymptotic regime. In other words, *singular models may be less prone to overfitting.*

The recent empirical successes of deep neural networks constitute perhaps an interesting instance of this phenomenon. In their most common form, deep neural networks (LeCun et al., 2015; Goodfellow et al., 2016) are models for supervised learning involving a predictor of the form

$$F(x) = \sigma_1 \circ f_1 \circ \sigma_2 \circ f_2 \circ ... \circ f_M(x), \tag{2.32}$$

where $\sigma_1, ..., \sigma_M$ are simple nonlinear pointwise functions chosen beforehand, and $f_1, ..., f_M$ are learnt affine functions. Empirical evidence strongly suggests that, if the number of observations is very large, using an important number M of layers leads to better generalization performance – state-of-the-art visual recognition systems usually involve hundreds of layers (He et al., 2016). However, this hypothesis still has little theoretical foundation, and the generalization prowesses of deep neural networks remain largely mysterious (Zhang et al., 2017). Asymptotic Bayesian model uncertainty provides a heuristic interpretation. While a one-layer network is a regular model, as the number of layers grows, networks become less and less identifiable. Specifically, the Hessian matrix of the log-likelihood of deep networks appears to have many null eigenvalues (Sagun et al., 2017), and at a given number of parameters, deeper networks have fewer degrees of freedom in Ye's (1998) sense (Gao and Jojic, 2016). It appears therefore reasonable to conjecture that the learning coefficient shrinks when the number of layers grows. If this is true, then, for a given number of parameters, a deeper network will have a higher marginal likelihood provided that there is enough data. This might explain why deep learning resists much more efficiently to overfitting than other more traditional techniques.

EXAMPLE: BIC VERSUS SBIC FOR REDUCED-RANK REGRESSION Consider the reduced-rank regression framework, as described by Drton and Plummer (2017). The problem is to linearly predict a multivariate response using some covariate. Each model corresponds to assigning a rank constraint on the regression matrix parameter. Since prediction is the objective, it would appear natural to perform BMA. Given a new covariate value, the BMA estimate of the response is a weighted average of the posterior means obtained for each model. The weights are posterior model probabilities, but are often replaced by BIC-based approximations (Hoeting et al., 1999). However, since this is a singular case, sBIC approximations may be more sensible. To empirically check if this is true, we use three real data sets: "eyedata" (Scheetz et al., 2006), "feedstock" (Liebmann et al., 2009) and "vélibs" (Bouveyron et al., 2015). To obtain multivariate regression problems, the following preprocessing step was used. The variables were ranked according to the unsupervised feature selection technique presented in Chapter 4. The first 20 variables were considered as response and the 30 last were considered as covariates. The data are then split equally between training and test set and the performance is assessed (Table 2.2) using the mean-squared error (MSE). Five estimators are considered: the ordinary least-squares estimator (OLSE) obtained with the full model, OLSEs obtained with models selected by BIC and sBIC, and two BMA estimators. The sBIC-based BMA estimator outperforms all other competitors, illustrating that sBIC provides a more reliable proxy for posterior probabilities than does BIC.

2.3.4 Approximate methods for high-dimensional and implicit models

The last decades have brought about wilder and wilder statistical models. In this subsection, we focus on two kinds of models for which Bayesian model uncertainty is particularly challenging, and has witnessed important advances in recent years: implicit models and high-dimensional models.

2.3.4.a Handling implicit models with likelihood-free inference

The models that have been studied so far are *explicit* in the sense that, given a parameter value, we have full access to a candidate distribution with density $p(\cdot|\theta, \mathcal{M}_d)$ over the data space, leading to the computation of the likelihood function $\theta \mapsto p(\mathcal{D}|\theta, \mathcal{M}_d)$ which plays a major role within the Bayesian machinery. However, more and more attention is devoted to families of models for which the likelihood function is not available. This context arises when, given a parameter θ, rather than knowing the corresponding candidate distribution $p(\cdot|\theta, \mathcal{M}_d)$, *we are merely able to simulate data from* $p(\cdot|\theta, \mathcal{M}_d)$. Often, the nonavailability of

Table 2.2 – BIC versus sBIC for reduced-rank regression: MSE over 1000 replications.

	OLSE	BIC	sBIC	BMA-BIC	BMA-sBIC
eyedata	10.8 (1.07)	8.67 (0.536)	8.67 (0.541)	8.67 (0.536)	**8.60 (0.584)**
feedstock	10.5 (1.72)	10.5 (1.53)	**9.79 (1.42)**	10.4 (1.52)	**9.79 (1.42)**
vélibs	14.9 (0.980)	16.5 (0.672)	14.7 (0.624)	16.4 (0.694)	**14.5 (0.612)**

the likelihood comes from the presence of a latent variable that is difficult to integrate. This is for instance the case of popular population genetics models where genealogical histories are unobserved (see e.g. Tavaré et al., 1997). Other examples include Markov random fields and related models (see e.g. Stoehr, 2017, for a recent review). While the likelihood is extremely hard to compute in these contexts, it also sometimes does not exist whatsoever. This occurs when dealing with *generative adversarial networks* (GANs, Goodfellow et al., 2014), deep learning models that have vastly improved the state-of-the-art in pseudonatural image generation. GANs essentially assume that the data is generated by passing noise through a neural network parametrized by θ. In this case, while it is easy to sample from $p(\cdot|\theta, \mathcal{M}_d)$, this distribution has no density (Arjovsky and Bottou, 2017), which makes the likelihood not only intractable, but nonexistent.

General-purpose inference within implicit models has been subject to much attention, dating at least back to Diggle and Gratton (1984). From a Bayesian perspective, the first important contribution came from population genetics with the seminal paper of Tavaré et al. (1997), who proposed a scheme for drawing samples from an approximation of the posterior distribution. The fruitful line of work that followed (see e.g. Csilléry et al., 2010, for a review of applications and Marin et al., 2012, for a methodological overview) has been called *approximate Bayesian computation* (ABC). While parameter inference in implicit models is already extremely challenging, recent efforts have also been concentrated towards accounting for model uncertainty. Model-specific methodologies has led to efficient schemes for estimating the marginal likelihood in several frameworks, such as exponential random graph models (Friel, 2013; Bouranis et al., 2017). In a more general setting, several techniques have been proposed to estimate posterior model probabilities using the ABC rationale (see e.g. Marin et al., 2015, for a review). In particular, Pudlo et al. (2015) proposed a scalable approach based on Breiman's (2001) random forests. Several papers have also tried to apply variational inference to general implicit models (Huszár, 2017; Tran et al., 2017a,b). Althought model uncertainty was not the primary focus of these works, such variational approaches lead to the computation of lower-bounds of the marginal likelihood (see next subsection), and can be therefore used to approximate posterior model probabilities.

2.3.4.b Handling high-dimensional models with large-scale deterministic inference

Families of high-dimensional models combine two major difficulties when accounting for model uncertainty:

1. The marginal likelihood of a high-dimensional model $\mathcal{M}_d$, as a $\dim \Theta_d$-dimensional integral, might be extremely difficult to compute, especially using MCMC methods.

2. Sparse modeling, which is extremely popular in high-dimensional settings because it can lead to increased interpretability and better performance, usually involves a number of candidate models of order 2^p, where p is the (large) total number of variables. In this setting, it appears impossible to compute posterior probabilities of all models within the family.

We choose to focus here specifically on sparsity, which has arguably constituted the most popular field of statistical research of the last two decades, culminating perhaps with the monograph of Hastie et al. (2015) and Candès's (2014) plenary lecture at the International Congress of Mathematicians (see Chapter 1).

In a sparse modeling context, there is a largest model $\mathcal{M}$ with parameter space Θ within which all other models are embedded. A convenient way to write models in this context in through the use of binary vectors $\mathbf{v} \in \{0,1\}^{\dim \Theta}$ that can index each model $\mathcal{M}_{\mathbf{v}}$, such that

$$\Theta_{\mathbf{v}} = \{\boldsymbol{\theta} \in \Theta | \text{Supp}(\boldsymbol{\theta}) = \mathbf{v}\}. \tag{2.33}$$

Accounting for model uncertainty now all comes down to studying the posterior distribution of this high-dimensional binary vector v, and model selection can be recast as the following discrete optimization problem

$$\mathbf{v}^* = \text{argmax}_{\mathbf{v}} p(\mathcal{D}|\mathbf{v}), \text{ such that } \mathbf{v} \in \{0,1\}^{\dim \Theta}. \tag{2.34}$$

Of course, so far, the problem remains exactly as difficult as before, and both the exact posterior of $\mathbf{v}$ and the best model $\mathbf{v}^*$ remain very difficult to compute because of the large number of models. However, using this formalism, we can now make use of the particular *structure* of the model space $\{0,1\}^{\dim \Theta}$ to efficiently approximate these quantities. There are several ways of building on this structural knowledge to perform approximate but fast model selection. We review here two particularly efficient ones: variational approximations and continuous relaxations.

First, although knowing the exact posterior distribution of $\mathbf{v}$ would require estimating a prohibiting $(2^p - 1)$-dimensional parameter, we can use the binary vector structure to derive a computationally cheaper approximation of the posterior. Specifically, we can consider a *mean-field approximation* $q_\rho(\mathbf{v})$ of the posterior that factorizes as a product of Bernoulli distributions with parameters $\boldsymbol{\rho} = (\rho_1, ..., \rho_p)$:

$$p(\mathbf{v}|\mathcal{D}) \approx q_{\boldsymbol{\rho}}(\mathbf{v}) = \prod_{i=1}^{p} q_{\rho_i}(v_i) = \prod_{i=1}^{p} \mathcal{B}(v_i|\rho_i). \tag{2.35}$$

Knowing this approximate posterior distribution conveniently requires to determine only a p-dimensional parameter. To insure that the approximation is close to the true posterior, *variational inference* minimizes the Kullback-Leibler divergence between $q_{\boldsymbol{\rho}}(\mathbf{v})$ and $p(\mathbf{v}|\mathcal{D})$. This is equivalent to maximizing a quantity known as the *evidence lower bound* (ELBO)

$$\text{ELBO}(\boldsymbol{\rho}) = \mathbb{E}_{\mathbf{v} \sim q_{\boldsymbol{\rho}}}[\log p(\mathcal{D}, \mathbf{v})] + \text{H}(q_{\boldsymbol{\rho}}), \tag{2.36}$$

with respect to $\boldsymbol{\rho}$. With this approximation, the very challenging computation of all 2^p posterior probabilities has been recast as a much simpler continuous p-dimensional optimization problem. This idea has been successfully applied to sparse high-dimensional linear and logistic regression (Logsdon et al., 2010; Carbonetto and Stephens, 2012; Huang et al., 2016). A similar approach, based on a related variational setting called *expectation propagation* (Minka, 2001), was also used for group-sparse regression (Hernández-Lobato et al., 2013). For more details on variational inference in general, and notably on optimization strategies for the ELBO, see Bishop (2006, Chapter 10) and Blei et al. (2017).

REMARK: THE ELBO AS AN APPROXIMATION OF THE MARGINAL LIKELIHOOD We have seen that the ELBO appears naturally if one wants to approximate a complex posterior using a parametric surrogate that minimizes the Kullback-Leibler divergence. But, as its name suggests, the ELBO also bounds the marginal likelihood (or evidence) and can therefore be seen as an approximation of it. This leads to an approximate procedure to compute posterior model probabilities, which has proven useful in many contexts involving complex posteriors, such as hidden Markov models (Watanabe et al., 2003), Gaussian mixture models (Bishop, 2006, Section 10.2.4) or stochastic block models (Latouche et al., 2012, 2014). As a non-asymptotic approximation, the ELBO usually compares favorably in small-sample scenarios with the Laplace-like approximations described in Section 2.3.3.a.

While variational inference provides a scalable way of tackling the variable selection problem, the mean-field assumption, which states that variable relevances are independent a posteriori, appears quite restrictive, especially when features are very correlated. Another more direct approach to transform the discrete optimization problem into a continuous one is trough making a continuous relaxation and replacing the condition $\mathbf{v} \in \{0,1\}^{\dim \Theta}$ by a continuous constraint $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^p$. Using the parameter set $\mathcal{V} = \mathbb{R}_+^p$ was the first proposal in that line of work. Introduced in the context of feed-forward neural networks by MacKay (1994) and Neal (1996) as *automatic relevance determination* (ARD), it led to efficient and sparse high dimensional learning in several contexts, including kernel machines (Tipping, 2001) and sparse probabilistic projections (Archambeau and Bach, 2009). Although the original motivation for ARD was mostly heuristic, similarly to the lasso, good theoretical properties were discovered later on (Wipf and Nagarajan, 2008; Wipf et al., 2011). In this thesis, we introduce the new heuristic relaxation $\mathbf{v} \in \mathcal{V} = [0,1]^p$ in the contexts of linear regression (Chapter 3) and PCA (Chapter 4). Maximizing the marginal likelihood using this hypercube constraint allows us to use the coefficients of $\mathbf{v}$ as relevance scores for the variables. This leads to the determination of a small subfamily of models over which the marginal likelihood is eventually discretely optimized. The key advantage of this technique is that while it has the scalability of both the variational approaches and ARD, it still performs exact Bayesian model selection at the end, the only approximation being the fact that only a small subfamily is considered.

2.4 Conclusion

Bayesian model uncertainty provides a systematized approach of many of the challenges modern statistics has to face: a large number of variables, a potentially low number of observations, and an ever-growing toolset of new statistical models. This framework, which was the subject of this review chapter, will constitute the main tool used in this thesis. As a concluding and temperating note, it is worth reminding and emphasizing that the paradigm of model uncertainty presented in this chapter has also been subject to much critique. For a philosophical overview of frequentists objections to Bayesian model uncertainty, see the monograph of Mayo and Spanos (2009). Even within the Bayesian community, several lines of work have criticized Jeffreys's framework, both from foundational (e.g. Gelman and Shalizi, 2013) and technical (e.g. Robert, 2016) grounds, leading to alternative paradigms for

model uncertainty (see e.g. the mixture approach of Kamary et al., 2014). We believe that such constructive criticism is vital for Bayesian model uncertainty to tackle the challenges offered by modern data. In particular, being able to diagnose cases where all models are irrelevant is not possible using model uncertainty, but is precisely the point of the *model criticism* advocated by Gelman and Shalizi (2013). We think that it will be customary in the future to combine model uncertainty with model criticism, in order to design these "sophisticatedly simple models" described and desired by Zellner (2001).

3

Combining a Relaxed EM Algorithm with Occam's Razor for Bayesian Variable Selection in High-Dimensional Regression

3.1 Introduction

As detailed in Chapters 1 and 2, over the past decades, sparsity has emerged as a very natural way to deal with high-dimensional (Candès, 2014; Hastie et al., 2015) data spaces. In the context of linear regression, finding a sparse parameter vector can both prevent overfitting, make an ill-posed problem (such as a "large p, small n" situation) tractable, and allow to interpret easily the data by finding which predictors are relevant. The problem of finding such predictors is referred to as *sparse regression* or *variable selection* and has mainly been considered either by likelihood penalization of the data, or by using Bayesian models.

3.1.1 Penalized likelihood

The most natural sparsity-inducing penalty, the ℓ_0-pseudonorm, is linked to the Akaike information criterion (Akaike, 1973) and to optimal subset selection. As proven by Natarajan (1995), it unfortunately leads to an NP-hard optimization problem that is intractable as soon as the number of predictors exceeds a few dozens. To overcome this restriction, convex relaxation of the ℓ_0-pseudonorm, that is, ℓ_1-regularization, have become a basic tool in modern statistics. The most spread formulation of the ℓ_1-penalized linear regression was introduced by Tibshirani (1996) as the "least absolute shrinkage and selection operator" (lasso) and by Chen et al. (1998) as "basis pursuit" in a signal processing framework. Several algorithms allow fast computations of the lasso, even when the number of predictors largely exceeds the number of observations. Among them is the popular least angle angle regression algorithm (LARS, Efron et al., 2004). The Dantzig selector, introduced by Candès and Tao (2007) as a refined ℓ_1-regularization problem, gives good variable selection performances while simply involving the resolution of a linear program. However, as shown by Zhao and Yu (2006), the crude lasso is not model-consistent unless some cumbersome conditions on the design matrix. Moreover, Zou and Hastie (2005) showed that it can be sensitive to highly correlated predictors and Pötscher and Leeb (2009) warned that its distributional properties can be surprisingly complex. A large number of proposals have been made to enhance the lasso as a selection operator. The adaptive lasso of Zou (2006) is a weighted version enjoying nice oracle properties that works extremely well in practice. "Bolasso", introduced by Bach

(2008), achieves model consistency by combining the lasso with a bootstrap step. In a similar fashion, the stability selection of Meinshausen and Bühlmann (2010) applies many lasso procedures with randomized weights on subsamples of the original data. This technique leads to effecient model selection, even in the presence of correlated predictors.

3.1.2 Bayesian modelling

Bayesian models have also been widely studied in a variable selection context (see e.g. O'Hara and Sillanpää, 2009, for a recent review). Bayesian procedures are supported by favorable empirical comparisons (Celeux et al., 2012) and strong theoretical analysis (Johnson and Rossell, 2012; Narisetty and He, 2014). However, most Bayesian techniques have difficulties in treating the case where the number of observations is smaller than the number of predictors (the so called "large p, small n" situation), mostly because of the exponential growth of the number of possible models (p predictors lead to 2^p models). Another drawback is the fact that the most classical linear regression prior, Zellner's (1986) g-prior (see Chapter 2), involves to invert the Fisher information matrix which is impossible in a "large p, small n" situation. Even though some regularization attempts of the g prior have been made by Baragatti and Pommeret (2012), the most efficient high-dimensional Bayesian techniques essentially rest on spike-and-slab procedures. Spike-and-slab models, first introduced by Mitchell and Beauchamp (1988), use mixtures of two distributions as priors for the regression coefficients: a thin one, corresponding to irrelevant predictors (the *spike*, typically a Dirac law or a Gaussian distribution with small variance) and a thick one, corresponding to the relevant variables (the *slab*, typically a uniform or Gaussian distribution of large variance). Notably, the refined spike-and-slab model of Ishwaran and Rao (2005a) or the PAC-Bayesian approach of Rigollet and Tsybakov (2011) have been particularly efficient even in very high-dimensional settings. Markov chain Monte Carlo (MCMC) methods have been usually chosen to select models with the highest posterior distributions. MCMC techniques, reviewed for example by Robert and Casella (2004), have an important computational cost and may suffer, as underlined by O'Hara and Sillanpää (2009), from poor mixing properties in the case of spike-and-slab-like priors. As mentioned in Chapter 2, a few deterministic methods have also recently been proposed to tackle this issue. The expectation propagation algorithm (EP, Minka, 2001) was applied to perform approximate inference for group feature selection with a spike-and-slab model by Hernández-Lobato et al. (2013). Related mean-field variational inference techniques have also been explored (Logsdon et al., 2010; Carbonetto and Stephens, 2012; Huang et al., 2016). The expectation maximization algorithm (EM, Dempster et al., 1977) was used by Ročková and George (2013) in the case of a hierarchical Bayesian model or by Yengo et al. (2014) in the case of a multi-slab empirical Bayes framework.

3.1.3 Our approach

As an alternative, our approach uses spike-and-slab-like priors induced by a binary vector which segregates the relevant from the irrelevant predictors. Such vectors, introduced by George and McCulloch (1993) have been widely used in the Bayesian literature, but have

always been considered as random parameters. In most Bayesian contexts like the (hierarchical) ones of George and McCulloch (1993), and Ishwaran and Rao (2005b) or the (empirical Bayes) one of George and Foster (2000), such a binary vector would be classically endowed with a product of Bernoulli prior distributions. In a PAC-Bayesian perspective, more complex prior distributions used for example by Alquier and Lounici (2011) or Rigollet and Tsybakov (2011) led to precise oracle inequalities and competitive predictive performances. In our work, the originality is to consider a deterministic binary vector, and to relax it in order to rely on an EM algorithm. This relaxed procedure allows us to find a family of p nested models, ordered by sparsity. Model selection is performed afterwards by maximizing the marginal likelihood over this family of models. This way to treat some parameters in a Bayesian way, and others in a frequentist one, is particularly motivated by the unifying multi-level inference approach advocated by Guyon et al. (2010) and by recent advances in Bayesian theory on the merging of partly frequentist empirical Bayes methods and classical hierarchical Bayesian approaches (Scott and Berger, 2010; Petrone et al., 2014).

The remainder of this document is organized as follows. In Section 3.2, a sparse generative model is defined and the general properties of its posterior distribution are exhibited. Section 3.3 shows how a relaxation of this model is considered in order to perform inference through an EM algorithm. Section 3.4 explains the model selection procedure of our approach and gives details about Occam's razor automatic selection as well as a link with classical frequentist penalized estimators. In Section 3.5, a new algorithm, called "SpinyReg", for variable selection in high-dimensional regression is introduced. Section 3.6 presents a benchmark comparison between SpinyReg and classical frequentist and Bayesian variable selection procedures, real and simulated data sets are considered. In Section 3.7, an original functional regression database, called OrsayVelib, is introduced and is used as a multivariate high-dimensional data set to demonstrate the efficiency of our approach.

3.2 A sparse generative model

This section introduces a sparse generative model based on a spike-and-slab-like prior, and describes the general properties of its posterior distribution. Links with related models are also discussed.

3.2.1 The model

Let us consider the following regression model

$$\begin{cases} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\beta} &= \mathbf{z} \odot \mathbf{w}, \end{cases} \tag{3.1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the vector of n observed responses, $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ is the design matrix with p input variables. The vector $\boldsymbol{\varepsilon}$ is a noise term with $p(\boldsymbol{\varepsilon}|\gamma) = \mathcal{N}(\boldsymbol{\varepsilon}|0, \mathbf{I}_n/\gamma)$. A prior distribution $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \mathbf{I}_p/\alpha)$ with an isotropic covariance matrix is further assumed. Moreover, we denote by $\mathbf{z} \in \{0,1\}^p$ a binary deterministic parameter vector, whose nonzero

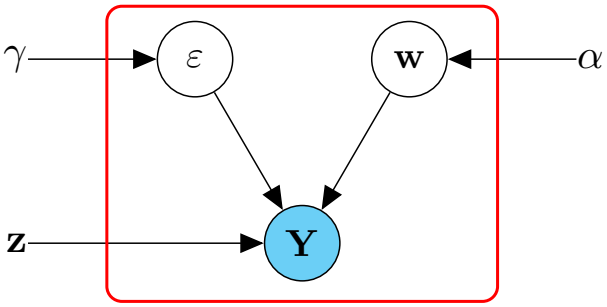


Figure 3.1 – Graphical representation of the sparse generative model.

entries correspond to the active variables of the regression model. It is worth noticing that such modeling induces a spike-and-slab-like prior distribution for $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\mathbf{z},\alpha) = \prod_{j=1}^{p} p(\beta_j|z_j,\alpha) = \prod_{j=1}^{p} \delta_0(\beta_j)^{1-z_j}\mathcal{N}(\beta_j|0,1/\alpha)^{z_j}. \tag{3.2}$$

However, we emphasize that, contrarily to standard spike-and-slab models (Mitchell and Beauchamp, 1988) which assume a Bernoulli prior distribution over $\mathbf{z}$, we see $\mathbf{z}$ here as a deterministic parameter to be inferred from the data. As we shall see, this allows us to work with a marginal log-likelihood which involves an Occam's razor term, allowing model selection afterwards. In the same spirit, we do not put any prior distribution on γ nor α. Finally, the graphical model is presented in Figure 3.1 and we denote by $q = ||\mathbf{z}||_0$ the number of relevant variables and $\mathbf{Z} = \text{diag}(\mathbf{z})$.

3.2.2 Posterior distribution

From now on, to simplify notations, the dependency on $\mathbf{X}$ in conditional distributions will be omitted.

Proposition 3.1. *The posterior distribution of* $\mathbf{w}$ *given the data is given by*

$$p(\mathbf{w}|\mathbf{Y},\mathbf{Z},\alpha,\gamma) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S}), \tag{3.3}$$

where $\mathbf{S} = (\gamma\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z} + \alpha\mathbf{I}_p)^{-1}$ *and* $\mathbf{m} = \gamma\mathbf{S}\mathbf{Z}\mathbf{X}^T\mathbf{Y}$.

Proof. Using Bayes' rule, we have

$$\begin{aligned}
\log p(\mathbf{w}|\mathbf{Y},\mathbf{Z},\alpha,\gamma) &= \log p(\mathbf{Y}|\mathbf{w},\mathbf{Z},\gamma) + \log p(\mathbf{w}|\alpha) + K_1 \\
&= -\frac{\gamma}{2}\left\|\mathbf{Y} - \mathbf{X}\mathbf{Z}\mathbf{w}\right\|_2^2 - \frac{\alpha}{2}\left\|\mathbf{w}\right\|_2^2 + K_2 \\
&= -\frac{\gamma}{2}\mathbf{w}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{w} + \gamma\mathbf{w}^T\mathbf{Z}\mathbf{X}^T\mathbf{Y} - \frac{\alpha}{2}\left\|\mathbf{w}\right\|_2^2 + K_3 \\
&= -\frac{1}{2}\mathbf{w}^T\mathbf{S}^{-1}\mathbf{w} + \mathbf{w}^T\mathbf{S}^{-1}\mathbf{m} + K_3.
\end{aligned}$$

where K_1, K_2 and K_3 are quantities that do not depend on $\mathbf{w}$. Therefore $p(\mathbf{w}|\mathbf{Y},\mathbf{Z},\alpha,\gamma) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$. □

The vector $\mathbf{m}$ is both the posterior mean the maximum a posteriori (MAP) point estimate of $\boldsymbol{\beta}$. Next proposition assures that it recovers the support of the parameter vector. Moreover, its nonzero coefficients correspond to ridge estimates with regularization parameter α/γ of the model where only the q predictors corresponding to the support of $\mathbf{z}$ have been kept.

Proposition 3.2. *We have* $\text{Supp}(\mathbf{m}) = \text{Supp}(\mathbf{z})$ *almost surely and*

$$\mathbf{m}_\mathbf{z} = \left(\mathbf{X}_\mathbf{z}^T\mathbf{X}_\mathbf{z} + \frac{\alpha}{\gamma}\mathbf{I}_p\right)^{-1}\mathbf{X}_\mathbf{z}^T\mathbf{Y}. \tag{3.4}$$

Proof. Using (3.3), one can write

$$\mathbf{S}^{-1}\mathbf{m} = \gamma\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{m} + \alpha\mathbf{m} = \gamma\mathbf{Z}\mathbf{X}^T\mathbf{Y},$$

which leads, by separating the lines corresponding to zero and nonzero coefficients of $\mathbf{z}$, to $\mathbf{m}_\mathbf{z} = 0$ and to (3.4). Notice that $\mathbf{m}_\mathbf{z} = 0$ implies $\text{Supp}(\mathbf{m}) \subset \text{Supp}(\mathbf{z})$. The vector $\mathbf{m}_\mathbf{z}$ therefore corresponds to the ridge estimator of the model where only the q predictors corresponding to the support of $\mathbf{z}$ have been kept. As a particular instance of a strictly convex bridge estimator, the coefficients of $\mathbf{m}_\mathbf{z}$ are almost surely nonzero (Fu, 1998, Theorem 1), therefore $\text{Supp}(\mathbf{m}) \subset \text{Supp}(\mathbf{z})$ implies that $\mathbf{m}$ and $\mathbf{z}$ have almost surely same support. □

3.2.3 Links with spike-and-slab models

Let us briefly link the proposed model to typical spike-and-slab models. The corresponding frameworks (Mitchell and Beauchamp, 1988; Hernández-Lobato et al., 2013) would add a hierarchical layer above the model of Figure 3.1 by using a multivariate Bernoulli prior of the form

$$p(\mathbf{z}) = \prod_{j=1}^{p} \tau_j^{z_j}(1-\tau_j)^{1-z_j},$$

where $\boldsymbol{\tau} = (\tau_1, ..., \tau_p) \in [0,1]^p$. However, as emphasized by Scott and Berger (2010), the estimation of $\boldsymbol{\tau}$ using empirical Bayes techniques can be extremely delicate and is likely to lead to poor variable selection performances. For instance, Hernández-Lobato et al. (2013) underline the fact that, in the case of their spike-and-slab model, the maximization of the evidence led to a sub-optimal choice of the hyper-parameter $\boldsymbol{\tau}$, and therefore to poor variable selection. To avoid such drawbacks, the use of Bernoulli priors is not considered in this chapter.

3.3 Inference

This section now focuses on inferring the model proposed above. To this end, $\mathbf{w}$ is seen as a latent variable while $\mathbf{Z} = \text{diag}(\mathbf{z})$, α, γ are parameters to be estimated from the data $(\mathbf{X}, \mathbf{Y})$ using an empirical Bayes framework.

3.3.1 Inference strategy and relaxation

The estimators of $\mathbf{z}$, α and γ will be the ones that maximize the *evidence* (or *type-II likelihood*) of the data:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{z}, \alpha, \gamma) = \int_{\mathbb{R}^p} p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{z}, \alpha, \gamma) p(\mathbf{w}|\alpha) d\mathbf{w}. \tag{3.5}$$

Seing $\mathbf{w}$ as a latent variable, a natural optimization procedure is the expectation-maximization (EM) algorithm introduced by Dempster et al. (1977). However, the maximization of (3.5) would be problematic for two reasons – both linked to the discreteness of the model parameter. First, because the optimization problem in $\mathbf{z}$ is combinatorial and 2^p values of $\mathbf{z}$ are possible. Then, because in this case, the parameter space is partly discrete and all theoretical convergence properties of the EM algorithm require a continuous parameter space (Wu, 1983; McLachlan and Krishnan, 2008).

To overcome these issues, we propose to use a simple relaxation by replacing the model parameter by a vector $\mathbf{z}^{\text{relaxed}}$ in $[0, 1]^p$. This relaxation allows us to efficiently maximize the new, relaxed version of (3.5) using an EM approach.

From now on, and until the end of this section, we will only consider the relaxed model with $\mathbf{z}^{\text{relaxed}} \in [0, 1]^p$. In order to simplify notations, we denote $\mathbf{Z} = \text{diag}(\mathbf{z}^{\text{relaxed}})$.

3.3.2 E-step

At the E-step of the relaxed EM algorithm, one has to compute the expectation of the complete data log-likelihood $\mathbb{E}_{\mathbf{w}}(\log p(\mathbf{Y}, \mathbf{w}, |\mathbf{Z}, \alpha, \gamma))$ with respect to the posterior distribution $p(\mathbf{w}|\mathbf{Y}, \mathbf{Z}, \alpha, \gamma)$. Consequently, the parameters $\mathbf{S}$ and $\mathbf{m}$ of the Gaussian posterior (3.3) have to be computed at each step.

Proposition 3.3. *Denoting $\boldsymbol{\Sigma} = \mathbf{S} + \mathbf{m}\mathbf{m}^T$, the expected complete data log-likelihood is given by*

$$\begin{aligned}\mathbb{E}_{\mathbf{w}}(\log p(\mathbf{Y}, \mathbf{w}|\mathbf{Z}, \alpha, \gamma)) = \frac{n}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{Y}^T\mathbf{Y} - \frac{\alpha}{2}\text{Tr}(\boldsymbol{\Sigma}) + \frac{p}{2}\log(\alpha) - \frac{p+n}{2}\log(2\pi) \\ + \gamma \mathbf{z}^{\text{relaxed}^T}(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y})) - \frac{\gamma}{2}\mathbf{z}^{\text{relaxed}^T}(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Sigma})\mathbf{z}^{\text{relaxed}}.\end{aligned} \tag{3.6}$$

Proof. By directly computing the integrand of (3.5), we find

$$\begin{aligned}\log p(\mathbf{Y}|\mathbf{Z}, \alpha, \gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) + \frac{p}{2}\log(\alpha) \\ + \log \int_{\mathbb{R}^p} \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{\gamma}{2}\mathbf{Y}^T\mathbf{Y} + \gamma\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbf{w} - \frac{\gamma}{2}\mathbf{w}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{w} - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) d\mathbf{w},\end{aligned}$$

which leads to

$$\begin{aligned}\log p(\mathbf{Y}|\mathbf{Z}, \alpha, \gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) + \frac{p}{2}\log(\alpha) - \frac{\gamma}{2}\|\mathbf{Y}\|_2^2 \\ + \log \int_{\mathbb{R}^p} \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{1}{2}\mathbf{w}^T\mathbf{S}^{-1}\mathbf{w} + \mathbf{w}^T\mathbf{S}^{-1}\mathbf{m}\right) d\mathbf{w},\end{aligned}$$

which allows us to conclude. □

3.3.3 M-step

Maximizing the expectation of the complete data log-likelihood (3.6) with respect to the parameter γ, α, $\mathbf{z}^{\text{relaxed}}$ leads to the following M-step updates.

Proposition 3.4. *The values of* γ, α, $\mathbf{z}^{\text{relaxed}}$ *maximizing* (3.6) *are*

$$\hat{\gamma}^{-1} = \frac{1}{n}\left\{\mathbf{Y}^T\mathbf{Y} + \mathbf{z}^{\text{relaxed}^T}(\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma})\mathbf{z}^{\text{relaxed}} - 2\mathbf{z}^{\text{relaxed}^T}(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))\right\} \tag{3.7}$$

$$\hat{\alpha} = \frac{p}{\text{Tr}(\mathbf{\Sigma})} \tag{3.8}$$

$$\hat{\mathbf{z}}^{\text{relaxed}} = \underset{\mathbf{u}\in[0,1]^p}{\arg\max}\left\{-\frac{1}{2}\mathbf{u}^T(\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma})\mathbf{u} + \mathbf{u}^T(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))\right\} \tag{3.9}$$

Proof. We have $\log p(\mathbf{Y}, \mathbf{w}|\mathbf{Z}, \alpha, \gamma) = \log p(\mathbf{Y}|\mathbf{w}, \mathbf{Z}, \alpha, \gamma) + \log p(\mathbf{w}|\alpha)$. Thus, since both the prior on $\mathbf{w}$ and the noise are Gaussian, we can write

$$\begin{aligned}\log p(\mathbf{Y}, \mathbf{w}|\mathbf{Z}, \alpha, \gamma) = \frac{n}{2}\log\gamma + \frac{p}{2}\log(\alpha) - \frac{p+n}{2}\log(2\pi) \\ - \frac{\gamma}{2}(\mathbf{Y} - \mathbf{XZw})^T(\mathbf{Y} - \mathbf{XZw}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}.\end{aligned}$$

Therefore, by expanding and computing the expectation of the expression, we find :

$$\begin{aligned}\mathbb{E}_{\mathbf{w}}(\log p(\mathbf{Y}, \mathbf{w}|\mathbf{Z}, \alpha, \gamma)) = \frac{n}{2}\log(\gamma) + \frac{p}{2}\log(\alpha) - \frac{p+n}{2}\log(2\pi) - \frac{\gamma}{2}\mathbf{Y}^T\mathbf{Y} \\ - \frac{\gamma}{2}\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{ZX}^T\mathbf{XZw}) + \gamma\mathbf{Y}^T\mathbf{XZ}\mathbb{E}_{\mathbf{w}}(\mathbf{w}) - \frac{\alpha}{2}\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{w}).\end{aligned}$$

We have $\mathbb{E}_{\mathbf{w}}(\mathbf{w}) = \mathbf{m}$ and, by using the properties of the trace operator,

$$\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{w}) = \mathbb{E}_{\mathbf{w}}(\text{Tr}(\mathbf{ww}^T)) = \text{Tr}(\mathbb{E}_{\mathbf{w}}(\mathbf{ww}^T)) = \text{Tr}(\mathbf{S} + \mathbf{mm}^T) = \text{Tr}(\mathbf{\Sigma}).$$

Thus, we will also have

$$\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{ZX}^T\mathbf{XZw}) = \mathbb{E}_{\mathbf{w}}(\text{Tr}(\mathbf{ZX}^T\mathbf{XZww}^T)) = \text{Tr}(\mathbf{ZX}^T\mathbf{XZ\Sigma}).$$

Moreover, since $\mathbf{Z} = \text{diag}(\mathbf{z}^{\text{relaxed}})$, we can compute

$$\mathbf{Y}^T\mathbf{XZm} = \mathbf{z}^{\text{relaxed}^T}(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))$$

and

$$\text{Tr}(\mathbf{ZX}^T\mathbf{XZ\Sigma}) = \mathbf{z}^{\text{relaxed}^T}(\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma})\mathbf{z}^{\text{relaxed}}.$$

By replacing the values of the terms we have just computed, we eventually find the appropriate value of the evidence. □

Notice that the $\mathbf{z}^{\text{relaxed}}$ update (3.9) is a quadratic program (QP) which is strictly concave if, and only if $\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}$ is positive definite. In fact, the next proposition assures that it is the case if and only if $\mathbf{X}$ has no null column. Therefore, in all practical cases, the objective function of this program is strictly concave and fast convex optimization procedures such as the L-BFGS-B method of Byrd et al. (1995) can be used.

Proposition 3.5. *The matrix* $\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma}$ *is positive definite if and only if* $\mathbf{X}$ *has no null column.*

Proof. According to the Schur product theorem (Bapat and Raghavan, 1997, Chapter 3), since $\mathbf{X}^T\mathbf{X}$ and $\mathbf{\Sigma}$ are positive semidefinite, $\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma}$ is also positive semidefinite. Therefore, $\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma}$ is positive definite if and only if its determinant is different from zero.

If one of the columns of $\mathbf{X}$ is null, then the same column of $\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}$ is also null and $\det(\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}) = 0$. The proposed condition is therefore necessary.

If none of the columns $\mathbf{x_1}, ...\mathbf{x_p}$ of $\mathbf{X}$ are null, then Oppenheim's (1930) inequality leads to

$$\det(\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}) \geq ||\mathbf{x_1}||_2^2 ... ||\mathbf{x_p}||_2^2 \det(\mathbf{\Sigma}).$$

Since $\mathbf{\Sigma} = \mathbf{S} + \mathbf{m}^T\mathbf{m}$, the determinant matrix lemma assures that

$$\det(\mathbf{\Sigma}) = (1 + \mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}) \det(\mathbf{S}),$$

and, since $\mathbf{S}$ and $\mathbf{S}^{-1}$ are positive definite, $\det(\mathbf{S}) > 0$ and $\mathbf{m}^T\mathbf{S}^{-1}\mathbf{m} \geq 0$. Therefore, we find

$$\det(\mathbf{\Sigma}) = (1 + \mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}) \det(\mathbf{S}) \geq \det(\mathbf{S}) > 0,$$

which, combined to Oppenheim's inequality, leads to $\det(\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}) > 0$. The condition is therefore also sufficient. □

3.3.4 Links with automatic relevance determination

Interestingly, this relaxed model is somehow related to the automatic relevance determination (ARD) which uses a prior of the form $p(\boldsymbol{\beta}|\mathbf{a}) = \mathcal{N}(\boldsymbol{\beta}|0, \text{diag}(\mathbf{a}))$ and for which the most classical way of inference is also an EM algorithm (MacKay, 1999; Tipping, 2001).

However, our method avoids several drawbacks of this technique. First, we do not assume any hyperprior on $\mathbf{z}^{\text{relaxed}}$ while Tipping (2001) uses a product of flat Gamma priors. More importantly, as pointed out by Wipf and Nagarajan (2008), the convergence of the EM algorithm is extremely slow and not theoretically guaranteed in the case of the ARD model. However, with our approach, since we only need the *ordering* of the coefficients of $\mathbf{z}^{\text{relaxed}}$ (see Section 3.4), we do not have to wait for the full convergence of this parameter. In practice, in all the experiments that we carried out, we only had to perform less than a few hundreds of iterations of the algorithm to obtain convergence of the evidence in order to perform variable selection. Notice that the fact that the evidence converges faster than the parameters of the model is a quite general property of EM algorithms (Xu and Jordan, 1996). Moreover, conversely to ARD-like models, our model additionally includes a "ridge parameter" α which, according to Occam's razor (see Section 3.4), also controls the sparsity. This also leads to an objective function different from the classical ARD one.

3.4 Model selection

In practice, the vector $\mathbf{z}^{\text{relaxed}}$ has to be binarized in order to select the relevant input variables. A common choice would consist in relying on a threshold τ such that z_j is set to 1

if $z_j \geq \tau$, and to 0 otherwise. However, numerical experiments showed that such a procedure would lead to poor estimates of $\mathbf{z}$. In order to perform an efficient variable selection, we will use the outputs of the relaxed EM algorithm to create a path of models and, relying on Occam's razor (see Chapter 2), we will afterward maximize the type-II likelihood over this path to finally select the relevant variables.

3.4.1 Occam's Razor

One of the key advantages of the approach proposed is that it maximizes a marginal log-likelihood, which automatically penalizes the model complexity by adding a term to the sum of squared errors.

Proposition 3.6. *Up to unnecessary additive constants, the negative type-II log-likelihood can be written as*

$$\begin{aligned} -\log p(\mathbf{Y}|\mathbf{z},\alpha,\gamma) &= -\log p(\mathbf{Y}|\mathbf{m},\mathbf{z},\gamma) + \text{pen}(\mathbf{z},\alpha,\gamma) \\ &= \frac{\gamma}{2}||\mathbf{Y}-\mathbf{X}_\mathbf{z}\mathbf{m}_\mathbf{z}||_2^2 + \text{pen}(\mathbf{z},\alpha,\gamma) \end{aligned} \tag{3.10}$$

where

$$\text{pen}(\mathbf{z},\alpha,\gamma) = -\log p(\mathbf{m}|\alpha) - \frac{1}{2}\log\det\mathbf{S} \tag{3.11}$$

$$= \frac{\alpha}{2}\left\|\mathbf{m}\right\|_2^2 - \frac{\log\alpha}{2}\left\|\mathbf{m}\right\|_0 - \frac{1}{2}\log\det(\gamma\mathbf{X}_\mathbf{z}^T\mathbf{X}_\mathbf{z}+\alpha\mathbf{I}_q) \text{ a.s} \tag{3.12}$$

is the Occam factor.

Proof. First, replacing $\mathbf{w}$ by $\mathbf{m}$ in the log-likelihood leads to

$$\log p(\mathbf{Y}|\mathbf{m},\mathbf{Z},\alpha,\gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) - \frac{\gamma}{2}\left\|\mathbf{Y}\right\|_2^2 - \frac{\gamma}{2}\mathbf{m}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{m} + \gamma\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbf{m}$$

therefore, since $\mathbf{m}^T\mathbf{S}^{-1}\mathbf{m} = \gamma\mathbf{m}^T\mathbf{Z}\mathbf{X}^T\mathbf{Y} = \gamma\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbf{m}$, we have

$$\log p(\mathbf{Y}|\mathbf{m},\mathbf{Z},\alpha,\gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) - \frac{\gamma}{2}\left\|\mathbf{Y}\right\|_2^2 - \frac{\gamma}{2}\mathbf{m}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{m} + \mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}.$$

Furthermore, $\log p(\mathbf{m}|\alpha) = -\frac{p}{2}\log(2\pi) + \frac{p}{2}\log(\alpha) - \frac{\alpha}{2}\mathbf{m}^T\mathbf{m}$. By summing the terms of the right-hand side of (3.11), we find the expression of the type-II log-likelihood that we already derived, which proves (3.11). To prove (3.10), let us note that

$$-\frac{1}{2}\log\det\mathbf{S} = \frac{1}{2}\log\det(\gamma\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}+\alpha\mathbf{I}_p) = \frac{\log\alpha}{2}(p-\left\|\mathbf{z}\right\|_0) - \frac{1}{2}\log\det(\gamma\mathbf{X}_\mathbf{z}^T\mathbf{X}_\mathbf{z}+\alpha\mathbf{I}_q).$$

Then, since $\left\|\mathbf{z}\right\|_0 = \left\|\mathbf{m}\right\|_0$ almost surely (see Proposition 2), we find

$$-\frac{1}{2}\log\det\mathbf{S} = \frac{\log\alpha}{2}(p-\left\|\mathbf{m}\right\|_0) - \frac{1}{2}\log\det(\gamma\mathbf{X}_\mathbf{z}^T\mathbf{X}_\mathbf{z}+\alpha\mathbf{I}_q) \text{ a.s.}$$

which leads to (3.11). □

The sparse generative model therefore automatically adds a ℓ_0-ℓ_2 penalty to the likelihood of the model at the MAP value of $\mathbf{w}$. This is somehow similar to the "elastic net" penalty of Zou and Hastie (2005), combined with a penalty linked to the volume of the Gaussian posterior $\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S})$. Notice that, when α is small, the Occam factor will be extremely sparsity-inducing but the coefficients will have a large variance. When α is close to 1, this penalty will lead to moderately sparse but notably shrinked solution. Moreover, if we write $\lambda = (\alpha - \log \alpha)/2$ and $\kappa = \alpha/(\alpha - \log \alpha)$, we obtain almost surely the expression

$$\text{pen}(\mathbf{z}, \alpha, \gamma) = \lambda \left((1-\kappa) \|\mathbf{m}\|_0 + \kappa \|\mathbf{m}\|_2^2 \right) - \frac{1}{2} \log \det(\gamma \mathbf{X}_\mathbf{z}^T \mathbf{X}_\mathbf{z} + \alpha \mathbf{I}_q),$$

involving a convex combination of the ℓ_0 and ℓ_2 penalties in an elastic net fashion. The elastic net can therefore be seen as some kind of strictly convex approximation of Occam's automatic penalty. Interestingly, recent theoretical work has also shown that hierarchical spike-and-slab-based model selection procedures can be equivalent to a ℓ_0 (Narisetty and He, 2014) or a ℓ_0-ℓ_2 penaly (Yen, 2011).

The term $\text{pen}(\mathbf{z}, \alpha, \gamma)$ exactly corresponds to Gull's (1988) Occam factor. Accordingly, the minimization of Equation (3.10) insures that the selected model realizes a tradeoff between the log-likelihood and an automatic penalty term. For more insight on this Occam factor, see Chapter 2. Let us simply remark that $\text{pen}(\mathbf{z}, \alpha, \gamma)$ is also related to the penalization term of the Bayesian information criterion (BIC, see Chapter 2). Indeed, if a broad Gaussian prior distribution for the vector $\mathbf{w}$ is considered and if the corresponding matrix $\mathbf{S}$ is assumed to have full rank, then Occam's razor is approximately $(-1/2)q \log n$. Contrarily to the BIC which relies on an asymptotic Laplace approximation, we obtained here an analytical expression of the evidence.

3.5 SpinyReg: an algorithm for sparse regression

We called our algorithm, which successively runs the EM algorithm for the relaxed model and performs model selection over the path of models, SpinyReg. Algorithms 1 and 2 present a pseudo-code for these two steps. An implementation of this algorithm is available via the R package `spinyreg` (available on CRAN).

3.5.1 Prediction

The SpinyReg algorithm is essentially a variable selection algorithm. In order to perform prediction, the natural estimator of the model is $\hat{\mathbf{z}}$ where

$$\hat{\mathbf{m}} = \hat{\gamma}(\hat{\gamma}\text{diag}(\hat{\mathbf{z}})\mathbf{X}^T\mathbf{X}\text{diag}(\hat{\mathbf{z}}) + \hat{\alpha}\mathbf{I}_p)^{-1}\text{diag}(\hat{\mathbf{z}})\mathbf{X}^T\mathbf{Y}.$$

However, as stated at the end of Section 3.3.1, this estimator is exactly the ridge estimator performed on a small model where only the predictors corresponding to nonzero coefficients of $\hat{\mathbf{z}}$ are kept. Since we do not wait for the full convergence of the parameters in the EM algorithm, we would rather recommend to perform an ordinary least squares (OLS) estimation or a ridge regression with only a small amount of regularization on the same small model. This is the choice we made in the numerical simulations hereafter.

Algorithm 1 EM algorithm for the relaxed model

Input: $\mathbf{X}$,$\mathbf{Y}$
Output: $\hat{\mathbf{z}}^{\text{relaxed}}$
  Initialize $\gamma = 1$, $\alpha = 1$, $\mathbf{z}^{\text{relaxed}} = (1, ..., 1)$
  repeat
    // E-step
    $\mathbf{S} = \gamma(\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z} + \alpha I_p)^{-1}$
    $\mathbf{m} = \gamma\mathbf{S}\mathbf{Z}\mathbf{X}^T\mathbf{Y}$; $\mathbf{\Sigma} = \mathbf{S} + \mathbf{m}\mathbf{m}^T$
    // M-step
    Compute $\hat{\alpha}$ and $\hat{\gamma}$ using (3.8) and (3.7)
    Compute $\hat{\mathbf{z}}^{\text{relaxed}}$ using (3.9) and the L-BFGS-B method
  until convergence of the evidence

Algorithm 2 Model selection algorithm

Input: $\mathbf{X}$,$\mathbf{Y}$,$\hat{\alpha}$,$\hat{\gamma}$,$\hat{\mathbf{z}}^{\text{relaxed}}$
Output: $\hat{\mathbf{z}}$
  for $k = 1$ to p **do**
    Compute $\hat{\mathbf{z}}^{(k)}$
  end for
  $\hat{q} = \arg\max_{1 \leq k \leq p} p(\mathbf{Y}|\hat{\mathbf{z}}^{(k)}, \hat{\alpha}, \hat{\gamma})$
  $\hat{\mathbf{z}} = \hat{\mathbf{z}}^{(\hat{q})}$

3.5.2 Initialization

The choice of initialization $\mathbf{z}^{\text{relaxed}} = (1, ..., 1)$ appears particularly natural because it helps to avoid the unwanted apparition of true zero coefficients in $\mathbf{z}^{\text{relaxed}}$. Indeed, if a coefficient of $\mathbf{z}^{\text{relaxed}}$ by the M-step update (3.9), then it can not go back to a positive value. This behavior is typical of ARD-like iterative procedures (MacKay, 1999; Tipping, 2001).

Contrarily to ARD models, we do not need true zeros in the vector $\mathbf{z}^{\text{relaxed}}$. Therefore, another solution to avoid their apparition would be to perform the quadratic program (3.9) over $[\eta_n, 1 - \eta_n]$ were $(\eta_n)_{n \leq 1}$ is a vanishing real sequence. The resulting algorithm would be a generalized EM (GEM) algorithm satisfying Wu's convergence conditions (Wu, 1983), contrary to the classical EM algorithm for ARD (Tipping, 2001; Wipf and Nagarajan, 2008). However, because we do not wait for the convergence of $\mathbf{z}^{\text{relaxed}}$, setting the initial coefficients at 1 is sufficient in practice to avoid true zeros. Regarding the parameter α, the form of the Occam factor suggests that using a small value such as $\alpha = 10^{-3}$ will lead to sparse solutions. This is the choice we made in the numerical simulations hereafter.

3.5.3 Computational cost

At each iteration, the most expensive step is the inversion of the $p \times p$ matrix $\mathbf{S}$ during the E-step. It would imply a $O(p^3)$ complexity, not allowing us to deal with high-dimentional data. However, using the Woodbury identity, one can write when $p > n$,

$$\mathbf{S} = \frac{1}{\alpha}\mathbf{I}_p + \frac{1}{\alpha^2}\left(\mathbf{Z}\mathbf{X}^T\right)\left(\frac{1}{\gamma}\mathbf{I}_n + \frac{1}{\alpha}\mathbf{X}\mathbf{Z}^2\mathbf{X}^T\right)^{-1}\left(\mathbf{X}\mathbf{Z}\right).$$

Thus, the final computational cost has therefore a $O(p^2 \min(n,p))$ complexity, which is more suitable for high-dimensional problems.

Overall, MCMC-based Bayesian variable selection methods for regression have a very large computational cost. To the best of our knowledge, the fastest efficient spike-and-slab algorithm for linear regression is the EP procedure of Hernández-Lobato et al. (2013). Each iteration of the EP algorithm costs $O(n^2 p)$ operations, and in practice it needs more iterations than our relaxed EM algorithm to converge. The complexity of the LARS algorithm is $O(pqn + pq^2 + q^3)$ (Bach et al., 2012). SpinyReg therefore realizes a complexity tradeoff between slow MCMC Bayesian techniques and fast ℓ_1-based methods. SpinyReg is consequently particularly suitable when p is moderately large. Screening procedures, similar to the sure independence screening (SIS) of Fan and Lv (2008) for instance, can be used to reduce the dimensionality to a reasonable level when p exceeds a few thousands. Such an approach is usual in the Bayesian literature (e.g. Narisetty and He, 2014).

Let us also emphasize that, whereas frequentist methods use cross-validation to optimize the prediction performance, SpinyReg automatically estimates its hyper-parameters. In particular, its inference procedure includes the estimation of the penalty term α which is linked to the sparsity level. Therefore, the computational cost of SpinyReg has to be compared to the one of ℓ_1-based methods with the cross-validation included.

3.5.4 Path of Models

We rely on $\hat{\mathbf{z}}^{\text{relaxed}}$ to find a path of models which are likely to have a high evidence. We build a path by assuming that the larger the coefficients of $\hat{\mathbf{z}}^{\text{relaxed}}$ are, the more likely they are to correspond to relevant variables.

We define the set of vectors $(\hat{\mathbf{z}}^{(k)})_{k\leq p}$ as the binary vectors such that, for each k, the k top coefficients of $\hat{\mathbf{z}}^{\text{relaxed}}$ are set to 1 and the others to 0. For example, $\hat{\mathbf{z}}^{(1)}$ contains only zeros and a single 1 at the position of the highest coefficient of $\hat{\mathbf{z}}^{\text{relaxed}}$. The set of vectors $(\hat{\mathbf{z}}^{(k)})_{k\leq p}$ defines a path of models to look at for model selection. Note that this path allows us to deal with a family of p models (ordered by sparsity) instead of 2^p, allowing our approach to deal with a large number of input variables. Thus, the evidence is evaluated for all $\hat{\mathbf{z}}^{(k)}$ and the number $\hat{q}$ of relevant variables is chosen such that the evidence is maximized:

$$\hat{q} = \underset{1\leq k\leq p}{\arg\max}\, p(\mathbf{Y}|\hat{\mathbf{z}}^{(k)}, \hat{\alpha}, \hat{\gamma}) \quad \text{and} \quad \hat{\mathbf{z}} = \hat{\mathbf{z}}^{(\hat{q})}. \tag{3.13}$$

3.6 Numerical comparisons

In this section, we illustrate the behavior of SpinyReg on simulated and real data sets, and compare it to the most efficient state-of-the-art methods.

3.6.1 Simulation setup

In order to consider a wide range of scenarios, we use three different simulation scenarios: "uniform", "Toeplitz" and "blockwise". The simulation of the parameter $\mathbf{w}$ and of the noise $\boldsymbol{\varepsilon}$ is common for the three schemes: $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_p/\alpha)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}_n/\gamma)$. The design matrix $\mathbf{X}$ is simulated according to a Gaussian distribution with zero mean and a covariance matrix R depending on the chosen scheme. The correlation structure of $R = (r_{ij})_{i,j=1,\dots,p}$ is as follows:

- "uniform": $r_{ii} = 1$ for all $i = 1, ...p$ and $r_{ij} = \rho$ for $i, j = 1, \dots, p$ and $i \neq j$,
- "Toeplitz": $r_{ii} = 1$ for all $i = 1, ...p$ and $r_{ij} = \rho^{|i-j|}$ for $i, j = 1, \dots, p$ and $i \neq j$,
- "blockwise": $R = \text{diag}(R_1, ..., R_4)$ is a 4-blocks diagonal matrix where R_ℓ is such that $r_{\ell ii} = 1$ and $r_{\ell ij} = \rho$ for $i, j = 1, \dots, p/4$ and $i \neq j$.

Then, $\mathbf{Z}$ is simulated by randomly picking q active variables among p. The predictive vector Y is finally computed according to Equation (3.1).

3.6.2 An introductory example

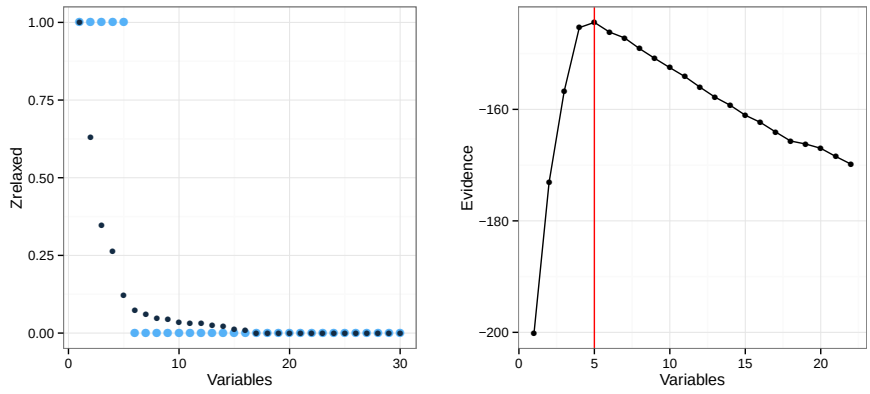
We consider here an introductory example which aims at highlighting the main features of the proposed approach. For this experiment, the Toeplitz simulation setup is used with $p = 30$, $q = 5$, $\rho = 0.25$, $\alpha = 1$ and $\gamma = 1$. From this setup, two data sets were simulated with respectively $n = 100$ and $n = 30$ observations. The second setting corresponds to a difficult scenario where $n = p$ whereas the first one should be easier to fit. Notice that the dimensionality is kept relatively low mainly for visualization purposes. Figure 3.2 presents the results of the application of SpinyReg on those two data sets. The left panels present in dark blue the values of $\hat{\mathbf{z}}^{\text{relaxed}}$ (sorted in decreasing order) and the corresponding true values of $\mathbf{z}$ (pale blue points) used in the simulations. The right panels show the values of evidence computed on the path of models.

Regarding the first example, one can see that the five largest values of $\hat{\mathbf{z}}^{\text{relaxed}}$ actually correspond to the five active variables. This confirms that SpinyReg succeeds here in finding the relevant variables in the regression model. The second panel confirms that SpinyReg would select five variables among the 30 original ones. On this quite simple example, SpinyReg yields a true positive rate (TPR) equals to 1 and a false positive rate (FPR) equals to 0.

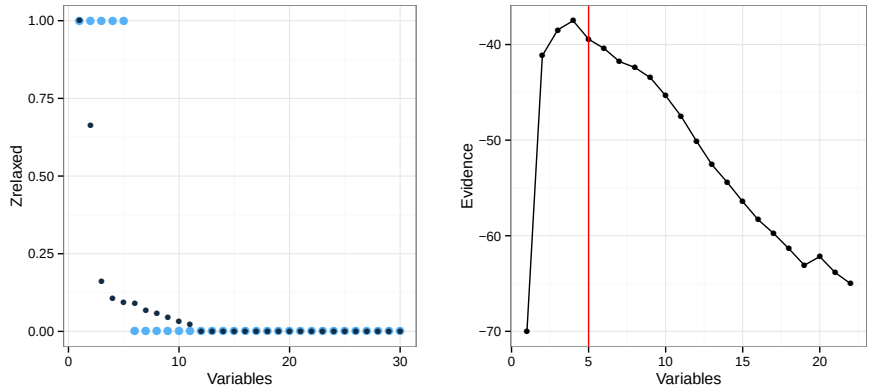
For the second and much more difficult situation (bottom row of Figure 3.2), the estimated values for $\mathbf{z}^{\text{relaxed}}$ are less discriminative. Indeed, the values of $\hat{\mathbf{z}}^{\text{relaxed}}$ are smaller than in the simpler case. However, even though the ranking of variables induced by $\hat{\mathbf{z}}^{\text{relaxed}}$ respects the partition between active and inactive variables, Occam's razor leads to a too conservative choice and misses one active variable. On this more difficult data set, SpinyReg yields a true positive rate (TPR) equals to 0.8 and a false positive rate (FPR) equals to 0.

3.6.3 Benchmark study on simulated data

We now compare the performance of SpinyReg with three of the most recent and popular variable selection methods based on ℓ_1 regularization: the lasso of Tibshirani (1996), the



Toeplitz setup with $\rho = 0.25$, $p = 30$ and $n = 100$



Toeplitz setup with $\rho = 0.25$, $p = 30$ and $n = 30$

Figure 3.2 – Variable selection with SpinyReg on the two introductory examples ($p = 30$ and $n = 150$ or $n = 30$). The left panels present the values of $\hat{\mathbf{z}}^{\text{relaxed}}$ (dark blue) and the actual binary values of $\mathbf{z}$ (pale blue). The right panels show the values of evidence computed on the path of models.

adaptive lasso of Zou (2006) and the stability selection of Meinshausen and Bühlmann (2010). We also added two recent spike-and-slab approaches: the multi-slab framework of CLERE (Yengo et al., 2014) and the EP procedure of Hernández-Lobato et al. (2013). To this end, we simulated 100 data sets for each of the three simulations schemes (uniform, Toeplitz and blockwise), for three data set sizes ($n = p/2$, $n = p$, $n = 2p$) and two values for the correlation parameter ($\rho = 0.25$ and $\rho = 0.75$). The other simulation parameters were $p = 100$, $q = 40$, $\alpha = 1$ and $\gamma = 1$. The measures used to evaluate the method performances are the prediction mean square error on test data (MSE, hereafter), the F-score (the harmonic mean of precision and recall, which provides a good summary of variable selection performances) and the estimated value of q (number of relevant predictors).

Lasso and Stability selection were trained using the R package `quadrupen` (Grandvalet et al., 2016). We used the package `parcor` (Kraemer et al., 2009) to train the adaptive lasso and the package `clere` (Yengo et al., 2016) to train CLERE. The spike-and-slab approach

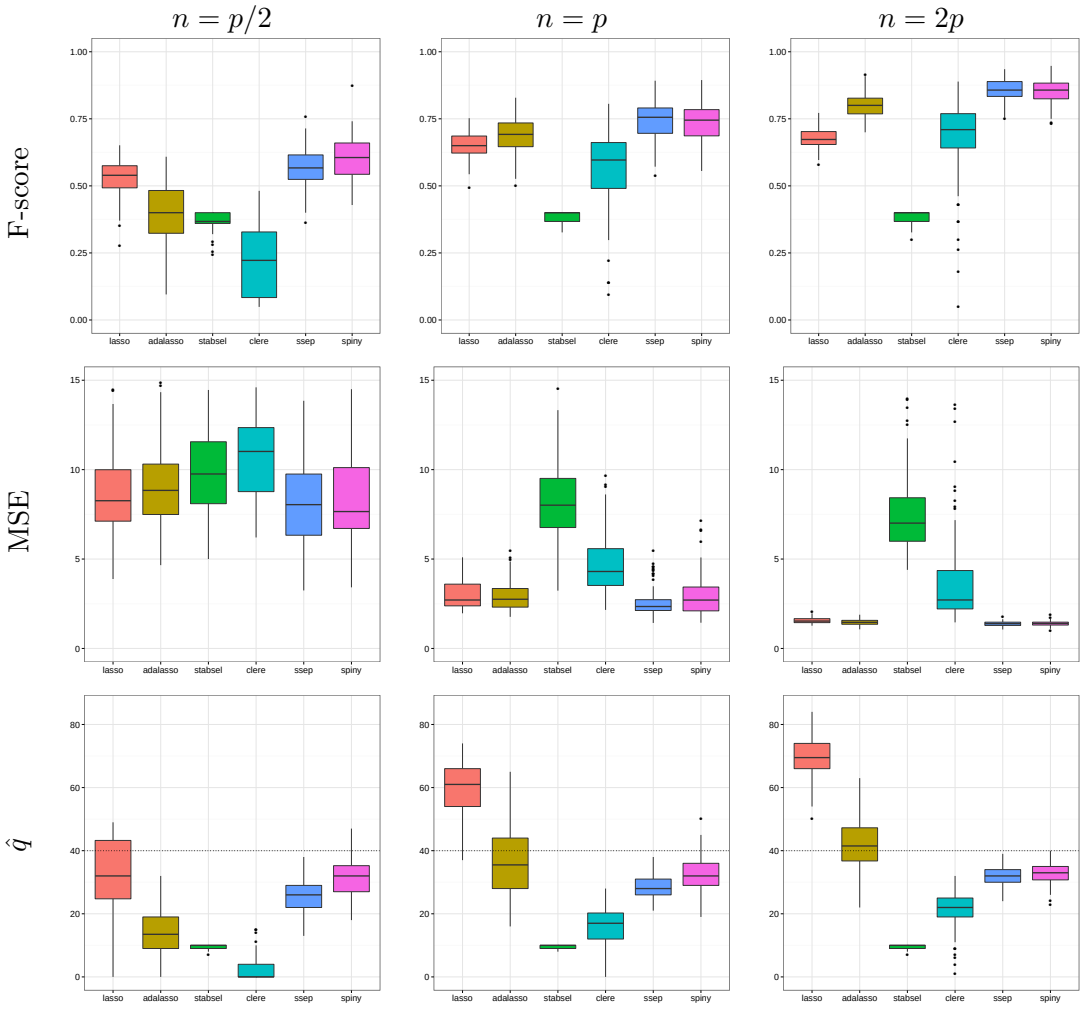


Figure 3.3 – Scenario "blockwise" with $\rho = 0.75$.

of Hernández-Lobato et al. (2013), which uses expectation propagation, will be subsequently denoted SSEP and was trained using the code available on the authors' web pages.

We present here only the results for two simulation setups: the "blockwise" one with $\rho = 0.75$ and the "Toeplitz" one with $\rho = 0.25$. All the other results are available in Appendix B. Note that similar conclusions can be drawn on these other scenarios. Figure 3.3 presents the F-score, MSE and $\hat{q}$ of the 6 studied methods for the blockwise simulation setup with $\rho = 0.75$ and for the three data set sizes, while Figure 3.4 presents these measures for the Toeplitz simulation setup with $\rho = 0.25$ and for the three data set sizes.

The first row of Figure 3.3 and Figure 3.4 gives the F-score. This measure allows us to figure out how the methods behave in terms of detection of the relevant variables. We can see that SpinyReg and SSEP outperform other methods and have close variable selection performances. SpinyReg appears to be at his best in the "$n = p/2$" case on these runs.

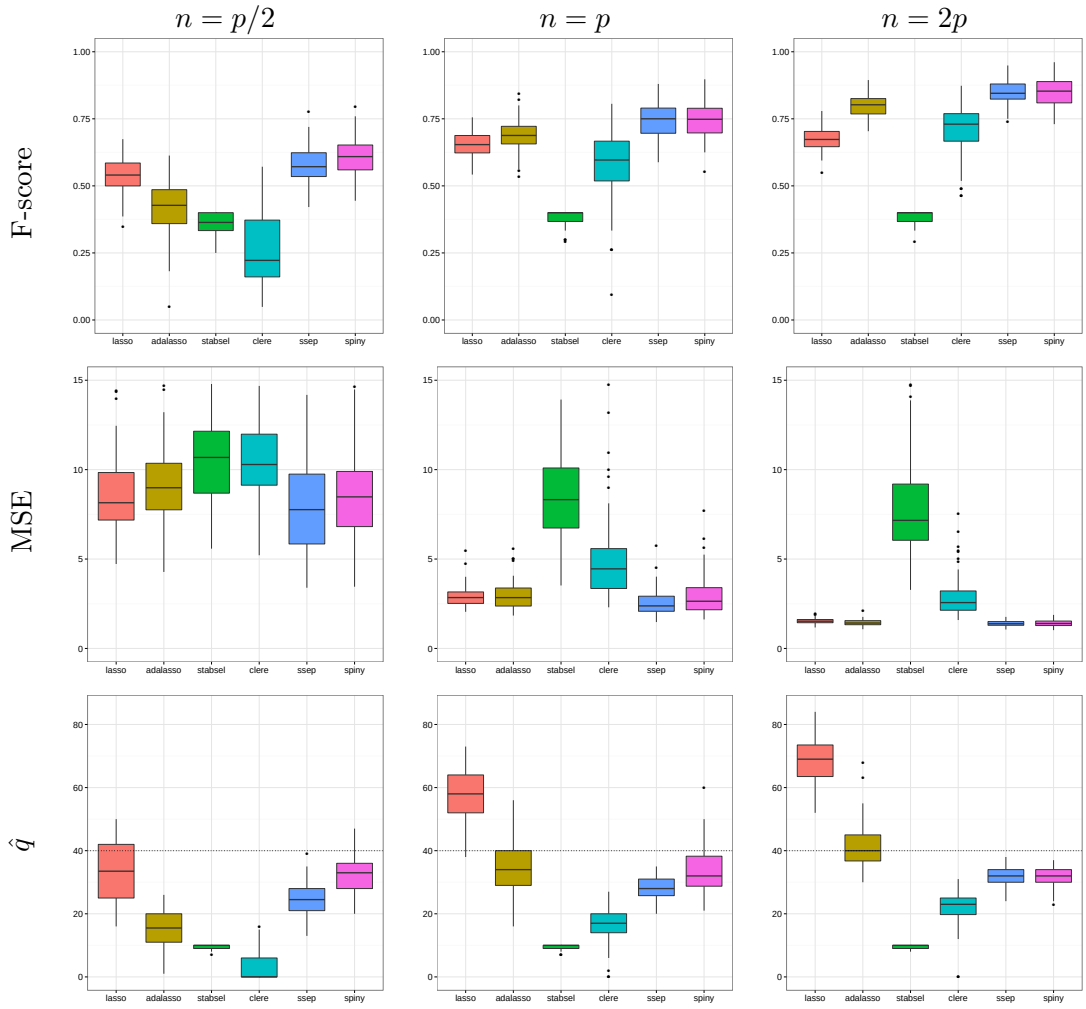


Figure 3.4 – Scenario "Toeplitz" with $\rho = 0.25$.

The second row of Figure 3.3 and Figure 3.4 provides the MSE values for the studied methods. Most of the methods perform well except stability selection and CLERE when $n \leq p$. In particular, SpinyReg has the best prediction performance for $n = p/2$ with the highly correlated blockwise case.

The last row of Figure 3.3 and Figure 3.4 gives the number q of active variables estimated by the 6 methods. We remind that the actual number of active variables is $q = 40$ for these simulations (represented by the dashed lines on Figure 3.3). It is worth noticing that lasso has a clear tendency to overestimate the number of active variables, particularly when n becomes large. Conversely, stability selection has the opposite behavior and underestimates q. Its very conservative behavior has the advantage of avoiding false positives. It turns out that SpinyReg provides consistently a good estimate of the actual value of q.

	Prostate ($n = 77$, $p = 8$)		Eyedata ($n = 96$, $p = 200$)	
	MSE×100	Selected variables	MSE×100	Selected variables
Lasso	63.6 ± 21.8	3.33 ± 0.877	1.26 ± 0.964	16.7 ± 5.56
Adalasso	58.4 ± 15.9	4.42 ± 1.57	1.50 ± 1.248	2.4 ± 0.700
Stability Selection	61.6 ± 14.4	1.94 ± 0.239	1.58 ± 0.850	1.7 ± 0.823
Clere	59.8 ± 19.7	2.87 ± 0.825	-	-
SSEP	56.6 ± 15.0	2.76 ± 0.474	-	-
SpinyReg	58.3 ± 15.4	3.34 ± 0.607	1.25 ± 0.920	143 ± 9
	OzoneI ($n = 162$, $p = 134$)		DiabetesI ($n = 353$, $p = 64$)	
	MSE	Selected variables	MSE/1000	Selected variables
Lasso	18.9 ± 4.96	10.3 ± 2.27	3.22 ± 0.407	7.43 ± 2.41
Adalasso	16.84 ± 4.48	8.32 ± 3.16	3.02 ± 0.395	9.31 ± 2.25
Stability Selection	17.9 ± 5.25	9.68 ± 1.10	2.97 ± 0.387	7.77 ± 0.423
Clere	19.6 ± 5.48	5.43 ± 2.55	3.15 ± 0.384	2.33 ± 0.587
SSEP	29.6 ± 10.2	74.8 ± 5.45	3.70 ± 0.647	62.0 ± 1.36
SpinyReg	18.9 ± 5.46	10.79 ± 2.69	3.13 ± 0.376	8.5 ± 1.45

Table 3.1 – Results on real-world data sets

3.6.4 Study on classical regression data sets

We now consider four real-world data sets: the classical `prostate` data set used for example by Tibshirani (1996), the `eyedata` data set of Scheetz et al. (2006), which contains gene expression data of mammalian eye tissue samples, the `OzoneI` data set included in the `spikeslab` package (Ishwaran et al., 2010) and which uses the ozone data set of Breiman and Friedman (1985) with some additional interactions and the `DiabetesI` data set which is also available in the `spikeslab` package and uses the diabetes data set of Efron et al. (2004) with some additional interactions. Applying the same methods as before, we trained our data randomly using 80% of the observations and computed the test error on the remaining data. Repeating this procedure 100 times, we computed the mean and the standard deviation of the test error and of the number of variables selected. Results are reported in Table 3.1. We did not compute the test error for methods which did not succeed in selecting variables.

We can see that SpinyReg obtains competitive predictive results on all data sets. Moreover, we can note that it is less conservative than most other algorithms. On the challenging `eyedata` data set for example, while the two other Bayesian methods fail to select at least one variable, SpinyReg selects three quarters of the predictors and has the lowest MSE. The three ℓ_1 based methods select only a few variables and have higher MSE. It is worth noticing that we tried to apply the elastic net of Zou and Hastie (2005) (which, using a ℓ_1-ℓ_2 regularization, is able to select more variables than most classical ℓ_1 procedures) to this data set. Elastic net selected all variables. This behavior is close to the one of SpinyReg and reminds the interesting analogy between the Occam factor (3.10) used in SpinyReg and the elastic net penalty.

Let us finally highlight that the medium prediction rank of SpinyReg is the second best, behind the adaptive lasso. Let us also emphasize that all frequentist methods were trained using cross-validation which optimizes prediction performance. Conversely, SSEP, CLERE and SpinyReg automatically estimate their hyper-parameters. In particular, the inference

procedure of SpinyReg includes the estimations of the penalty term α which is linked to the sparsity level.

3.7 Prediction of the frequentation of the Orsay museum using bike-sharing data

In this section, we introduce a new regression problem, which aims at predicting the number of visitors of the Orsay museum (Paris) using the activity of the Paris bike-sharing system (*Vélib'*).

3.7.1 Predicting a touristic index using open data

The emergence of open data systems has brought about a surge of complex data illustrating various social behaviors. In this challenging context, the analysis of bike-sharing systems (BSSs) provides a new insight into the touristic patterns of a city. We therefore wanted to see how well, in a city like Paris, bike-sharing data could predict a touristic index, such as the number of visitors of an important museum. With nearly three million annual visitors, the Orsay museum is one of the ten most visited museums in the world (Skeggs, 2014). Known for having the vastest collection of impressionist paintings in the world, it holds for example Manet's *Le Déjeuner sur l'herbe* or Van Gogh's *Nuit étoilée sur le Rhône*. The frequentation of the museum at each hour was given as a courtesy by the museum services. The Paris BSS, called *Vélib'*, was launched by JCDecaux and the city of Paris in 2007 and is nowadays certainly the most active BSS in Europe. Statistical studies of the *Vélib'* system have been for example conducted by Njato Randriamanamihaga et al. (2014) and Bouveyron et al. (2015). The predictive variables that will interest us for our regression problem are the percentages of parked bikes (or *loadings*) for all the *Vélib'* stations of Paris. These percentages are available through the open data API provided by JCDecaux (real-time data are available at https://developer.jcdecaux.com/ with an API key).

3.7.2 The OrsayVelib database

At each hour, the number of visitors present in the museum constitutes the response variable of our regression problem. The predictors are the loadings at each hour of the $p = 1158$ *Vélib'* stations in Paris. Only the hours corresponding to opening days (from 8am to 6pm, except Mondays) of the museum are kept. The month of September 2014 constitutes the learning set (with $n = 316$ observations), and the first two weeks of October 2014 the test set (see Figure 3.5).

This data set, thereafter called the OrsayVelib database, has several interesting aspects. First, while most "large p, small n" regression problems inherit their dimensionality from genomics or signal processing, this data set is purely related to social sciences. This illustrates the fact that modern social data can also lead to high-dimensional challenging statistical problems. Second, since the variables are the *Vélib'* stations, a sparse solution can be easily interpretable and visualizable. We would expect the relevant predictors to correspond – at

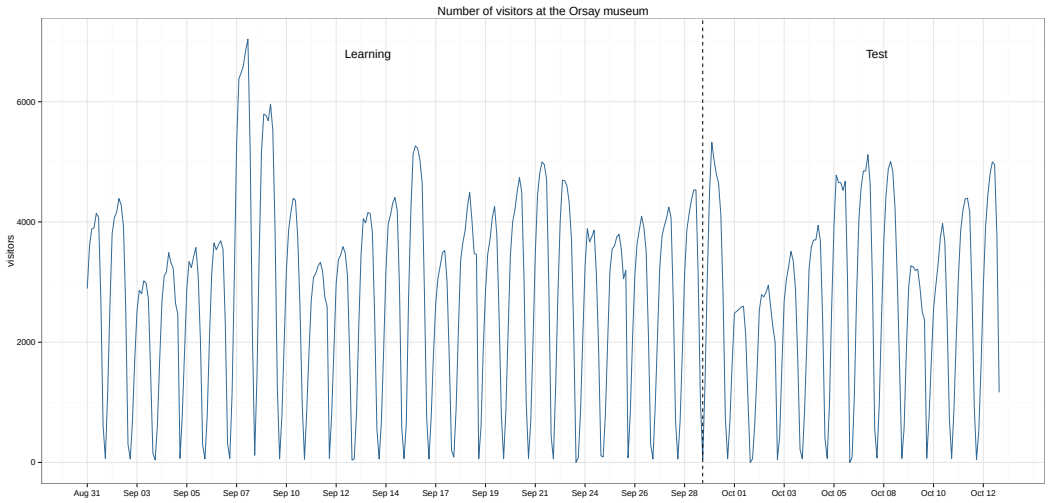


Figure 3.5 – Number of visitors during the learning and test phases. Only opening hours of the museum (8am to 7pm, from Tuesday to Sunday) are shown.

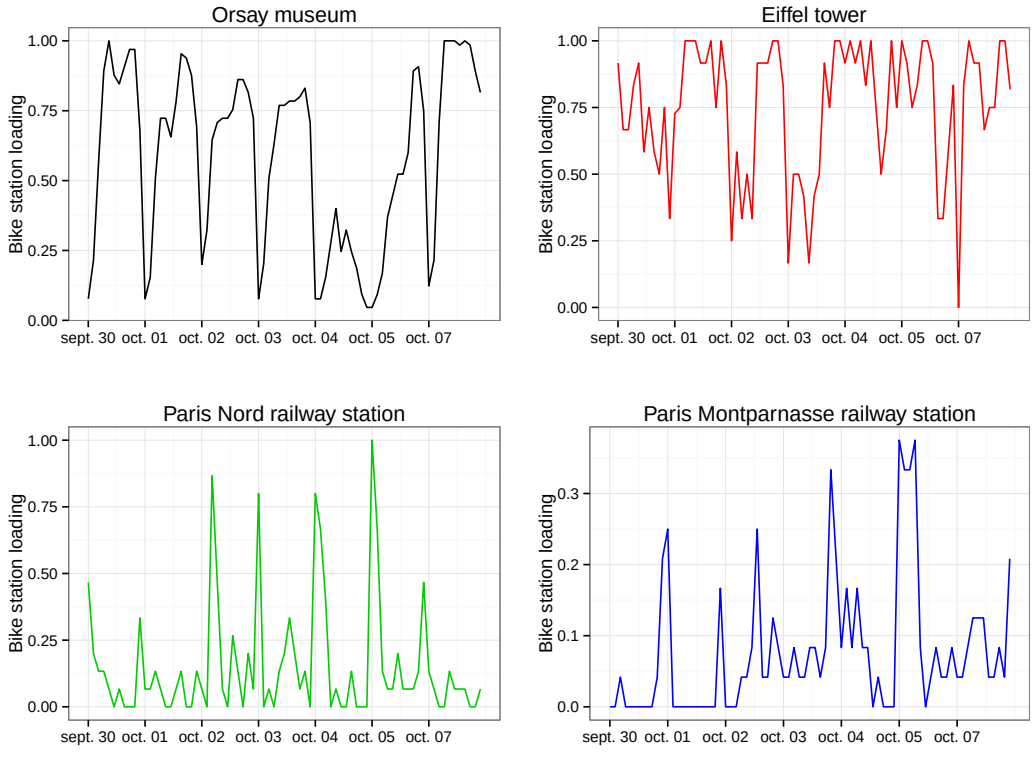


Figure 3.6 – Loadings of four *Vélib'* stations during the first week of September. Only opening hours of the museum (8am to 7pm, from Tuesday to Sunday) are shown.

	Ridge	SSEP	Lasso	Adalasso	SpinyReg
MSE$\times 10^4$	145.66	144.38	132.08	159.17	127.36
Selected variables	1158	1146	167	155	45

Table 3.2 – Test error and number of selected predictors for each method.

least to some extend – to stations used by the visitors of the Orsay museum. In particular, the behavior of the stations closest to the museum are expected to be of important interest. For visualization purposes, one can plot on a map the location of the selected variables, being able to efficiently interpret the selection. Finally, the learning/test segregation of the data harshly punishes overfitting. Indeed, while September 2014 (the learning month) corresponded to exceptionally good weather conditions in Paris, October had some rainy days. Since BSS data are naturally heavily linked to the weather, this means that overfitting algorithms will struggle with predicting the number of predictors on rainy days (such as October 8th). This interesting behavior is exhibited in the next subsection.

To illustrate the behavior of the data, Figure 3.5 provides the curve of the number of visitors during the learning and test phases and Figure 3.6 shows the loadings of four *Vélib'* stations during the first week of September. Two of these stations correspond to touristic areas with different behaviors: one is the closest one to the Orsay museum and one is one of the closest ones to the Eiffel tower. The other two correspond to large railway stations (which also happen to be large subway stations). We will show in the next subsection that these stations are of particular interest if we aim at predicting the number of visitors of the museum.

3.7.3 Results

We applied the algorithms of Section 3.6.1 to the OrsayVelib database. Since the sparsity of this regression problem is not absolutely certain, we also added a non-sparse method to the benchmark: ridge regression with a cross-validated regularization parameter. The test errors and sparsity patterns obtained are detailed in Table 3.2 (for the sake of clarity, only the five best methods are displayed). One can notice that SpinyReg has the lowest generalization error and that it selects fewer variables than its competitors.

Figure 3.7 allows to compare the true number of visitors during the test phase with the predicted values of the four methods. We can notice that, as expected, all algorithms struggle with October 8th, which was a rainy day. On this specific day, SpinyReg is (especially in the afternoon) the closest one to the truth. In a similar fashion, SpinyReg is the only method that accurately predicts the small augmentation of the first three days of October.

Eventually, one can plot the location of the selected variables on the map of Paris. For the sake of clarity, we only did it for the two best methods: lasso and SpinyReg. Figure 3.8 presents the maps of selected stations by both methods. Green dots correspond to positive coefficients and red dots to negative coefficients. The dot size indicates the magnitude of the coefficient (the larger the dot, the larger the absolute value of the coefficient). The black dot corresponds to the location of the Orsay museum.

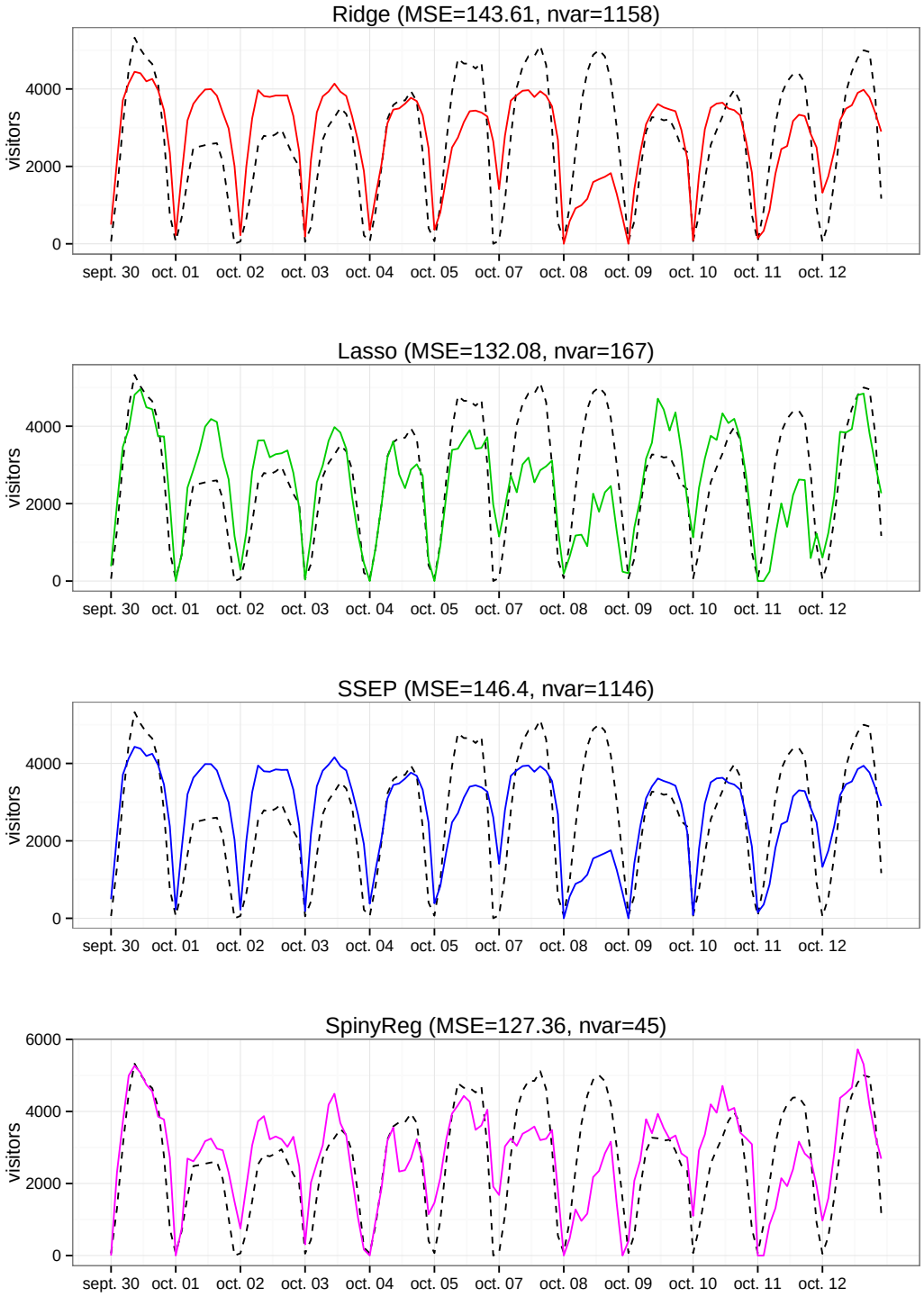
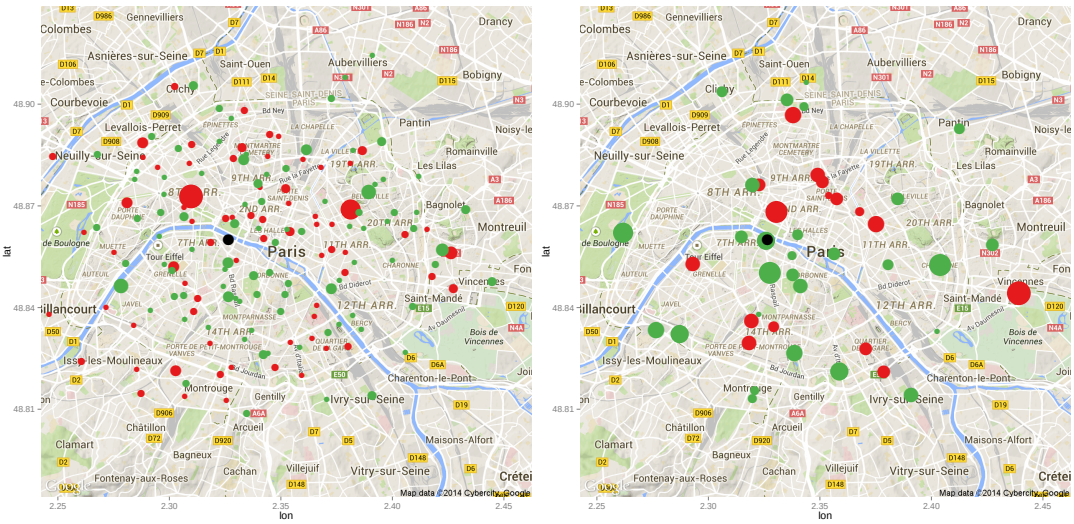


Figure 3.7 – Ground truth (dashed line) and predicted values for the number of visitors at each hour.



Lasso (MSE=132.08, 167 var.) SpinyReg (MSE=127.36, 45 var.)

Figure 3.8 – Stations selected by the lasso (left) and by SpinyReg (right). Green dots correspond to positive coefficients and red dots to negative coefficients (the larger the dot, the larger the absolute value of the coefficient). The black dot corresponds to the location of the Orsay museum.

The lasso selection appears to be very broad and difficult to interpret. In particular, the lasso does not select the closest station to the museum. Conversely, the SpinyReg selection is more interpretable: one can see that it does select the closest stations to the museum, and that their regression coefficients are positive (which means that these stations are likely to be full when the museum is crowded). Around the neighborhood of the museum, there is a ring of stations with almost exclusively negative coefficients (Eiffel tower, Paris Nord and Montparnasse railway stations, place de la Bastille) which can be interpreted as stations from where the visitors of the museum rent their bikes. Beyond this ring, the selected stations essentially correspond to popular public parks (bois de Vincennes, parc Montsouris, parc André Citro"en, bois de Boulogne). This is not surprising since their frequentation is also linked to the touristic activity of the city.

As a summary, SpinyReg both succeeds in providing an interpretable selection of *Vélib'* stations while having the most effective prediction performance.

3.8 Conclusion

We considered the problem of Bayesian variable selection for high-dimensional linear regression through a sparse generative model. The sparsity is induced by a deterministic binary vector which multiplies with the Gaussian regressor vector. The originality of the work was to consider its inference through relaxing the model and using a type-II log-likelihood maximization based on an EM algorithm. Model selection can be performed relying on Occam's razor and on a path of models found by the EM algorithm. Numerical experiments on simulated data have shown that SpinyReg performs well compared to the most recent

competitors both in terms of prediction and of selection, especially in moderately sparse cases and with highly correlated predictors.

SpinyReg was finally applied for the prediction of a touristic index from open data. OrsayVelib, a new high-dimensional regression database, was introduced to this end and allowed us to illustrate the powerful aspects of the proposed method. It is worth noticing that, even though it was used here as a multivariate high-dimensional data set, OrsayVelib is by nature functional. It would be therefore interesting to investigate the extension of SpinyReg to functional regression (Silverman and Ramsay, 2005, chap. 12). The variable selection will, in such a case, operate on the basis functions. For instance, if we consider Fourier bases, it would allow to recover the periodicity of the data.

4

Bayesian Variable Selection for Globally Sparse Probabilistic PCA

4.1 Introduction

From the children test results of the seminal paper of Hotelling (1933) to the challenging analysis of microarray data (Ringnér, 2008) and the recent successes of deep learning (Chan et al., 2015), principal component analysis (PCA) has become one of the most popular tools for data-preprocessing and dimension-reduction. The original procedure consists in projecting the data onto a "principal" subspace spanned by the leading eigenvectors of the sample covariance matrix. It was later shown that this subspace could also be retrieved from the maximum-likelihood estimator of a parameter, in a particular factor analysis model called probabilisitic PCA (PPCA) (Roweis, 1998; Tipping and Bishop, 1999). This probabilistic framework led to diverse Bayesian analysis of PCA (Bishop, 1999a; Minka, 2000; Nakajima et al., 2011).

4.1.1 Local and global sparsity

A potential drawback of PCA is that the principal components are linear combinations of every single original variable, and can therefore be difficult to interpret. To tackle this issue, several procedures have been designed to project the data onto subspaces generated by sparse vectors while retaining as much variance as possible. Many of them were based on convex or partially convex relaxations of cardinality-constrained PCA problems – among these techniques are the popular ℓ_1-based SPCA algorithm of Zou et al. (2006) or the semidefinite relaxation of d'Aspremont et al. (2008). Another strategy is to use a sparsity-inducing prior distributions on the coefficients of the projection matrix (Archambeau and Bach, 2009; Guan and Dy, 2009; Khanna et al., 2015).

However, when several principal components are computed, these various techniques do not enforce them to have the same sparsity pattern (i.e. the same active variables), and each component has to be interpreted individually. While individual interpretation is particularly natural in several cases – when PCA serves visualization, for example –, it is not adapted to situations where the practitioner aims at *globally* selecting which features are relevant. In these situations, a simple and popular approach has been to consider that the relevant variables correspond to the sparsity pattern of the first principal component (Zou et al., 2006; Zhang et al., 2012). However, this procedure is limited, and several important aspects of the data may lie in the next principal components. For example, in the colon cancer data set studied by d'Aspremont et al. (2008), the most relevant genes were the ones selected not by the first but by the *second* principal component. Another motivation for global sparsity is the fact that, in many real-life situations, the sparsity pattern of the axes computed by a sparse PCA algorithm are extremely close. This is for example the case of the three axes of the template attacks application considered by Archambeau and Bach (2009). In this setting, forcing these patterns to be equal will give the practitioner a precise idea of which variables are relevant. Another interesting feature of global sparsity is the fact that, once the common sparsity pattern has been determined, performing PCA on the relevant variables yields orthogonal and uncorrelated principal components – conversely to most sparse PCA procedures.

4.1.2 Related work

Since the seminal papers of Jolliffe (1972, 1973) and Robert and Escoufier (1976), several methods have been designed to discard features in PCA (see e.g. Brusco (2014) for a recent review). However, these techniques were designed to eliminate redundant, rather that irrelevant variables, and are based on combinatorial algorithms that are not really suitable for high-dimensional problems.

A simple and scalable way of performing variable selection for PCA is to simply keep the features that have the largest marginal variance. In certain cases, this technique is theoretically sound, and was applied for instance to the analysis of electrocardiogram (ECG) data (Johnstone and Lu, 2009). Zhang and El Ghaoui (2011) also proved that it could be used as an efficient preprocessing technique to reduce the dimensionality of ultra-high dimensional problems before applying a traditional sparse PCA algorithm. However, this technique has two main drawbacks. First, it is not robust to simple transformations of the data since simply multiplying a variable by a constant may wrongfully select (or discard) it. An unfortunate consequence of this is the fact that this technique can not be applied to scaled data. Moreover, since it ignores non-marginal information, this technique will behave badly in the case of correlated features.

A more refined approach to global sparsity is ℓ_1-based regularization, which has imposed itself as one of the most versatile and efficient approaches to sparse statistical learning (Hastie et al., 2015). In a context of *structured* sparse PCA, Jenatton et al. (2009) proposed to recast sparse PCA as a penalized matrix factorization problem and suggested that limiting the number of sparsity patterns allowed within the principal vectors could improve the feature extraction quality – particularly in face recognition problems. Using the $\ell_1 - \ell_2$ norm, they derived an algorithm (hereafter referred as SSPCA) that allows to compute d sparse components with exactly $m \leq d$ sparsity patterns. However, they only considered cases where m is larger than 2 and therefore did not focus on global sparsity. They were followed by Khan et al. (2015) who, in a very close framework, argued that global sparsity (which they called *joint sparsity*) led to better representations of hyperspectral images. Other similar approaches based on structured composite norms have been conducted by Masaeli et al. (2010), Gu et al. (2011) and Xiaoshuang et al. (2013). Ulfarsson and Solo (2008a, 2011) used sparsity inducing penalties together with a PPCA model to enforce global sparsity. They proposed an algorithm called *sparse variable noisy PCA* (hereafter refered as svnPCA) and fixed the amount of penalization using the Bayesian information criterion (BIC) of Schwarz (1978).

Eventually, it is worth mentioning that global sparsity has also been investigated in other contexts, such as partial least squares regression (Liu et al., 2013) or electroencephalography (EEG) imaging (Wipf and Nagarajan, 2009; Gramfort et al., 2013).

4.1.3 Contributions and organization of the chapter

We present in Section 4.2 a Bayesian approach that allows to project the data onto a *globally sparse subspace* (i.e a subspace spanned by vectors with the same sparsity pattern) while preserving a large part of the variance. To this end, we use the noiseless PPCA model

introduced by Roweis (1998) together with an isotropic gaussian prior on the projection matrix and a binary vector that segregates relevant from irrelevant variables. While past Bayesian PCA frameworks relied on variational (Bishop, 1999b; Archambeau and Bach, 2009; Guan and Dy, 2009) or Laplace (Bishop, 1999a; Minka, 2000; Sobczyk et al., 2017) methods to approximate the marginal likelihood, we derive here a closed-form expression for the evidence based on the multivariate Bessel distribution. In order to avoid the drawbacks of discrete model selection and to treat high-dimensional data, we also present a relaxation of our model by replacing the binary vector with a continuous one. Inference of this relaxed model can be performed using a variational expectation-maximization (VEM) algorithm. Such a procedure allows to find a path of models. The exact evidence is eventually maximized over this path, relying on Occam's razor (MacKay, 2003, chap. 28), to select the relevant variables.

We illustrate the behaviour of our algorithm and compare it to other methods in Section 4.3. In particular, we show that Bayesian model selection empirically outperforms ℓ_1-ℓ_2-based regularization on a series of tasks.

Sections 4.4 and 4.5 are devoted to two applications showcasing the features of our method. The first one concerns signal denoising with wavelets, and shows how global sparsity can surpass traditional sparse PCA algorithms within this context. The second one treats about unsupervised gene selection. Given an (unlabeled) microarray data matrix, we show how GSPPCA can select biologically relevant subsets of genes. Interestingly, we exhibit an important correlation between our exact marginal likelihood expression and a criterion of biological relevance based on pathway enrichment.

4.2 Bayesian variable selection for PCA

Let us assume that a centered i.i.d. sample $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ is observed which one wishes to project onto a d-dimensional subspace while retaining as much variance as possible. All the observations are stored in the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$.

4.2.1 Probabilistic PCA

The PPCA model assumes that each observation is driven by the following generative model

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}, \tag{4.1}$$

where $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_d)$ is a low-dimensional Gaussian latent vector, $\mathbf{W}$ is a $p \times d$ parameter matrix called the *loading matrix* and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$ is a Gaussian noise term.

This model is a particular instance of factor analysis and was first introduced by Lawley (1953). Following Theobald (1975), Tipping and Bishop (1999) confirmed that this generative model is equivalent to PCA in the sense that the principal components of $\mathbf{X}$ can be retrieved using the maximum likelihood (ML) estimator $\mathbf{W}_{\text{ML}}$ of $\mathbf{W}$. Indeed, if $\mathbf{A}$ is the $p \times d$ matrix of ordered principal eigenvectors of $\mathbf{X}^T\mathbf{X}$ and if $\boldsymbol{\Lambda}$ is the $d \times d$ diagonal matrix with corresponding eigenvalues, we have

$$\mathbf{W}_{\text{ML}} = \mathbf{A}(\boldsymbol{\Lambda} - \sigma^2 \mathbf{I}_d)^{1/2}\mathbf{R}, \tag{4.2}$$

where $\mathbf{R}$ is an arbitrary orthogonal matrix.

Several Bayesian treatments of this model have been conducted by using different priors on the loading matrix. However, the marginal likelihood of these models appeared to be untractable. To tackle this issue, several computational techniques were considered. The automatic relevance determination (ARD) prior was used together with Laplace (Bishop, 1999a) or variational (Bishop, 1999b; Archambeau and Bach, 2009) approximations. Minka (2000) introduced more complex conjugate priors to perform Bayesian model selection on the dimension d of the latent space using the Laplace approximation. Combined with variational inference, several sparsity inducing priors such as the Laplace (Guan and Dy, 2009), the generalized hyperbolic (Archambeau and Bach, 2009) or the spike-and-slab (Lázaro-Gredilla and Titsias, 2011) prior were also chosen for $\mathbf{W}$.

In this work, we aim at avoiding these approximations. Our approach is to investigate in which cases the marginal likelihood can be analytically computed. To this end, we will use the fact that, within the PPCA model (5.1), the limit noiseless setting $\sigma \to 0$ also allows to recover the principal components. This convenient framework was first studied by Roweis (1998) and has proven to be useful in several situations. The noiseless PPCA model was used for instance to facilitate inference in the presence of missing data (Yu et al., 2010; Ilin and Raiko, 2010). More importantly in our context, it was successfully used by Sigg and Buhmann (2008) to enforce sparsity within an ℓ_1-penalized PPCA framework – which means that getting rid of the noise term is likely to be compatible with variable selection.

4.2.2 A general framework for globally sparse PPCA

In a classical (locally) sparse PCA context, the loading matrix $\mathbf{W}$ would be expected to contain few nonzero coefficients. However, to reach global sparsity, *several entire rows* of $\mathbf{W}$ have to be further constrained to be null. In this work, we handle variable selection using a binary vector $\mathbf{v} \in \{0,1\}^p$ whose nonzero entries correspond to relevant variables. For technical purposes, we also denote by $\bar{\mathbf{v}}$ the binary vector of $\{0,1\}^p$ whose support is exactly the complement of $\mathrm{Supp}(\mathbf{v})$. We denote $q = ||\mathbf{v}||_0$ the number of relevant variables. In the PPCA framework, this leads to the following model for each observation

$$\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}, \tag{4.3}$$

where $\mathbf{V} = \mathrm{diag}(\mathbf{v})$. Notice that the rows of $\mathbf{VW}$, corresponding to the zero entries of $\mathbf{v}$, are null. Therefore, the principal subspace will be generated by a basis of vectors which shares the sparsity pattern of $\mathbf{v}$. Such spaces spanned by a family of vectors sharing the same sparsity pattern will be called *globally sparse subspaces*. This definition of global sparsity is closely related to the notion of *row sparsity* of Vu and Lei (2013).

We further assume that the coefficients of the matrix $\mathbf{W}$ are endowed with the Gaussian priors $w_{ij} \sim \mathcal{N}(0, 1/\alpha^2)$, for all i, j. Following the parametric empirical Bayes framework (Kass and Steffey, 1989) leads to seeking the parameters $\mathbf{v}$, α and σ that maximizes the *marginal likelihood* or *evidence*

$$p(\mathbf{X}|\mathbf{v},\alpha,\sigma) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{v},\alpha,\sigma) = \prod_{i=1}^{n} \int_{\mathbb{R}^{p\times d}} p(\mathbf{x}_i|\mathbf{W},\mathbf{v},\alpha,\sigma)p(\mathbf{W})d\mathbf{W}.$$

In previous Bayesian PCA models, the marginal likelihood was never derived because it was too difficult to compute in practice or even intractable. Here, specifically, the evidence of the model can be expressed analytically as a univariate integral using the isotropy of the prior on $\mathbf{W}$. In the following, $\mathbf{x_v}$ denotes the subvector of $\mathbf{x}$ where only the variables corresponding to the nonzero indexes of $\mathbf{v}$ are kept. Given a real order ν, we denote repectively by J_ν and K_ν the Bessel function of the first kind and the modified Bessel function of the second kind (Abramowitz and Stegun, 1965, chap. 10 and 11).

Theorem 4.1. *The density of* $\mathbf{x}$ *is given by*

$$p(\mathbf{x}|\mathbf{v},\alpha,\sigma) = \frac{e^{-\frac{||\mathbf{x_v}||_2^2}{2\sigma^2}}}{(2\pi)^{p/2}\sigma^{p-q}}||\mathbf{x_v}||_2^{1-q/2}\int_0^\infty \frac{u^{q/2}e^{-\sigma^2u^2}}{(1+(u/\alpha)^2)^{d/2}}J_{q/2-1}(u||\mathbf{x_v}||_2)du. \quad (4.4)$$

Proof. Let us first consider the case where all variables are active and assume that $\mathbf{v} = (1,1,...,1)$. Therefore, $\mathbf{V} = \mathbf{I}_p$ and the considered model reduces to probabilistic PCA. In this framework, we will derive the density of $\mathbf{x}$ by computing the Fourier transform of its characteristic function.

In order to compute the characteristic function of $\mathbf{x}$, we first decompose the latent vector $\mathbf{y}$ in the canonical base

$$\mathbf{y} = y_1\mathbf{e_1} + ... + y_d\mathbf{e_d},$$

where $(\mathbf{e_i})_{i\geq d}$ is the canonical base of $\mathbb{R}^d$. We can now write the vector $\mathbf{Wy}$ as a sum of of d i.i.d variables

$$\mathbf{Wy} = y_1\mathbf{We_1} + ... + y_d\mathbf{We_d}.$$

Its characteristic function will consequently be

$$\varphi_{\mathbf{Wy}} = (\varphi_{y_1\mathbf{We_1}})^d.$$

Now, for all $\mathbf{u} \in \mathbb{R}^d$, we have

$$\varphi_{y_1\mathbf{We_1}}(\mathbf{u}) = \mathbb{E}[\exp(iy_1{\mathbf{e_1}}^T\mathbf{W}^T\mathbf{u})] \quad (4.5)$$

$$= \mathbb{E}\left[\exp\left(iy_1\sum_{k=1}^p w_{k1}u_k\right)\right], \quad (4.6)$$

but, since $w_{st} \sim \mathcal{N}(0,\alpha)$ for all s,t, we will have

$$\frac{1}{\sqrt{\alpha}||\mathbf{u}||_2}\sum_{k=1}^p w_{k1}u_k \sim \mathcal{N}(0,1),$$

thus, since $\mathbf{y}$ and $\mathbf{W}$ are independent, the law of $(\sqrt{\alpha}||\mathbf{u}||_2)^{-1}y_1\sum_{k=1}^p w_{k1}u_k$ will be the one of a product of two standard Gaussian random variables, whose density is $1/\pi K_0(|.|)$ (Wishart and Bartlett, 1932). Therefore, we find that

$$\varphi_{y_1\mathbf{We_1}}(\mathbf{u}) = \frac{1}{\pi}\int_{-\infty}^{+\infty} K_0(|t|)e^{i\sqrt{\alpha}||\mathbf{u}||_2 t}dt$$

$$= \frac{2}{\pi}\int_0^{+\infty} K_0(t)\cos(\sqrt{\alpha}||\mathbf{u}||_2 t)dt,$$

is simply the cosine Fourier transform of a univariate Bessel function. Using a formula in Abramowitz and Stegun (1965, p. 486), we eventually find that

$$\varphi_{y_1 \mathbf{W}}(\mathbf{u}) = \frac{1}{\sqrt{1+\alpha||\mathbf{u}||_2^2}},$$

which leads to

$$\varphi_{\mathbf{Wy}}(\mathbf{u}) = \frac{1}{(1+\alpha||\mathbf{u}||_2^2)^{d/2}}.$$

Finally, since the noise term and $\mathbf{Wy}$ are independent, the characteristic function of $\mathbf{x}$ will be

$$\varphi_{\mathbf{x}}(\mathbf{u}) = \varphi_{\mathbf{Wy}}(\mathbf{u})\varphi_{\boldsymbol{\varepsilon}}(\mathbf{u}) = \frac{e^{-\sigma^2||\mathbf{u}||_2^2}}{(1+\alpha||\mathbf{u}||_2^2)^{d/2}}.$$

The density of $\mathbf{x}$ is then given by the Fourier transform of its characteristic function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^p}\int_{\mathbb{R}^p} \varphi_{\mathbf{x}}(\mathbf{u}) e^{i\mathbf{x}^T\mathbf{u}} d\mathbf{u},$$

but, since $\varphi_{\mathbf{x}}(\mathbf{u})$ is a radial function (i.e a function that only depends on the norm of its argument), its Fourier transform can be expressed as a univariate integral (Schaback and Wu, 1996) and we can write

$$p(\mathbf{x}) = \frac{||\mathbf{x}||_2^{1-p/2}}{(2\pi)^{p/2}}\int_0^{+\infty} \frac{u^{p/2}e^{-\sigma^2u^2}}{(1+\alpha u^2)^{d/2}} J_{p/2-1}(u||\mathbf{x}||_2)du, \tag{4.7}$$

which is the desired form for the case with no inactive variable.

In the general case, $\mathbf{v}$ is not necessarily equal to $(1,1,...,1)$ but we can notice that, since $\mathbf{x_v}$ and $\mathbf{x_{\bar{v}}}$ are independent, we can write $p(\mathbf{x}) = p(\mathbf{x_{\bar{v}}})p(\mathbf{x_v})$. Applying (4.7) to $\mathbf{x_v}$ allows us to compute $p(\mathbf{x_v})$ and to eventually obtain the expression of the density given by the theorem. □

While reducing the dimension of the integration domain to one appears to be a valuable improvement, the integral of Equation (4.4), albeit univariate, falls within the category of Hankel-like integrals known to be particularly delicate to compute, even numerically. This is due to the fact that the integrand has singularities near the real axis (Ogata, 2005). To overcome this limitation, we investigate in the following subsection the use of the noiseless PPCA model to obtain a tractable expression.

4.2.3 A closed-form evidence for globally sparse noiseless PPCA

To obtain a closed-form expression of the marginal likelihood, we consider the following modification of Model (4.3). For the relevant variables, we use the noiseless PPCA model, and we assume that the irrelevant variables are generated by a Gaussian white noise. More specifically, we write

$$\mathbf{x} = \mathbf{VWy} + \bar{\mathbf{V}}\boldsymbol{\varepsilon_1} + \mathbf{V}\boldsymbol{\varepsilon_2}, \tag{4.8}$$

where $\bar{\mathbf{V}} = \text{diag}(\bar{\mathbf{v}})$, $\boldsymbol{\varepsilon_1} \sim \mathcal{N}(0, \sigma_1^2\mathbf{I}_p)$ is the noise of the inactive variables and $\boldsymbol{\varepsilon_2} \sim \mathcal{N}(0, \sigma_2^2\mathbf{I}_p)$ is the noise of the active variables, having in mind that we aim at investigating

the noiseless limit $\sigma_2 \to 0$. We will see that, with this particular formulation of the problem, the evidence has a closed form expression which involves the multivariate Bessel distribution, introduced by Fang et al. (1990, Def. 2.5).

Definition 4.1. *A random vector is said to have a* ***symmetric multivariate Bessel distribution*** *with parameters $\beta > 0$ and $\nu > -k/2$ if its density is*

$$\forall \mathbf{z} \in \mathbb{R}^k,\ \text{Bessel}(\mathbf{z}|\beta,\nu) = \frac{2^{-k-\nu+1}\beta^{-k-\nu}}{\Gamma(\nu+k/2)\pi^{k/2}}||\mathbf{z}||_2^{\nu}K_{\nu}(||\mathbf{z}||_2/\beta).$$

Theorem 4.2. *In the noiseless limit $\sigma_2 \to 0$, $\mathbf{x}$ converges in probability to a random variable $\tilde{\mathbf{x}}$ whose density is*

$$p(\tilde{\mathbf{x}}|\mathbf{v},\alpha,\sigma_1^2) = \mathcal{N}(\tilde{\mathbf{x}}_{\bar{\mathbf{v}}}|0,\sigma_1\mathbf{I}_{p-q})\text{Bessel}(\tilde{\mathbf{x}}_{\mathbf{v}}|1/\alpha,(d-q)/2). \tag{4.9}$$

Proof. Let us first consider the case where all variables are active and assume that $\mathbf{v} = (1,1,...,1)$. Using Lévy's continuity theorem, $\boldsymbol{\varepsilon_2}$ weakly converges to zero when σ_2 vanishes. Since zero is a constant, this convergence also happens to be in probability (Van der Vaart, 2000, p. 10). The variable $\mathbf{x}$ therefore converges in probability to $\mathbf{Wy}$, which follows a $\text{Bessel}(1/\alpha,(d-q)/2)$ distribution according to the main result of Appendix A.

In the general case when $\mathbf{v}$ is not necessarily equal to $(1,1,...,1)$ we can prove (4.9) by invoking the independence between $\mathbf{x}_{\mathbf{v}}$ and $\mathbf{x}_{\bar{\mathbf{v}}}$, similarly to the proof of Theorem 1. □

The key tool used to prove this result is the exact distribution of the product of a Gaussian matrix and a Gaussian vector, which is derived in Appendix A. This theorem allows us to efficiently compute the noiseless marginal log-likelihood defined as

$$\mathcal{L}(\mathbf{X},\mathbf{v},\alpha,\sigma_1) = \sum_{i=1}^{n}\log p(\tilde{\mathbf{x}}_i|\mathbf{v},\alpha,\sigma_1).$$

It is worth noticing that Ando (2009) also obtained a closed-form expression for the marginal likelihood in the related, but different, context of factor analysis. More specifically, he considered heavy-tailed factors and a inverse Wishart prior for the (unconstrained) noise covariance matrix. Regarding hyper-parameter tuning, if we assume that $\mathbf{v}$ is known, the regularization parameter α can be optimized efficiently using univariate gradient ascent. In fact, as stated by next proposition, the marginal log-likelihood is even a strictly concave function of α.

Proposition 4.1. *The function $\alpha \mapsto \mathcal{L}(\mathbf{X},\mathbf{v},\alpha,\sigma_1)$ is strictly concave on $\mathbb{R}_+^*$.*

Proof. Since a sum of concave functions is concave, it is sufficient to prove that the function $g : \alpha \mapsto p(\tilde{\mathbf{x}}|\mathbf{v},\alpha,\sigma_1)$ is strictly concave. Up to unnecessary additive constants, we have for all $\alpha > 0$,

$$g(\alpha) = d\log\alpha + \log\left((\alpha||\tilde{\mathbf{x}}_{\mathbf{v}}||_2)^{\frac{q-d}{2}}K_{\frac{q-d}{2}}\left(||\tilde{\mathbf{x}}_{\mathbf{v}}||_2\alpha\right)\right).$$

Using standard results about Bessel functions derivatives (Abramowitz and Stegun, 1965, p. 376), it can be shown that

$$g'(u) = \frac{d}{\alpha} - ||\tilde{\mathbf{x}}_{\mathbf{v}}||_2 h(u),$$

where the h is the ratio

$$h(\alpha) = \frac{K_{\frac{q-d}{2}-1}\left(||\tilde{\mathbf{x}}_{\mathbf{v}}||_2\alpha\right)}{K_{\frac{q-d}{2}}\left(||\tilde{\mathbf{x}}_{\mathbf{v}}||_2\alpha\right)}.$$

As proven independently by Lorch (1967) and Hartman and Watson (1974), since $q-d \geq 0$, h is a increasing function on $\mathbb{R}_+^*$. Therefore g' is stricly decreasing and g is strictly concave. □

The unique optimal value $\hat{\alpha}$ can therefore be found easily using univariate convex programming.

The noise variance σ_1 can be estimated using (4.9) by computing the standard error of the variables which were not selected by $\mathbf{v}$. However, since model (4.3) is a particular instance of PPCA, it is possible to use any regular PPCA noise variance estimator. A discussion on which estimator to choose is provided in Section 4.2.7

4.2.4 High-dimensional inference through a continuous relaxation

In spite of the results of the previous subsection, maximizing the evidence, even in the noiseless case, is particularly difficult (because of the discreteness of $\mathbf{v}$ which can take 2^p possible values). We therefore consider a simple continuous relaxation of the problem by replacing $\mathbf{v}$ by a continuous vector $\mathbf{u} \in [0,1]^p$. This relaxation is close to the one we considered in the previous chapter in a sparse linear regression framework. Denoting $\mathbf{U} = \text{diag}(\mathbf{u})$, this relaxed model can be written as

$$\mathbf{x} = \mathbf{U}\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}. \tag{4.10}$$

We denote $\boldsymbol{\theta} = (\mathbf{u}, \alpha, \sigma)$ the vector of parameters. In order to maximize the evidence $p(\mathbf{X}|\boldsymbol{\theta})$, we adopt a variational approach (Bishop, 2006, chap. 10). We view $\mathbf{y}_1, ...\mathbf{y}_n$ and $\mathbf{W}$ as latent variables.

Given a (variational) distribution q over the space of latent variables, the variational free energy is given by

$$\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = -\mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta})] - H(q), \tag{4.11}$$

where H denotes the differential entropy, and is an upper bound to the negative log-evidence

$$-\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) - \text{KL}(q||p(\cdot|\boldsymbol{\theta})) \leq \mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}).$$

To minimize $\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta})$, similarly to Bishop (1999b) and Archambeau and Bach (2009), the following mean-field approximation is made on the variational distribution

$$q(\mathbf{Y}, \mathbf{W}) = q(\mathbf{Y})q(\mathbf{W}). \tag{4.12}$$

With this factorization, a variational expectation-maximization (VEM) algorithm can be derived. For the E-step, the variational posterior distribution q^*, which minimizes the free energy, is computed.

Proposition 4.2. *The variational posterior distribution of the latent variables which minimizes the free energy is given by*

$$q^*(\mathbf{Y}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \tag{4.13}$$

and

$$q^*(\mathbf{W}) = \prod_{k=1}^{p} \mathcal{N}(\mathbf{w}_k | \mathbf{m}_k, \mathbf{S}_k), \tag{4.14}$$

where, for all $i \in \{1, ..., n\}$ *and* $k \in \{1, ..., p\}$

$$\boldsymbol{\mu}_i = \frac{1}{\sigma^2} \boldsymbol{\Sigma} \mathbf{M}^T \mathbf{U} \mathbf{x}_i, \; \mathbf{m}_k = \frac{u_k}{\sigma^2} \mathbf{S}_k \sum_{i=1}^{n} x_{i,k} \boldsymbol{\mu}_i,$$

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I}_d + \frac{1}{\sigma^2} \mathbf{M}^T \mathbf{U}^2 \mathbf{M} + \frac{1}{\sigma^2} \sum_{k=1}^{p} u_k^2 \mathbf{S}_k, \; \mathbf{S}_k^{-1} = \alpha^2 \mathbf{I}_d + \frac{n u_k^2}{\sigma^2} \boldsymbol{\Sigma} + \frac{u_k^2}{\sigma^2} \boldsymbol{\mathcal{M}}^T \boldsymbol{\mathcal{M}},$$

$$\mathbf{M} = (\mathbf{m}_1, ... \mathbf{m}_p)^T \;\; \textit{and} \;\; \boldsymbol{\mathcal{M}} = (\boldsymbol{\mu}_1, ... \boldsymbol{\mu}_n)^T.$$

Proof. Variational distribution of the latent vectors. Using a standard result in variational mean-field approximations (Bishop, 2006, chap. 10), we can write

$$\ln q^*(\mathbf{y}) = \mathbb{E}_{q(\mathbf{W})}[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W} | \boldsymbol{\theta})]$$

which leads to the factorization $q^*(\mathbf{y}) = \prod_{i \leq n} q^*(\mathbf{y}_i)$. Then, for each $i \leq n$, we can write, up to unnecessary additive constants,

$$\ln q^*(\mathbf{y}_i) = \mathbb{E}_{q(\mathbf{W})}[\ln p(\mathbf{x}_i, \mathbf{y}_i, \mathbf{W} | \boldsymbol{\theta})] = \mathbb{E}_{q(\mathbf{W})} \left[\frac{-1}{2\sigma^2} ||\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{y}_i||_2^2 \right] - \frac{1}{2} ||\mathbf{y}_i||_2^2,$$

thus

$$\ln q^*(\mathbf{y}_i) = \frac{-1}{2\sigma^2} \mathbf{y_i}^T \mathbb{E}_{q(\mathbf{W})}[\mathbf{W}^T \mathbf{U}^2 \mathbf{W}] \mathbf{y}_i + \frac{1}{\sigma^2} \mathbf{y}_i^T \mathbb{E}_{q(\mathbf{W})}[\mathbf{W}]^T \mathbf{U} \mathbf{x}_i - \frac{1}{2} ||\mathbf{y}_i||_2^2,$$

which leads to the desired form.

Variational distribution of the loading matrix. Similarly, up to unnecessary additive constants,

$$\ln q^*(\mathbf{W}) = \frac{-1}{2\sigma^2} \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{y}_i)}[||\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{y}_i||_2^2] - \frac{\alpha^2}{2} \sum_{i=1}^{p} ||\mathbf{w}_i||_2^2,$$

$$\ln q^*(\mathbf{W}) = \sum_{i=1}^{n} \left(\frac{-1}{2\sigma^2} \sum_{j=1}^{p} u_j^2 \mathbf{w}_j^T \mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i \mathbf{y}_i^T] \mathbf{w}_j + \frac{1}{\sigma^2} \sum_{j=1}^{p} x_{i,j} u_j \mathbf{w}_j^T \mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i] \right) - \frac{\alpha^2}{2} \sum_{i=1}^{p} ||\mathbf{w}_i||_2^2,$$

and

$$\ln q^*(\mathbf{W}) = \sum_{i=1}^{p} \left(\frac{-1}{2\sigma^2} \sum_{j=1}^{p} u_j^2 \mathbf{w}_j^T \mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i \mathbf{y}_i^T] \mathbf{w}_j + \frac{1}{\sigma^2} \sum_{j=1}^{p} x_{i,j} u_j \mathbf{w}_j^T \mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i] \right) - \frac{\alpha^2}{2} \sum_{i=1}^{p} ||\mathbf{w}_i||_2^2,$$

which leads to the factorization $q^*(\mathbf{W}) = \prod_{j \leq p} q^*(\mathbf{w}_i)$ and to the desired expression. □

It is worth noticing that two factorizations arise naturally. The four equations of Proposition (4.2) will constitute the E-step of the VEM algorithm used to minimized the free energy.

We can now compute the negative free energy which will be maximized during the M-step.

Proposition 4.3. *Up to unnecessary additive constants, the negative free energy is given by*

$$\begin{aligned}-\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k| - np\ln\sigma + dp\ln\alpha - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) \\ - \frac{1}{2\sigma^2}\sum_{k=1}^{p}u_k^2\mathrm{Tr}[(n\boldsymbol{\Sigma} + \boldsymbol{\mathcal{M}}^T\boldsymbol{\mathcal{M}})(\boldsymbol{S}_k + \mathbf{m}_k\mathbf{m}_k^T)] + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i \\ + \sum_{k=1}^{p} -\frac{\alpha^2}{2}\mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T). \quad (4.15)\end{aligned}$$

Proof. By definition, we have

$$-\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = \mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta})] + H(q),$$

therefore

$$\begin{aligned}-\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = -np\ln\sigma - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}_q[\mathbf{y}_i\mathbf{W}^T\mathbf{U}^2\mathbf{W}\mathbf{y}_i] + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i \\ + \sum_{k=1}^{p}\left(d\ln\alpha - \frac{\alpha^2}{2}\mathbb{E}_q[\mathbf{w}_k^T\mathbf{w}_k]\right) - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_q[\mathbf{y}_i^T\mathbf{y}_i] + \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k|,\end{aligned}$$

and computing the expectations leads to

$$\begin{aligned}-\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = -np\ln\sigma + dp\ln\alpha - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{p}u_k^2\mathrm{Tr}[(\boldsymbol{\Sigma}+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k+\mathbf{m}_k\mathbf{m}_k^T)] \\ + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i + \sum_{k=1}^{p} -\frac{\alpha^2}{2}\mathrm{Tr}(\mathbf{S}_k+\mathbf{m}_k\mathbf{m}_k^T) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}(\boldsymbol{\Sigma}+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) + \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k|, \quad (4.16)\end{aligned}$$

which allows us to conclude. □

Minimizing the free energy leads to the following M-step updates

$$\alpha^* = \left(\frac{1}{dp}\sum_{k=1}^{p}\mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T)\right)^{-1/2}, \quad (4.17)$$

$$\sigma^* = \sqrt{\frac{\mathrm{Tr}(\mathbf{X}\mathbf{X}^T - 2\mathbf{X}\mathbf{U}\mathbf{M}\boldsymbol{\mathcal{M}})}{np} + \frac{1}{np}\sum_{i=1}^{n}\sum_{k=1}^{p}u_k^2\mathrm{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \mathbf{m}_i\mathbf{m}_i^T)]}, \quad (4.18)$$

and, for $k \in \{1, ..., p\}$,

$$u_k^* = \text{argmin}_{u \in [0,1]} \frac{u^2}{2\sigma^2} \sum_{i=1}^{n} \text{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \mathbf{m}_i \mathbf{m}_i^T)] - u \sum_{i=1}^{n} x_{i,k} \mathbf{m}_k^T \boldsymbol{\mu}_i. \tag{4.19}$$

Note that the objective function of the optimization problem (4.21) is simply a univariate polynomial. Denoting

$$\xi_k = \frac{\sum_{i=1}^{n} x_{i,k} \mathbf{m}_k^T \boldsymbol{\mu}_i}{\sum_{i=1}^{n} \text{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \mathbf{m}_i \mathbf{m}_i^T)]}, \tag{4.20}$$

the solution can be written as

$$u_k^* = \min\{\max\{\xi_k, 0\}, 1\}. \tag{4.21}$$

4.2.5 The GSPPCA algorithm

Once the VEM algorithm has converged, the continuous vector $\mathbf{u}$ still needs to be transformed into a binary one. To do so, we rely on a technique close to the one introduced in the previous chapter in a sparse linear regression framework. Specifically, the following simple procedure (summarized in Algorithm 3) is considered:

- a family of p nested models is built using the order of the coefficients of $\mathbf{u}$ as a way of ranking the variables. Specifically, for each $k \leq p$, the k-th element of this family is the binary vector $\mathbf{v}^{(k)}$ such that the k top coefficients of $\mathbf{u}$ are set to 1 and the others to 0.
- the marginal likelihood $\mathcal{L}$ of the noiseless model (computed using the formula of Theorem 3) is then maximized over this family of models.
- the model $\mathbf{v}$ with the largest marginal likelihood is kept.

Once the model is estimated, the globally sparse principal components of $\mathbf{X}$ can be computed by simply performing PCA on $\mathbf{X}_\mathbf{v}$. This type of post-processing is similar to the *variational renormalization* introduced by Moghaddam et al. (2005). In the case of local sparsity, variational renormalization can be achieved using an alternating maximization scheme (Journée et al., 2010). However, the global sparsity structure greatly simplifies this procedure by reducing it to performing PCA on the relevant variables. An implementation of the GSPPCA algorithm is available on Github (*https://github.com/pamattei/GSPPCA*).

4.2.6 Links with other sparsity-inducing Bayesian procedures

SPIKE-AND-SLAB MODELS Model (4.3) may be rewritten $\mathbf{x} = \tilde{\mathbf{W}}\mathbf{y} + \boldsymbol{\varepsilon}$ where $\tilde{\mathbf{W}} = \mathbf{V}\mathbf{W}$. The prior distribution for the parameter $\tilde{\mathbf{W}}$ is similar to the spike-and-slab prior introduced by Mitchell and Beauchamp (1988) in a linear regression framework. Indeed, each coefficient $\tilde{w}_{ij}$ follows *a priori* either a Dirac distribution with mass at zero (if $v_i = 0$) which is usually called the *spike* or a Gaussian distribution with variance $1/\alpha^2$ (if $v_i = 1$) which is usually called the *slab*. However, contrary to standard spike-and-slab models which would assume a

Algorithm 3 GSPPCA algorithm for unsupervised variable selection

Input: data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, dimension of the latent space $d \in \mathbb{N}^*$
Output: sparsity pattern $\mathbf{v} \in \{0,1\}^p$

// VEM algorithm to infer the path of models
Initialize $\mathbf{u}, \alpha, \sigma, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n, \mathbf{m}_1, ..., \mathbf{m}_p, \mathbf{S}_1, ..., \mathbf{S}_p$ and $\boldsymbol{\Sigma}$
repeat
  E-step from Proposition 4.2
  M-step from equations (4.17),(4.18),(4.21)
until convergence of the free energy

// Model selection using the exact marginal likelihood
Compute σ_1
for $k = 1$ to p **do**
  Compute $\mathbf{v}^{(k)}$
  Find $\alpha_k = \arg\max_{\alpha>0}\{\alpha \mapsto \mathcal{L}(\mathbf{X}, \mathbf{v}^{(k)}, \alpha, \sigma_1)\}$ using gradient ascent
end for

$q = \arg\max_{1 \leq k \leq p} \mathcal{L}(\mathbf{X}, \mathbf{v}^{(k)}, \alpha_k, \sigma_1)$
$\mathbf{v} = \mathbf{v}^{(q)}$

product of Bernoulli prior distributions over $\mathbf{v}$, we see $\mathbf{v}$ here as a deterministic parameter to be inferred from the data. It is worth noticing that spike-and-slab priors have already been applied to locally sparse PCA by Lázaro-Gredilla and Titsias (2011) and Mohamed et al. (2012).

AUTOMATIC RELEVANCE DETERMINATION Introduced in the context of feedforward neural networks (MacKay, 1994; Neal, 1996), automatic relevance determination (ARD) is a popular empirical Bayes procedure to induce sparsity. ARD was applied to Bayesian PCA models together with VEM algorithms in order to obtain automatic dimensionality selection (Bishop, 1999b) of local sparsity (Archambeau and Bach, 2009). In order to obtain global sparsity, ARD may be built using Model (5.1) together with Gaussian priors $\mathbf{w}_i \sim \mathcal{N}(0, a_i \mathbf{I}_d)$ for $i \in \{1, ..., p\}$. Similarly to Tipping (2001), maximizing the marginal likelihood would discard irrelevant variables by leading several variance parameters a_i to vanish. Interestingly, this model is somehow related to the relaxed GSPPCA model. Indeed the relaxed model (4.10) assumes that the i-th line of the loading matrix $\mathbf{UW}$ follows *a priori* a $\mathcal{N}(0, u_i^2/\alpha^2 \mathbf{I}_d)$ distribution. The relaxed model will consequently inherit the good properties of ARD – listed for example by Wipf et al. (2011). However, similarly to the previous chapter, using the exact marginal likelihood to eventually obtain a sparse solution will avoid many classical drawbacks of ARD. First, as pointed out by Wipf and Nagarajan (2008), convergences of EM algorithms are extremely slow in the case of the ARD models. However, with our approach, since we only need the *ordering* of the coefficients of $\mathbf{u}$, we do not have to wait for the complete convergence of this parameter. In practice, in all the experiments that we carried out, we only had to perform less than a few hundreds of iterations of the algorithm

to obtain convergence of the free energy in order to perform variable selection. It is worth mentioning that the fact that the objective function converges faster than the parameters of the model is a quite general property of EM algorithms (Xu and Jordan, 1996). Our procedure also avoids the lack of flexibility of ARD by computing posterior probabilities of models rather than simply giving an estimate of the best sparse model. Combined with a greedy technique similar to Occam's window (Madigan and Raftery, 1994), this feature could allow for example to perform Bayesian model averaging, which is not possible with ARD. Eventually, in the context of Bayesian PCA, ARD models such as the ones of Bishop (1999a,b) or Archambeau and Bach (2009) have to rely on approximations of the marginal likelihood while we use an exact expression.

4.2.7 Computational considerations

INTRINSIC DIMENSION ESTIMATION Since model (4.3) is a particular instance of PPCA, any intrinsic dimension estimator for PCA can be applied to estimate beforehand the intrinsic dimension d (see e.g. Sobczyk et al. (2017) for a recent overview of existing estimators). Although the problem of finding d is of critical importance, we assume in this work that a reasonable choice of dimension has already been made by the practitioner. While it could be tempting to use the exact noiseless marginal likelihood to select d, the close relationship existing between the noise level and d in PPCA (Tipping and Bishop, 1999; Nakajima et al., 2011) suggests that loosing the noise information is likely to be prejudicial for intrinsic dimension estimation.

ESTIMATION OF THE NOISE VARIANCE As mentioned in Section 4.2.3, the standard error σ_1 of irrelevant predictors can be estimated using any regular PPCA estimator. Specifically, three important estimators are considered: the maximum likelihood estimator (Tipping and Bishop, 1999), its unbiased correction (Passemier et al., 2017), or simply the median of the variances of all features (Johnstone and Lu, 2009). Since the ML estimator is known to be biased in the high-dimensional regime, it is usually preferable to use its bias-corrected version. Both of these estimators can also be computed using the singular value decomposition (SVD) of $\mathbf{X}$. Note that since the median estimator does not need to perform this decomposition, it is therefore more suitable for large-scale inference.

INITIALIZATION STRATEGIES FOR THE VEM ALGORITHM Regarding the initialization of the relaxed model parameter $\mathbf{u}$, we chose to initialize all its coefficients to one. This allows to avoid premature vanishing of these coefficients which is a common drawback of ARD-like techniques (Wipf and Nagarajan, 2008). The noise standard error can be simply initialized using any classical PPCA noise estimator. Similarly to the previous chapter, the slab precision parameter α controls the sparsity of the VEM solution and a too small initial value is likely to lead to a too sparse solution such as the useless local optimum $\mathbf{u} = 0$. Following Biernacki et al. (2003), we chose to perform short VEM runs (with less than 5 iterations) on a small grid (typically $\alpha \in \{0.1, 1, 10\}$) and to select the value of α that led to the lowest free energy. The posterior means of the PCA loadings $\mathbf{m}_1, ..., \mathbf{m}_p$ and of the corresponding

scores $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n$ can be initialized using the singular vectors of $\mathbf{X}$. If the size of the data forbids to perform this SVD, using random Gaussian coefficients as starting points does not significantly alter the results. Finally, the initial values chosen for the posterior covariance matrices are $\boldsymbol{\Sigma} = \mathbf{I}_d$ and $\mathbf{S}_1 = ... = \mathbf{S}_p = \alpha^{-2}\mathbf{I}_d$.

COMPUTATIONAL COST OF VEM ITERATIONS Thanks to the factorizations that arised naturally during variational inference, the cost of each VEM iteration is of order $O(pnd^3)$ which is linear *both in sample size and dimensionality* and therefore particularly suitable for high-dimensional inference.

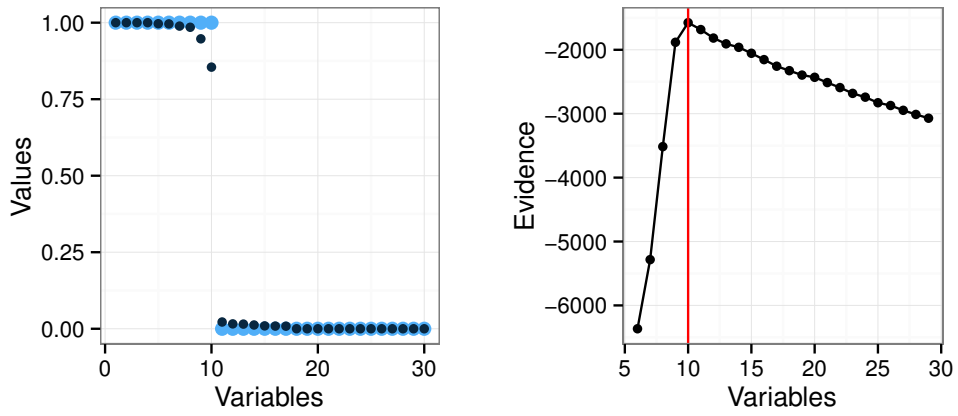
LARGE SCALE INFERENCE In the GSPPCA algorithm, SVD is used twice. Indeed, the top d singular vectors can be used to initiate the VEM algorithm and the $p - d$ smallest singular values can be used to estimate the noise variance (both as a VEM starting point for σ and as an estimator for σ_1). This can be done efficiently using a *truncated SVD algorithm*. We chose specifically the R interface (Qiu and Mei, 2016) of the Spectra[1] C++ library. However, for very large scale problems, even a fast truncated SVD algorithm appears computationally prohibitive. To tackle this issue, we offer two alternatives. First, the posterior parameters initialized using the eigenvectors can be initialized using random standard Gaussian coefficients. Moreover, following Johnstone and Lu (2009), the noise variance can be estimated using the median of the variable variances. This leads to a "SVD-free" version of the GSPPCA algorithm suitable for very large scale problems.

MODEL SELECTION SPEEDUP The model selection step of the GSPPCA algorithm requires to perform p univariate gradient ascents, which can be computationally expensive when p is large. A simple way to reduce the number of gradient ascents is to rely on the links between our relaxed model and ARD. Specifically, we can discard *before* the model selection step all the variables corresponding to the subset $\{i \in \{1, ..., p\} | u_i = 0\}$ where $\mathbf{u}$ is the relaxed model parameter obtained after convergence of the VEM algorithm. When $\mathbf{u}$ is sparse, this will bring about a substantial speedup. Notice that, since ARD is known to converge slowly, $\mathbf{u}$ is unlikely to be sparse enough and the model selection step is still necessary.

EVALUATION OF BESSEL FUNCTIONS The modified Bessel function of the second kind, which is used to compute the exact marginal likelihood and it gradient with respect to α, can be delicate to compute as soon as its order or its argument is large. In our experiments, we tackled this issue by using an asymptotic expansion based on Debye polynomials (Abramowitz and Stegun, 1965, formula 9.8.7). This is in particular implemented in the R package `Bessel` (Mäechler, 2013). We found this approximation to be extremely accurate in all the experiments that we carried out.

[1] http://yixuan.cos.name/spectra/index.html

Figure 4.1 – Variable selection with GSPPCA on the introductory example.



4.3 Numerical simulations

This section aims at highlighting the specific features and abilities of the proposed GSPPCA approach on simulated and real data sets.

4.3.1 An introductory example

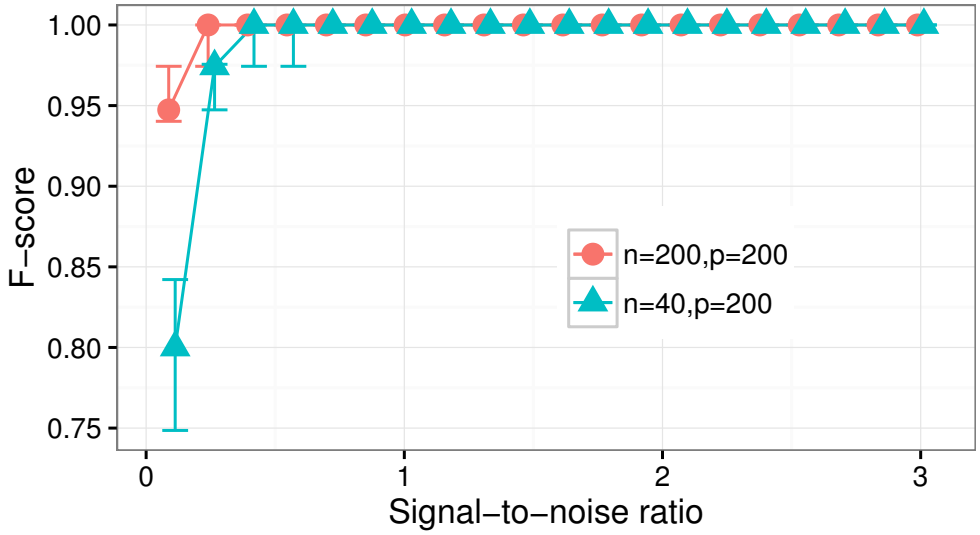
We consider here a simple introductory example to illustrate the proposed combination between a relaxed VEM algorithm and the closed-form expression of the marginal likelihood. For this experiment, $n = 50$ observations are simulated according to (4.3) with $p = 30$, $d = 5$ and $q = 10$. Each coefficient of $\mathbf{W}$ is drawn at random according to a standard Gaussian distribution and the noise variance is equal to 0.1. Figure 1 presents the results of GSPPCA on this toy data set. The left panel presents in dark blue the coefficients of the estimated $\mathbf{u}$ obtained after running the VEM algorithm (sorted in decreasing order) and the corresponding true values of $\mathbf{v}$ (pale blue points) used in the simulations. The right panel shows the values of evidence computed on the family of models inferred by the order of the coefficients of $\mathbf{u}$. On this simple example, $\mathbf{u}$ captures the true ranking of the variables and the model with the largest evidence is actually the true one.

4.3.2 Range of the noiseless assumption

In all the experiments that we carried out, since the noiseless PPCA model is not a true generative p-dimensional model (the random variable $\tilde{\mathbf{x}}$ belongs to a strict subspace of $\mathbb{R}^p$), we chose not to use it to generate data in our experiments. We rather chose the more realistic and natural Model (4.3). Since this model includes a nonzero noise, it is important to know the limits of the noiseless assumption.

We therefore simulated two scenarios according to Model (4.3): a first one with $n = 40$ observations and a second one with $n = 200$. In both scenarios, $p = 200$, $d = 10$, $q = 20$, and

Figure 4.2 – Median, first and third quartiles of the F-score for different noise levels, based on 100 replications



each coefficient of $\mathbf{W}$ is drawn according to a standard Gaussian distribution. The sparsity pattern chosen is simply

$$\mathbf{v} = (\overbrace{1, ..., 1}^{20 \text{ times}}, \overbrace{0, ..., 0}^{180 \text{ times}})^T. \tag{4.22}$$

In this simple simulation scheme, the signal-to-noise ratio (SNR) may be defined as $\text{SNR} = \frac{1}{p\sigma^2}\mathbb{E}_{\mathbf{W}}[(\mathbf{VW})^T\mathbf{VW}]p\sigma^2 = \frac{dq}{p\sigma^2}$. We chose a linear grid of 20 SNR ranging from 0.1 (most difficult scenario) to 3 (easiest scenario) and generated 100 datasets for each noise level. To evaluate the quality of the variable selection, we computed the F-score between $\hat{\mathbf{v}}$ and $\mathbf{v}$ on 100 runs. We recall that the F-score is the harmonic mean of precision and recall, and is closer to 1 when the selection is faithful. Unsurprisingly, when the SNR gets close to zero, the quality of the variable selection diminishes. However, GSPPCA appears to be quite robust to noise, even though the data are not generated according to the underlying noiseless model. Indeed, even in the case where $n = 40$, we observe an almost perfect recovery as long as SNR>0.5.

4.3.3 Model selection

In this subsection, we compare the model selection accuracies of two global methods – GSPPCA, SSPCA (Jenatton et al., 2009) – and a local one – SPCA (Zou et al., 2006).

SIMULATION SETUP While the simple simulation setup of Subsection 4.3.2 conveniently allowed to compute the SNR in closed formed in order to assess the range of the noiseless assumption, we introduce here a more realistic scheme by considering a finer correlation structure as well as a non-Gaussian noise. Specifically, first we generate n i.i.d observations

$(\mathbf{z}_1, ..., \mathbf{z}_n)$ following multivariate normal distribution $\mathcal{N}(0, \mathbf{R})$ where $\mathbf{R} = \text{diag}(\mathbf{R}_1, ..., \mathbf{R}_4)$ is a 4-blocks diagonal matrix where R_ℓ is such that $r_{\ell ii} = 0.3$ and $r_{\ell ij} = \rho$ for $i, j = 1, \ldots, p/4$ and $i \neq j$. Then, a globally sparse PCA model is obtained as followed. First, PPCA is performed on the sample $(\mathbf{z}_1, ..., \mathbf{z}_n)$, which leads to a non-sparse ML estimate $\mathbf{W}_{\text{ML}}$ for the loading matrix. Then, given a sparsity pattern $\mathbf{v} \in \{0, 1\}^p$ and denoting $\mathbf{V} = \text{diag}(\mathbf{v})$ as before, the loading matrix matrix is "globally sparsified" by considering $\mathbf{V}\mathbf{W}_{\text{ML}}$. The final observations are eventually generated according to the non-noiseless model

$$\forall i \leq n, \quad \mathbf{x}_i = \mathbf{V}\mathbf{W}_{\text{ML}}\mathbf{y}_i + \boldsymbol{\varepsilon}. \tag{4.23}$$

The simple sparsity pattern (4.22) is kept and the vectors $\mathbf{y_1}, ..., \mathbf{y}_n$ are standard Gaussian as in regular PPCA. Regarding the noise term $\boldsymbol{\varepsilon}$, we consider two scenarios. A first one with Gaussian noise and a second one with Laplacian noise, both centered with unit variance. We choose $p = 200$, $d = 10$, $q = 20$ and consider five cases for the sample size: $n = p/5$, $p/4$, $n = \lfloor p/3 \rfloor$, $n = p/2$ and $n = p$. More classical $n > p$ cases are not presented here since regular PCA is known to perform well in this context and variable selection thus may not be of great use (Johnstone and Lu, 2009). Each experiment was repeated 50 times.

MODEL SELECTION CRITERIA Regarding SSPCA, we used the Matlab code available at the main author's webpage and chose the tuning parameter using 5-fold cross-validation on the reconstruction error. We constrained the algorithm in order to obtain globally sparse solutions. For SPCA, we used the `elasticnet` R package and an *ad-hoc* method by selecting enough variables to explain 99% of the total variance. We also tried to apply another globally sparse algorithm, vsnPCA-ℓ_0 from Ulfarsson and Solo (2011). However, their use of the Bayesian information criterion (BIC) led to selecting very few variables. This is not very surprising: since BIC is an asymptotic sparsity criterion, it is thus likely to perform poorly when p is larger than n.

RESULTS Tables 4.1 and 4.2 reports the mean and standard error of the F-score for the experiments described is this subsection. The two globally sparse methods vastly outperform SPCA, which is unable to identify the particular structure of the data. When p is larger than $n/2$, both globally sparse algorithms perform very well, GSPPCA being slightly better in the Gaussian noise case. It is not surprising to see SSPCA adapt efficiently to Laplacian noise because cross-validation is a model-free technique and is more likely to outperform model-based techniques when the data is not generated according to the model distribution. However, when n is smaller than $p/2$, GSPPCA significantly outperforms SSPCA in both noise scenarios. This reminds the fact that, in many $p \gg n$ situations, Bayesian model selection empirically outperforms ℓ_1-based methods (Celeux et al., 2012).

4.3.4 Global versus local

Here, we illustrate on real data sets how using GSPPCA instead of computing the leading sparse principal component for model selection can lead to selecting more relevant variables – i.e variables that retain more variance or are more interpretable.

Table 4.1 – F-score×100 for the model selection experiment of Section 4.3.3 with Gaussian noise

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
SPCA	20.7 ± 0.7	21.2 ± 0.7	21.5 ± 0.7	21.7 ± 0.5	25.2 ± 2.1
SSPCA	66.7 ± 21.4	71.5 ± 20	86.7 ± 14.2	95.6 ± 8.9	98.2 ± 7.2
GSPPCA	$\mathbf{86.8 \pm 7.06}$	$\mathbf{93.9 \pm 3.66}$	$\mathbf{97.2 \pm 2.55}$	$\mathbf{99.2 \pm 1.4}$	$\mathbf{100 \pm 0}$

Table 4.2 – F-score×100 for the model selection experiment of Section 4.3.3 with Laplacian noise

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
SPCA	20.8 ± 0.6	21.3 ± 0.6	21.6 ± 0.8	21.8 ± 0.6	25.3 ± 1.7
SSPCA	60.6 ± 22.4	63.9 ± 25.2	82.7 ± 18.1	$\mathbf{94.2 \pm 10.2}$	97.4 ± 9.5
GSPPCA	$\mathbf{74.2 \pm 10}$	$\mathbf{77.6 \pm 9.09}$	$\mathbf{79.7 \pm 8.38}$	88 ± 5.95	$\mathbf{99.2 \pm 1.4}$

EXPLAINED VARIANCE We consider the data set from the `breastCancerVDX` R package (Schroeder et al., 2011), consisting in expression levels of $p = 5391$ genes for $n = 344$ breast cancer patients. This data set contains the gene expression data published by Wang et al. (2005) and Minn et al. (2007). It contains expression levels of 22.283 probes for 344 patients. In order to be able to provide an interpretation of feature selection, we reduced the data from probe-level to gene-level using the following procedure:

- first, the probes with no gene identifier were discarded
- then, the data was aggregated to gene-level using the `collapseRows` R function of Miller et al. (2011),
- among the genes obtained, only the genes listed in the Reactome database (Fabregat et al., 2016) were kept in order to eventually perform pathway enrichment,
- finally, the data was centered but not standardized.

The resulting data matrix contains 5391 variables (genes) and 344 observations (patients).

Given a cardinality q, we applied four methods to select relevant genes:

- we computed the first q-sparse principal component using SPCA (Zou et al., 2006) and GSPPCA with $d = 1$
- we computed the support of the globally q-sparse subspace of dimension $d = 10$ using GSPPCA and SSPCA.

For each method, we projected the data onto a 10-dimensional globally q-sparse subspace using the sparsity pattern found by the algorithm and computed the percentage of explained variance using the criterion introduced by Shen and Huang (2008) – for each method, we applied the post-processing technique of Moghaddam et al. (2005). The results are plotted on Figure 4.3. GSPPCA with $d = 1$ outperform its local competitor SPCA by a significant margin, which means that the VEM algorithm finds more relevant genes than ℓ_1 approach of Zou et al. (2006) – this is consistent with the experiments of Archambeau and Bach (2009). Both global methods explain consistently more variance than local ones. This fact

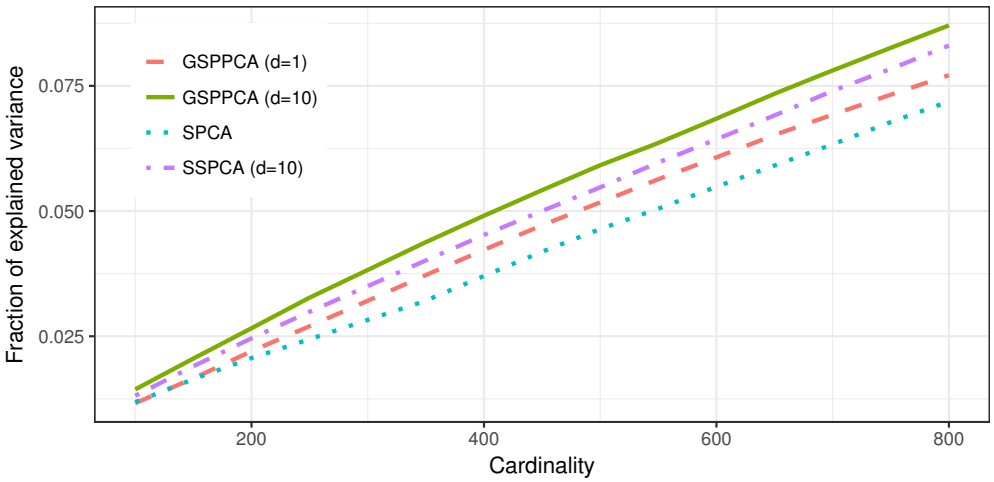


Figure 4.3 – Percentage of variance explained by projecting the data onto a 10-dimensional globally sparse subspace

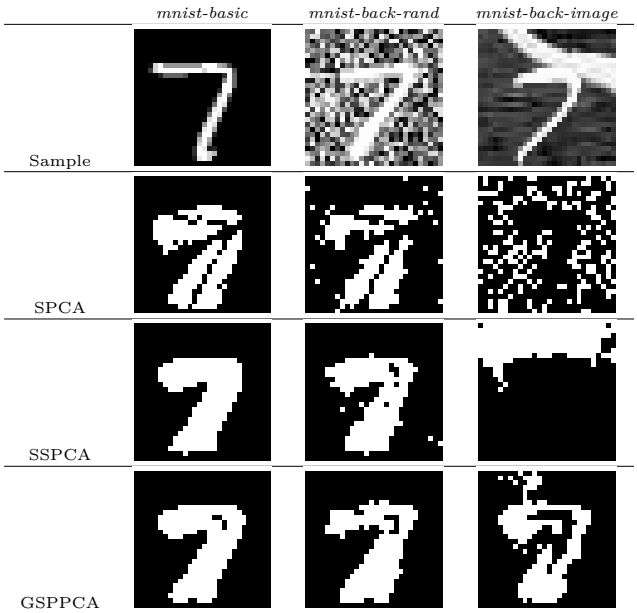
is not surprising since the data is indeed projected onto a globally sparse subspace, but the significance of this variance gap highlights the fact that different dimensions lead to very different sparsity patterns. This means that projecting the data onto a single sparse axis is likely to lead to an important information loss (this fact is confirmed in Section 4.5). The variables selected by GSPPCA retain significantly more variance than the ones selected by SSPCA, and may consequently be of superior interest.

INTERPRETABILITY Inspired by Hastie et al. (2015, section 8.2.3.1), we consider the problem of learning which features are relevant on three data sets of handwritten digits. We consider $n = 500$ gray-scale images (with $p = 758$ pixels) of handwritten sevens from three data sets introduced by Larochelle et al. (2007):

- *mnist-basic* which is simply a subsample of sevens from the original MNIST data set,
- *mnist-back-rand* in which random backgrounds were inserted in the images. Each pixel value of the background was generated uniformly between 0 and 255,
- *mnist-back-image* in which random patches extracted from a set of 20 grey-scale natural images were used as backgrounds for the sevens.

On these three data sets, we apply SPCA (with $d = 1$), SSPCA and GSPPCA (both with $d = 100$) in order to select $q = 200$ relevant pixels. On *mnist-basic*, even if SPCA's result is a little bit more erratic than the two others, all selections are interpretable and we can easily recognize a seven. On *mnist-back-rand* however, while the two globally sparse selections are still consistent, SPCA's pixels are more scattered and it is harder to recognize the shape of a seven. Eventually, on *mnist-back-image*, GSPPCA's selection is less smooth but a seven can still be recognized, whereas SPCA appears to randomly select pixels *almost everywhere*

Table 4.3 – Variable selection of SPCA and GSPPCA for the three datasets of Larochelle et al. (2007), selected variables are in white



but near the mean seven. SSPCA seems to notice that the zone occupied by the upper bars of the sevens is of interest, but its selection does not appear interpretable.

4.4 Application to signal denoising

In this section, we focus on a first possible application of GSPPCA for signal denoising through the sparsification of a wavelet decomposition. PCA is indeed a popular way to denoise multivariate signals (Aminghafari et al., 2006; Johnstone and Lu, 2009). To illustrate the potential interest of GSPPCA in this context, we consider hereafter two simulation scenarios, each using a specific form of signal and wavelet. The simulation scenarios are as follows:

- Scenario A: it consists in a square wave signal with 6 states of different lengths. The observed signal is sampled with a time step of 5×10^{-3} with an additional Gaussian noise with zero mean and 0.2 standard deviation. The Haar wavelet is used here for signal reconstruction.
- Scenario B: the original signal is here a mixture of 4 Gaussian densities. The observed signal is also sampled with a time step of 5×10^{-3} with an additional Gaussian noise with zero mean and 0.2 standard deviation. The Daubechies D8 wavelet is used here for signal reconstruction.

Figure 4.4 presents the original signals and observed signals for scenarios A and B. In both cases, $n = 100$ signals were sampled during the training phase and decomposed as

Scenario	Wavelet	PCA	tPCA	SPCA	GSPPCA
A	9.516±0.819	2.719±0.439	2.484±0.372	2.480±0.371	**2.283±0.344**
B	8.156±0.725	1.390±0.351	1.253±0.343	1.406±0.354	**1.193±0.337**

Table 4.4 – Reconstruction error (sum of squared errors) for wavelet signal denoising on the two simulation scenarios (results are averaged on 50 signal reconstructions). Standard deviations are also provided.

$p = 175$ wavelet coefficients. For signal denoising, GSPPCA is applied on the $n \times p$ wavelet coefficient matrix to extract $d = 10$ globally sparse principal axes. Then, a new sampled signal is projected on those extracted principal axes and back-projected in the original wavelet domain. It is worth mentioning that the estimated value for $q = \|v\|_0$ is 17 on scenario A and 15 on scenario B.

As an illustration, we plotted on Figure 4.4 the denoising results for newly sampled signals A and B with GSPPCA. We used the same projection-reconstruction protocol for PCA, thresholded PCA (PCA loading smaller than 1×10^{-3} are set to 0) and SPCA (λ is chosen such that 99% of the PCA projected variance is conserved). Denoising results obtained with those methods are also supplied on Figure 4.4. First, on both signal A and B, PCA achieves a very satisfying denoising and thus confirms his validity in this context. One can also show that a simple thresholding of the PCA loadings allows a clear denoising improvement and turns out to be competitive with the one performed by SPCA. The SPCA result is here somehow disappointing due to the fact that the sparsity is not global and most wavelet levels stay active in the final reconstruction. Finally, the global sparsity of GSPPCA retains only a few wavelet levels and achieves here the best reconstruction in both scenarios.

Finally, Table 4.4 presents the reconstruction error (sum of squared errors) averaged on 50 test signal reconstructions, on the two simulation scenarios. The results confirms the observations made on Figure 4.4. GSPPCA achieves particularly good performances on both scenarios and thus imposes itself as a competitive tool for signal denoising. Moreover, the GSPPCA reconstruction uses fewer wavelet levels and is therefore visually smoother.

4.5 Application to unsupervised gene selection

Considering again the breast cancer data set previously studied in Section 4.5, we address here the issue of the biological significance of the selected genes. To this end, we will use the *pathway enrichment index* (PEI) introduced by Teschendorff et al. (2007) and used in a sparse PCA framework by (Journée et al., 2010).

4.5.1 Pathway enrichment as a measure of biological significance

In this subsection, we briefly review how the PEI can be computed in order to evaluate the quality of a given subset of genes. For more details on the PEI, see Teschendorff et al. (2007) or Journée (2009), and on hypergeometric tests and enrichment, see Rivals et al. (2007).

Suppose that using a microarray data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ where each variable corresponds to a gene, an algorithm infers a subset $\mathbf{s} \subset \{1, ..., p\}$ of genes. A way to assess its biological

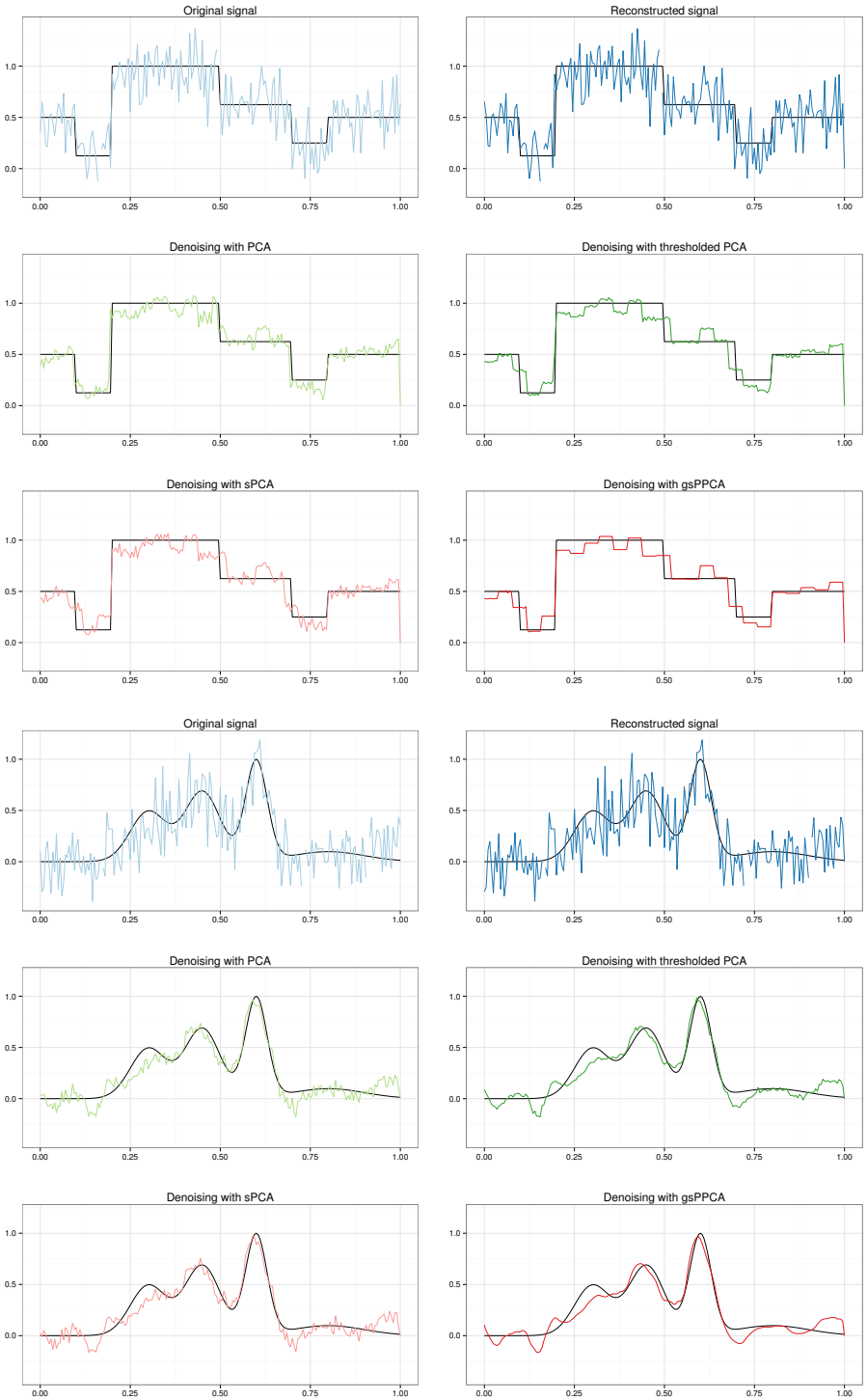


Figure 4.4 – Denoising results for signals A (top) and B (bottom) with PCA, thresholded PCA, SPCA and GSPPCA.

Table 4.5 – PEI for several fixed cardinalities

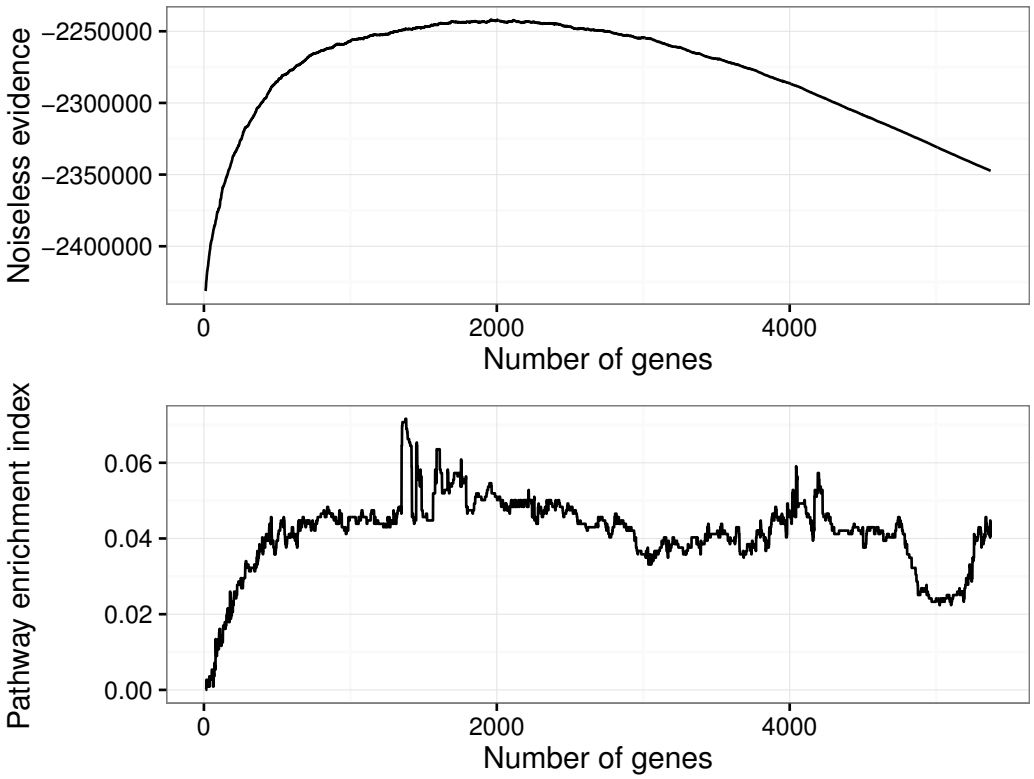
Cardinality		tPCA	SPCA	GSPPCA
290	*selected by tPCA*	0.09	0.09	**3.22**
1000		1.88	1.88	**4.57**
1965	*selected by GSPPCA*	1.7	1.61	**5.19**
3000		1.16	1.43	**3.58**
4466	*selected by SPCA*	3.04	3.22	**4.29**
5000		1.79	1.88	**2.42**

significance is to compare $\mathbf{s}$ to many other subsets *which are known to be biologically relevant*. In this case, the biologically relevant subsets are defined by *biological pathways*, and are therefore groups of genes involved in series of biochemical reactions linked to a certain biological function. Let us denote these known subsets $\mathbf{b}_1, ..., \mathbf{b}_N \subset \{1, ..., p\}$. For our breast cancer experiment, we use the $N = 1116$ pathways from the Reactome database (Fabregat et al., 2016) included in the R package `reactomePA` (Yu and He, 2016). For $k \leq N$, the *enrichment* of $\mathbf{s}$ in the k-th pathway of this list is the statistical significance of its overlap with $\mathbf{b}_k$, evaluated using the *hypergeometric test*. More specifically, for each $k \leq N$, the null hypothesis of this test is that the genes in $\mathbf{s}$ are chosen uniformly at random from the total gene population. Under this hypothesis, the test statistic $\#(\mathbf{s} \cap \mathbf{b}_k)$ follows a hypergeometric distribution and a p-value can be computed to assess the statistical significance of the overlap. Because we are conducting one test for each pathway considered, these p-value are then adjusted using the Benjamini-Hochberg procedure to control the false discovery rate (Benjamini and Hochberg, 1995). The subset $\mathbf{s}$ is eventually declared enriched for a certain pathway if the adjusted p-value of the corresponding hypergeometric test is lower than 0.01. The PEI is finally defined as the percentage of enriched pathways in the Reactome family.

4.5.2 Results

We compare in Table 4.5 the PEI obtained by GSPPCA with $d = 10$, SPCA and thresholded PCA for several fixed cardinalities. Similarly to Zou et al. (2006), the two local methods are computing a single sparse axis. As in Journée et al. (2010) SPCA appears to give slightly better results than thresholded PCA. GSPPCA significantly outperforms the two other methods. This means that the genes selected by GSPPCA are consistently more associated with the Reactome pathways, and are therefore more interpretable. This highlights the fact that projecting the data onto a globally sparse subspace of dimension higher than one leads to significantly more interpretable and biologically plausible results. Regarding the estimation of the sparsity level, choosing the one that explains 99% of the variance led SPCA to selecting 4466 genes, which is difficult to interpret. For thresholded PCA, we selected the sparsity level using a criterion proposed by Teschendorff et al. (2007). Even though it led to the sparsest solution, its PEI was very small. Regarding GSPPCA, the noiseless marginal log-likelihood and the PEI of the corresponding models are plotted on Figure 4.5. We can see

Figure 4.5 – Marginal likelihood and PEI for the gene selection problem



that the marginal likelihood peak corresponds to highly interpretable genes: more than 5% of the biological pathways in the Reactome family have a significant overlap with the genes selected by GSPPCA. Furthermore, models with a lower marginal likelihood have generally a lower PEI. To a certain extend, this shows that our marginal likelihood expression can stand as an indicator of biological significance.

4.6 Conclusion

Unsupervised feature selection is an hazy and exciting problem. It becomes particularly difficult and ill-posed when no specific learning task (such as clustering) is driving it. We have proposed in this chapter a new method for unsupervised feature selection based on the idea that the data may lie close to a subspace of moderate dimension spanned by a basis with a shared sparsity pattern. On several real data sets, this approach outperforms a popular method which consists in finding the sparsity pattern of the single leading principal vector of the data. These results suggest that, on many real-life high-dimensional data sets, an important part of the information cannot be captured by one-dimensional subspace approximations.

While building our framework, we derived the first closed-form expression of the marginal

likelihood of a Bayesian PCA model, using the noiseless model of Roweis (1998). Regarding future work, it would be interesting to see if more complex priors can be used. In the next chapter, we investigate how this closed-form expression can allow to perform intrinsic dimension estimation for PCA.

5

Exact Dimensionality Selection for Bayesian PCA

5.1 Introduction

The computer age is characterized by a surge of multivariate data, which is often difficult to explore or describe. A natural way to deal with such datasets is to reduce their dimensionality in a interpretable way, trying not to loose too much information. Accordingly, a wide range of dimension-reduction techniques have been developed over the years. Principal component analysis (PCA), perhaps the earliest of these techniques, remains today one of the most widely used (Jolliffe and Cadima, 2016). Introduced by Pearson (1901) and rediscovered by Hotelling (1933) in the beginning of the twentieth century, PCA has had indeed

an ubiquitous role in statistical analysis since the introduction of electronic computation in the 1950s. Recent exemples include climate research (Hannachi et al., 2006), genome-wide expression studies (Ringnér, 2008), massive text mining (Zhang and El Ghaoui, 2011), and deep learning (Chan et al., 2015). For a more exhaustive overview of past applications of PCA, we defer the reader to the monograph of Jolliffe (2002) or the recent review paper of Jolliffe and Cadima (2016).

Specifically, PCA consists in a simple procedure: the practitioner orthogonally projects his multivariate data on a space spanned by the eigenvectors associated with the largest eigenvalues of the empirical covariance matrix. The dimension of the representation learnt in this way is simply the number of eigenvectors – called principal components (PCs) – kept for the projection. However, it may come as a surprise that in spite of the popularity of this method, no authoritative solution has been widely accepted for choosing how many PCs should be computed. Common practice is to choose this dimension by considering the eigenvalues scree of the sample covariance matrix. This ad-hoc technique, popularized by Cattell (1966), has been largely modified and perfected over the last fifty years (Jackson, 1993; Zhu and Ghodsi, 2006), and is often chosen when PCA is used as a building block within a larger algorithmic framework – see e.g. Bouveyron et al. (2007b) for an example in cluster analysis or Evangelopoulos et al. (2012) in latent semantic analysis. However, more refined approaches have also been developed. Earlier works were based on hypothesis testing (Jolliffe, 2002, Section 6.1.4). Cross-validation, suggested by Wold (1978) and developed over the years (Bro et al., 2008), is known to be effective in a wide variety of settings (Josse and Husson, 2012). Another fruitful line of work follows the seminal article of Tipping and Bishop (1999), who recast PCA as a simple inferential problem. Their model, called probabilistic PCA (PPCA), led to several model-based methods for dimensionality selection, both from frequentist (Ulfarsson and Solo, 2008b; Bouveyron et al., 2011; Passemier et al., 2017) and Bayesian (Bishop, 1999a; Minka, 2000; Hoyle, 2008; Sobczyk et al., 2017) perspectives.

Most of the aforementioned methods are based on asymptotic considerations. However, it was recently proven that, in an asymptotic framework, hard thresholding the eigenvalues surprisingly suffices to provide an optimal dimensionality (Gavish and Donoho, 2014). Thus, the path to more efficient schemes for finding the number of PCs goes through the study of non-asymptotic criteria, which have been overlooked in the past. A natural non-asymptotic answer is provided by exact Bayesian model selection, which was previously used at the price of computationally expensive Markov chain Monte Carlo (MCMC) sampling (Hoff, 2007). We present here a prior structure based on the PPCA model that allows us to exhibit a closed-form expression of the marginal likelihood, leading to an efficient algorithm that selects the number of PCs without any asymptotic assumption. Specifically, we rely on a normal prior distribution over the loading matrix and a gamma prior distribution over the noise variance. Imposing a simple constraint on the hyperparameters of the respective distributions, we show that this allows the data to marginally follow a generalized Laplace distribution, leading to an efficient closed-form computation of the marginal likelihood. We also propose a heuristic based on the expected shape of the marginal likelihood curve in order to choose hyperparameters. With simulated data, we demonstrate that our approach is competitive compared to state-of-the-art methods, especially in non asymptotic settings

and with less observations than variables. This setting is at the core of many practical problems, such as genomics and chemometrics.

In Section 5.2, we briefly review PPCA and present several dimensionality selection techniques based on this model. The new normal-gamma prior is presented in Section 5.3 together with a derivation of the closed-form expression of the marginal likelihood. A heuristic to choose hyperparameters is also presented. Numerical experiments are provided in Section 5.4.

5.2 Choosing the intrinsic dimension in probabilistic PCA

Let us assume that a centered independent and identically distributed (i.i.d.) sample $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ is observed that we aim at projecting onto a d-dimensional subspace while retaining as much variance as possible. All the observations are stored in the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$.

5.2.1 Probabilistic PCA

The PPCA model $\mathcal{M}_d$ assumes that, for all $i \in \{1, ..., n\}$, each observation is driven by the following generative model

$$\mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \boldsymbol{\varepsilon}_i, \tag{5.1}$$

where $\mathbf{y}_i \sim \mathcal{N}(0, \mathbf{I}_d)$ is a low-dimensional Gaussian latent vector, $\mathbf{W}$ is a $p \times d$ parameter matrix called the *loading matrix* and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$ is a Gaussian noise term.

This model is an instance of factor analysis and was first introduced by Lawley (1953). Tipping and Bishop (1999) then presented a thorough study of this model. In particular, expanding a result of Theobald (1975), they proved that this generative model is indeed equivalent to PCA in the sense that the principal components of $\mathbf{X}$ can be retrieved using the maximum likelihood (ML) estimator $\mathbf{W}_{\text{ML}}$ of $\mathbf{W}$. More specifically, if $\mathbf{A}$ is the $p \times d$ matrix of ordered principal eigenvectors of $\mathbf{X}^T\mathbf{X}$ and if $\boldsymbol{\Lambda}$ is the $d \times d$ diagonal matrix with corresponding eigenvalues, we have

$$\mathbf{W}_{\text{ML}} = \mathbf{A}(\boldsymbol{\Lambda} - \sigma^2 \mathbf{I}_d)^{1/2}\mathbf{R}, \tag{5.2}$$

where $\mathbf{R}$ is an arbitrary orthogonal matrix.

Under this sound probabilistic framework, dimension selection can be recast as a *model selection problem*, for which standard techniques are available. We review a few important ones in the next subsection.

5.2.2 Model selection for PPCA

The problem of finding an appropriate dimension can be seen as choosing a "best model" within a family of models $(\mathcal{M}_d)_{d \in \{1,...,p-1\}}$. A first popular approach would be to use likelihood penalization, leading to the choice

$$d^* \in \text{argmax}_{d \in \{1,...,p-1\}} \{\log p(\mathbf{X}|\mathbf{W}_{\text{ML}}, \boldsymbol{\sigma}_{\text{ML}}, \mathcal{M}_d) - \text{pen}(d)\},$$

where pen is a penalty which grows with d. These methods include the popular Akaike information criterion (AIC, Akaike, 1974), the Bayesian information criterion (BIC, Schwarz, 1978), or other refined approaches (Bai and Ng, 2002). However, their merits are mainly asymptotic, and our main interest in this chapter is to investigate non-asymptotic scenarios. While the penalty term is usually necessary to avoid selecting the largest model, under a constrained PPCA model, called isotropic PPCA, Bouveyron et al. (2011) proved that regular maximum likelihood was suprinsingly consistent. While the theoretical optimality of this method is also asymptotic, the fact that it directly maximizes a likelihood criterion which is not derived based on asymptotic considerations makes it of particular interest within the scope of this chapter.

Another interesting set of techniques non-asymtotic in essence is Bayesian model selection (Kass and Raftery, 1995). While BIC does not actually approximates the marginal likelihood in the case of PPCA because of violated regularity conditions (Drton and Plummer, 2017), a more refined approach was followed by Minka (2000) who derived a Laplace approximation of the marginal likelihood. This technique, albeit asymptotic, has been proven empirically efficient in several small-sample scenarios.

Another interesting framework considered in the literature is the case where both n and p grow to infinity. Several consistent estimators have been proposed, both from a penalization point of view (Bai and Ng, 2002; Passemier et al., 2017), using Stein's unbiased risk estimator (Ulfarsson and Solo, 2008b) or in a Bayesian context (Hoyle, 2008; Sobczyk et al., 2017). While these high-dimensional scenarios are of growing importance, they fall beyond the scope of this chapter, which is focused on the non-asymptotic setting (with potentially fewer observations than variables), for which very few automatic dimension selection methods are available.

5.3 Exact dimensionality selection for PPCA under a normal-gamma prior

In this section, we present a prior structure that leads to a closed-form expression for the marginal likelihood of PPCA.

5.3.1 The model

We consider the regular PPCA model already defined in (5.1),

$$\forall i \in \{1, ..., n\},\ \mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{y}_i \sim \mathcal{N}(0, \mathbf{I}_d)$, $\mathbf{W}$ is a $p \times d$ parameter matrix, and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$. We rely on a Gaussian prior distribution over the loading matrix $\mathbf{W}$ and a gamma prior distribution over the noise variance σ^2. Specifically, we use a gamma prior $\sigma^2 \sim \text{Gamma}(a, b)$ with hyperparameters $a > 0$ and $b > 0$ together with i.i.d. Gaussian priors $w_{jk} \sim \mathcal{N}(0, \phi^{-1})$ for $j \in \{1, ..., p\}$ and $k \in \{1, ..., d\}$ with some precision hyperparameter $\phi > 0$.

Within the framework of Bayesian model uncertainty (Kass and Raftery, 1995), the posterior probabilities of models can be written as, for all $d \in \{1, ..., p\}$,

$$p(\mathcal{M}_d|\mathbf{X}, a, b, \phi) \propto p(\mathbf{X}|a, b, \phi, \mathcal{M}_d)p(\mathcal{M}_d), \tag{5.3}$$

where

$$p(\mathbf{X}|a, b, \phi, \mathcal{M}_d) = \prod_{i=1}^{n} \int_{\mathbb{R}^{d\times p}\times\mathbb{R}^+} p(\mathbf{x}_i|\mathbf{W}, \sigma, \mathcal{M}_d)p(\mathbf{W}|\phi)p(\sigma|a, b)d\mathbf{W}d\sigma,$$

is the *marginal likelihood* of the data under conditional independence (Kass and Steffey, 1989). Note that this expression also involves model prior probabilities – in this chapter, we will simply consider a uniform prior

$$\forall d \in \{1, ..., p\},\ p(\mathcal{M}_d) \propto 1.$$

Computing the high-dimensional integral of the marginal likelihood usually comes at the price of various approximations (Bishop, 1999a; Minka, 2000; Hoyle, 2008) or expensive sampling (Hoff, 2007). However, with our specific choice of priors, and imposing a constraint on their respective hyperparameters, we obtain a closed-form expression for the marginal likelihood.

Theorem 5.1. *Let $d \in \{1, ..., p\}$. Under the normal-gamma prior with $b = \phi/2$, the log-marginal likelihood of model $\mathcal{M}_d$ is given by*

$$\begin{aligned}\log p(\mathbf{X}|a, \phi, \mathcal{M}_d) &= \sum_{i=1}^{n} \log p(\mathbf{x}_i|a, \phi, \mathcal{M}_d) \\ &= -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(2\phi^{-1}) - n\log\Gamma(a + d/2) \\ &\quad + (a + \frac{d-p}{2})\sum_{i=1}^{n}\log(\frac{\sqrt{\phi}||\mathbf{x}_i||_2}{2}) + \sum_{i=1}^{n}\log K_{a+(d-p)/2}(\sqrt{\phi}||\mathbf{x}_i||_2),\end{aligned} \tag{5.4}$$

where K_ν is the modified Bessel function of the second kind of order $\nu \in \mathbb{R}$.

A detailed proof of this theorem is given in the next subsection.

To the best of our knowledge, this result is the first computation of the marginal likelihood of a PPCA model. It is worth mentioning that, in a slightly different context, Ando (2009) also derived the marginal likelihood of a factor analysis model, with Student factors. Similarly, we derived in the previous chapter derived the exact marginal likelihood of the noiseless PPCA model, in order to obtain sparse PCs.

While Gaussian priors for the loading matrix have been extensively used in the past (Bishop, 1999a; Archambeau and Bach, 2009), it is worth noticing that the use of a gamma prior for a variance parameter is rather peculiar. Indeed, most Bayesian hierarchical models choose *inverse-gamma* priors for variances. This choice is often motivated by its conjugacy properties (see e.g. George and McCulloch, 1993, for a linear regression example or Murphy, 2007, in a wider setting). The derivation provided in the next subsection notably explains why this gamma prior over σ^2 actually arises naturally.

5.3.2 Derivation of the marginal likelihood

We begin by shortly reviewing the generalized Laplace distribution, which will prove to be key within the PPCA framework. This distribution was introduced by Kotz et al. (2001, p. 257). For a more detailed overview, see Kozubowski et al. (2013).

Definition 5.1. *A random variable* $\mathbf{z} \in \mathbb{R}^p$ *is said to have a* ***multivariate generalized asymetric Laplace distribution*** *with parameters* $s > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ *and* $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$ *if its characteristic function is*

$$\forall \mathbf{u} \in \mathbb{R}^p,\ \phi_{\mathrm{GAL}_p(\boldsymbol{\Sigma},\boldsymbol{\mu},s)}(\mathbf{u}) = \left(\frac{1}{1 + \frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u} - i\boldsymbol{\mu}^T\mathbf{u}}\right)^s.$$

When $\boldsymbol{\mu} = 0$, the generalized Laplace distribution is elliptically contoured and is referred to as the *symmetric* generalized Laplace distribution. The elementary properties of the generalized Laplace distribution are discussed by Kozubowski et al. (2013). We list the ones that we consider in the proof of Theorem 1.

Proposition 5.1. *If* $\mathbf{z} \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s)$*, we have* $\mathbb{E}(\mathbf{z}) = s\boldsymbol{\mu}$ *and* $\mathrm{Cov}(\mathbf{z}) = s(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)$*. Moreover, if* $\boldsymbol{\Sigma}$ *is positive definite, the density of* $\mathbf{z}$ *is given by*

$$\forall \mathbf{x} \in \mathbb{R}^p,\ f_{\mathbf{z}}(\mathbf{x}) = \frac{2e^{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}}{(2\pi)^{p/2}\Gamma(s)\sqrt{\det\boldsymbol{\Sigma}}}\left(\frac{Q_{\boldsymbol{\Sigma}}(\mathbf{x})}{C(\boldsymbol{\Sigma},\boldsymbol{\mu})}\right)^{s-p/2} K_{s-p/2}\left(Q_{\boldsymbol{\Sigma}}(\mathbf{x})C(\boldsymbol{\Sigma},\boldsymbol{\mu})\right), \quad (5.5)$$

where $Q_{\boldsymbol{\Sigma}}(\mathbf{x}) = \sqrt{\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}$ *and* $C(\boldsymbol{\Sigma},\boldsymbol{\mu}) = \sqrt{2 + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$*.*

Proposition 5.2. *Let* $s_1, s_2 > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ *and* $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$*. If* $\mathbf{z}_1 \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s_1)$ *and* $\mathbf{z}_2 \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s_2)$ *are independant random variables, then*

$$\mathbf{z}_1 + \mathbf{z}_2 \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s_1 + s_2). \quad (5.6)$$

This proposition is a directed consequence of the expression of the characteristic function of the generalized Laplace distribution.

Another appealing property of the multivariate generalized Laplace distribution is that it can be interpreted as an infinite scale mixture of Gaussians with gamma mixing distribution (a property called *Gauss-Laplace transmutation* by Ding and Blitzstein, 2017).

Proposition 5.3 (Generalized Gauss-Laplace transmutation)**.** *Let* $s > 0$ *and* $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$*. If* $u \sim \mathrm{Gamma}(s, 1)$ *and* $\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ *is independent of* u*, we have*

$$\sqrt{u}\mathbf{x} \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, 0, s). \quad (5.7)$$

For a proof of this result, see Kotz et al. (2001, Chapter 6).

To prove Theorem 5.1, we first study the marginal distribution of the signal term. Following Appendix A, we can state the following lemma.

Lemma 5.1. *Let* $\mathbf{W}$ *be a* $p \times d$ *random matrix with i.i.d. columns following a* $\mathcal{N}(0, \phi^{-1}\mathbf{I}_p)$ *distribution,* $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_d)$ *be a Gaussian vector independent from* $\mathbf{W}$*. We obtain*

$$\mathbf{W}\mathbf{y} \sim \mathrm{GAL}_p(2\phi^{-1}\mathbf{I}_p, 0, d/2). \quad (5.8)$$

Proof. For each $k \in \{1, ..., d\}$ let $\mathbf{w}_k$ be the k-th column of $\mathbf{W}$, $u_k = y_k^2$ and $\boldsymbol{\xi}_k = y_k \mathbf{w}_k$. To prove the lemma, we demonstrate that $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_d$ follow a GAL distribution and use the decomposition

$$\mathbf{W}\mathbf{y} = \sum_{k=1}^{d} \boldsymbol{\xi}_k.$$

Let $k \in \{1, ..., d\}$. Since $\mathbf{y}$ is standard Gaussian, $u_k = y_k^2$ follows a $\chi^2(1)$ distribution, or equivalently a Gamma$(1/2, 1/2)$ distribution. Therefore, $u_k/2 \sim$ Gamma$(1/2, 1)$. Moreover, note that $\sqrt{u_k}\mathbf{w}_k = |y_k|\mathbf{w}_k = y_k \text{sign}(y_k)\mathbf{w}_k \overset{d}{=} y_k \mathbf{w}_k$ since $|y_k|$ and sign(y_k) are independent and sign$(y_k)\mathbf{w}_k \overset{d}{=} \mathbf{w}_k$. Therefore, according to Proposition 5.3, we have

$$\boldsymbol{\xi}_k = \sqrt{\frac{u_k}{2}}\sqrt{2}\mathbf{w}_k \sim \text{GAL}_p(2\phi^{-1}\mathbf{I}_p, 0, 1/2).$$

Since $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_d$ are i.i.d. and following a $\text{GAL}_p(2\phi^{-1}\mathbf{I}_p, 0, 1/2)$ distribution, we can use Proposition 5.2 to conclude that

$$\mathbf{W}\mathbf{y} = \sum_{k=1}^{d} \boldsymbol{\xi}_k \sim \text{GAL}_p(2\phi^{-1}\mathbf{I}_p, 0, d/2).$$

□

We now focus on the second term of (5.1) involving the noise vector.

Lemma 5.2. *Let* $\boldsymbol{\varepsilon}_i|\sigma^2 \sim \mathcal{N}(0, \sigma^2\mathbf{I}_p)$ *and* $\sigma^2 \sim \text{Gamma}(a, b)$ *then*

$$\boldsymbol{\varepsilon}_i \sim \text{GAL}_p\left(b^{-1}\boldsymbol{I}_p, 0, a\right).$$

Proof. Again, the Gauss-Laplace transmutation is considered. Indeed, the noise can be written as

$$\boldsymbol{\varepsilon}_i = \sqrt{b\sigma^2}\mathbf{e}_i,$$

where $\mathbf{e}_i \sim \mathcal{N}(0, b^{-1}\mathbf{I}_p)$. Therefore, the Gauss-Laplace transmutation allows to conclude. □

Now that we have proved that both the signal and the noise term follow marginally a generalized Laplace distribution, we use Proposition 5.2 which ensures that, assuming $b = \phi/2$, the sum of the two generalized Laplace random vectors is a generalized Laplace random vector:

$$\mathbf{x}_i \sim \text{GAL}_p(2\phi^{-1}\mathbf{I}_p, 0, a + d/2). \tag{5.9}$$

Using the expression of the density of the generalized Laplace distribution, we eventually end up with the closed-form expression of the marginal likelihood of Theorem 1.

5.3.3 Choosing hyperparameters

To obtain a closed-form expression of the marginal likelihood, we have shown that it is sufficient to assume that $b = \phi/2$. Two hyperparameters remain henceforth to be tuned: the shape parameter of the gamma prior a and the precision hyperparameter ϕ. We developed data-driven heuristics for this purpose.

A first observation is that, when d grows, σ is expected to decay because the signal part of the model can be more expressive. This prior information can be distilled into the model by roughly centering the gamma priors on estimates of $\hat{\sigma}$. More precisely, our heuristic is to choose a such that $\mathbb{E}(\sigma) \propto \hat{\sigma}$ for each d. In order for ϕ to control the diffusiveness of both the loading matrix and the variance, we specifically made the choice $a = \hat{\sigma}^2/\phi$. In our experiments, we chose the ML estimator $\hat{\sigma} = \sigma_{\mathrm{ML}}$ (which is the mean of the $p - d$ smallest eigenvalues of the covariance matrix, see Tipping and Bishop, 1999) but more complex estimates may be considered (Passemier et al., 2017).

Regarding the remaining parameter ϕ, we propose a heuristic based on the following statements which can be made regarding the problem of dimension selection:

- overestimation of d should be preferred to underestimation since loosing some information is much more damageable than having a representation not parsimonious enough,
- consequently, the marginal likelihood curve as a function of the dimension should have two distinct phases: a first one when "signal dimensions" are added (before the true value of d), and a second one, when "noise dimensions" are added.

Thus, we built a simple heuristic criterion to judge the relevance of a choice of ϕ by the shape of the marginal likelihood curve. First, if the slope of the first part of the curve (before the maximum) is lower than the slope of the second part, this means that this choice leads to underestimation and is therefore discarded. Second, the criterion is equal to the discrete second derivative of the marginal likelihood curve evaluated at the maximum, in order to select a hyperparameter leading to a strong distinction between the two phases. This criterion is eventually maximized over a grid of values of ϕ. This scheme for hyperparameter choice is illustrated in Fig. 5.1 using the simpler simulation scheme described in Subsection 5.4.2.

5.4 Numerical experiments

In this section, we perform some numerical experiments in order to highlight the main features of the proposed approach and to compare it with state-of-the-art methods.

5.4.1 Simulation scheme

To assess the performance of our algorithm (referred hereafter as NGPPCA or NG, for short), we consider the following simulation scheme in the following experiments. We follow the simulation setup proposed in Bouveyron et al. (2011) based on their isotropic PPCA model. We therefore simulate data sets following the isotropic PPCA model which assumes

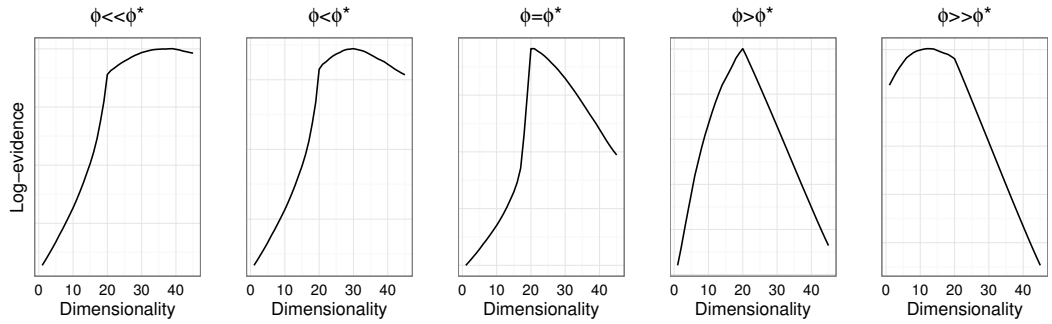


Figure 5.1 – Different shapes of the marginal likelihood curve for growing values of ϕ. ϕ^* corresponds to the maximum of the heuristic criterion that we describe is Subsection 5.3.3. The true dimensionality is 20.

that the covariance matrix of X has only two different eigenvalues a and b (instead of $d+1$ in the PPCA model). In this case, the signal-to-noise ratio (SNR hereafter) is simply defined by

$$\text{SNR} = \frac{ad}{p-d}.$$

In our simulation, b is set up to 1 and $a > 1$, which will control the strength of the signa, varies to explore different signal-to-noise ratios. Then, an orthonormal $p \times p$ matrix $\mathbf{Q}$ is drawn uniformly at random. The data is eventually generated according to a centered Gaussian distribution with covariance matrix

$$\mathbf{Q}^T \text{diag}(\overbrace{a, ..., a}^{d \text{ times}}, \overbrace{1, ..., 1}^{p-d \text{ times}})\mathbf{Q}.$$

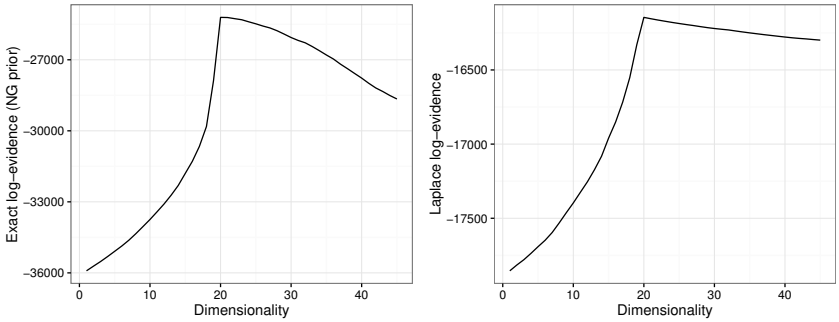
Finally, the number p of variables is fixed to 50 in all experiments and the number n of observations varies in the range $\{40, 50, 70, 100\}$.

5.4.2 Introductory examples

We first conduct two small simulations to illustrate the behavior of our algorithm and its difference with the Laplace approximation of Minka (2000). We consider two scenario: a simple case and a harder and more realistic one.

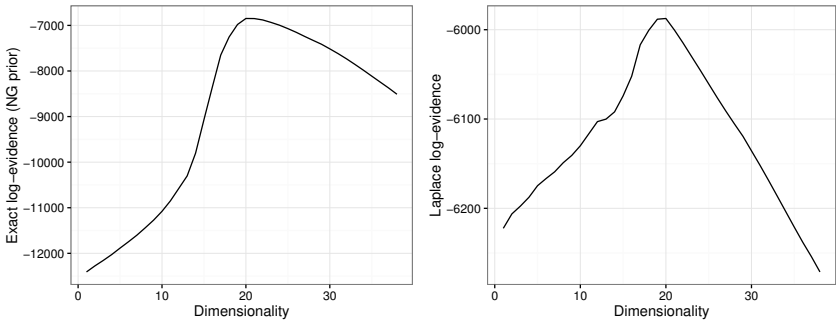
SIMPLE SCENARIO We consider a setup with $n = 100$ and SNR $= 20$. In this simple scenario, we first illustrate our heuristic for hyperparameter tuning by displaying marginal likelihood curves for different values of ϕ (Fig. 5.1). The heuristic criterion allows to find the desired shape, leading to a correct dimensionality estimation. A GIF animation displaying all values of the criterion for a large grid of 200 values of ϕ is provided as a online material[1]. On Fig. 5.2, we compare the results of our algorithm with the Laplace approximation of the marginal likelihood of Minka (2000). In this case, both methods recover the true dimensionality of the

[1] http://pamattei.github.io/animationeasy.gif



NG (our methodology) Laplace (Minka, 2000)

Figure 5.2 – Exact log-evidence for NGPPCA (left) and the Laplace approximation of Minka (2000) (right) for the simpler simulation scenario ($n = 100$). Both curves have the desirable properties detailed is Subsection 5.3.3 and find the correct dimensionality $d = 20$.



NG (our methodology) Laplace (Minka, 2000)

Figure 5.3 – Exact log-evidence for NGPPCA (left) and the Laplace approximation of Minka (2000) (right) for the more challenging simulation scenario ($n = 40$). While both methods select the correct dimensionality $d = 20$, the Laplace approximation prefers underestimation which is not a satisfactory behavior. Our exact computation gives a much more acceptable curve.

data and are very confident with their choice (the posterior probability of the true model is higher than 99% with both approaches). The two curves have a similar shape, in compliance with the expected shape, as detailed in Subsection 5.3.3.

CHALLENGING SCENARIO We now consider a setup with $n = 40$ and SNR $= 20$. A GIF animation illustrating hyperparameter tuning is provided online[2]. Again, our results are compared with the Laplace approximation (Fig. 5.3). Regarding our exact approach (left panel), the marginal likelihood curve has an extremely similar shape to the one of the first simulation. This shape is satisfactory, and the maximum of our heuristic criterion actually corresponds to the true dimensionality. Although it also finds the correct dimensionality, the Laplace approximation wrongfully prefers simpler models. More precisely, the top two models chosen by the Laplace approximation are $\mathcal{M}_{20}$ (with posterior probability 69.3%) and $\mathcal{M}_{19}$ (with posterior probability 30.7%). In contrast, our algorithm favors $\mathcal{M}_{20}$ (with posterior probability 86.7%) and $\mathcal{M}_{21}$ (with posterior probability 13.3%). By prefering overestimation over underestimation, the exact method appears less likely to destroy valuable information, which would be damaging in a dimensionality selection context.

As a summary, those experiments confirm the expected behaviors of NG *vs.* Laplace approximation: in the first scenario ($n = 100$), the asymptotic assumption of the Laplace approximation is much more relevant than in the second setup ($n = 40$). Our method, which does not rely on such an assumption, is much less impacted by the reduction of the sample size.

5.4.3 Benchmark comparison with other dimension selection methods

This section now focuses on the comparison of our methodology with other dimension selection methods. We here consider all possible scenarios with $n \in \{40, 50, 70, 100\}$ and a SNR grid going from 1.5 to 30 (50 repetitions are made for each case). We compare the performance of our technique based on the normal-gamma prior (NG) with the following four competitors:

- the Laplace approximation of Minka (2000) which is a benchmark Bayesian method for dimensionality selection,
- the generalized cross-validation approximation (GCV) of Josse and Husson (2012) which is known to give state of the art results in many scenarios (see the vast simulation study of Josse and Husson, 2012)
- the high-dimensional Laplace approximation of Sobczyk et al. (2017) called PESEL, which performs well even in scenarios that imply a large number of variables
- the ML approach of Bouveyron et al. (2011), which maximizes a non-asymptotic criterion (the likelihood). Notice that is specifically adapted to this simulation scheme and this advantage allow us to consider this technique a gold-standard for the simulated data.

[2] http://pamattei.github.io/animationhard.gif

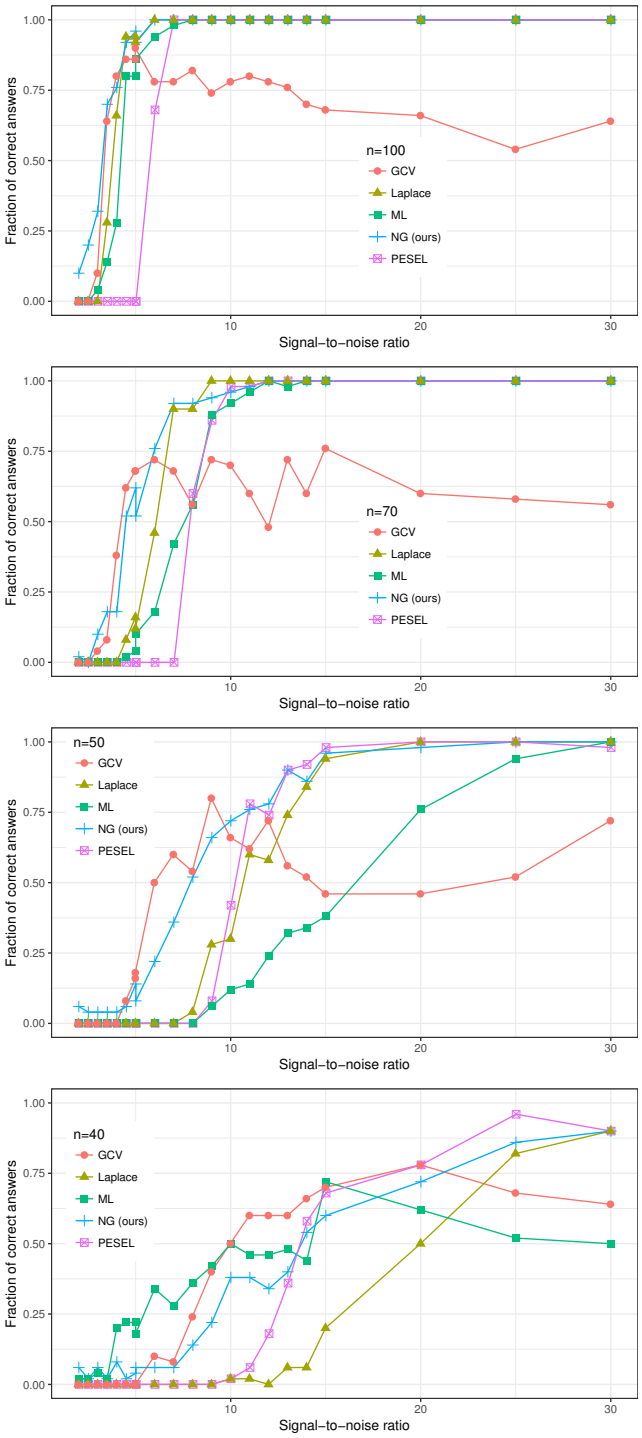


Figure 5.4 – Percentage of correctly estimated dimensions for different sample sizes (50 replications) of NGPPCA (NG) and its competitors for different signal-to-noise ratios. From top to bottom, the data sample sizes are respectively $n = 100,\ 70,\ 50$ and 40.

The performance metric that we chose is the percentage of correct answers given by each algorithm, which is a standard measure used in other simulations studies (see e.g. Minka, 2000; Hoyle, 2008; Ulfarsson and Solo, 2008b). Results are presented in Fig. 5.4.

One can first notice generalized cross-validation often gives satisfactory results, but fails to be competitive with model-based methods when the SNR is high. This is arguably a consequence of the fact that the data is actually generated according to a PPCA model. The ML approach has a good behaviour, especially when n is very small, this is partly explained by the fact that it is designed for this very simulation setup. As expected, PESEL especially outperforms the traditional Laplace approximation when p/n is large. Finally, our approach (NG), which consistently outperforms the other Bayesian method (the traditional Laplace approximation and PESEL), is the only method that gives satisfactory results in all settings (high and low SNR, moderate and small n).

5.5 Conclusion

PCA is more of a descriptive and exploratory tool than a model. Therefore, no unique dimension selection method should be uniquely preferred – sometimes, very relevant information may actually lie within the *last* PCs (Jolliffe, 2002, Section 3.4). However, PCA's ubiquity in the statistical world makes necessary the search for guidance procedures to help the practitioner choose the number of PCs. This need is even more critical when the data are scarce or particularly expensive. Our work, by deviating from usually adopted asymptotic settings, is a step in that direction. Regarding future work, our exact computation of model posterior probabilities may be used to perform Bayesian model averaging (Hoeting et al., 1999) in predictive contexts. Potential applications could involve principal component regression (Jolliffe, 2002, Chapter 8), image denoising (Deledalle et al., 2011), or deep learning (Chan et al., 2015). As a concluding note, these last two chapters come as an illustration that exactly computing the marginal likelihood is sometimes easier than expected. Although both recent asymptotic approximations (Drton and Plummer, 2017) and the MCMC arsenal (Friel and Wyse, 2012) are well-equipped to deal with marginal likelihoods, we argue, like Lin et al. (2009), that finding exact expressions is an important task that should not be deemed untractable too hastily.

6

Conclusion, Ongoing Work, and Perspectives

For twenty years, high-dimensional data has kept on providing countless challenges for our field, calling upon a constant renewal of statistical theory and practice. The aim of this thesis was to illustrate how the Bayesian framework of model uncertainty can provide a scalable way of dealing with high-dimensional data. In this conclusion, we briefly recall the main contributions of this thesis and give details about several ongoing projects. More distant perspectives are also eventually evoked.

6.1 Overview of the contributions

The first two chapters of this thesis reviewed the challenge this work attempts at tackling – high-dimensional machine learning – and the main tool used for that purpose – Bayesian model uncertainty. We proposed algorithms for high-dimensional linear regression (SpinyReg, Chapter 3) and principal component analysis (GSPPCA, Chapter 4). To this end, we introduced a new continuous relaxation of the traditional Bayesian model selection

problem. Both methods proved to be scalable and competitive with state-of-the-art approaches, and allowed to find relevant sets of variables hidden in real high-dimensional data, coming notably from social transportation or genomics. In particular, these techniques led to greater interpretability of several data sets involving much more variables than observations. From a theoretical perspective, we derived the first closed-form expression of the marginal likelihood of a principal component analysis (PCA) model. Besides being at the heart of the GSPPCA algorithm, this result allowed us to build an algorithm to estimate the intrinsic dimension of a high-dimensional data set (Chapter 5). To do so, we introduced a new normal-gamma (NG) prior structure for PCA, which led to a closed-form computation of the marginal likelihood, and to an exact assessment of model uncertainty. R code for both SpinyReg and GSPPCA is available online (via the `spinreg` package on CRAN and from *https://github.com/pamattei/GSPPCA*).

This work led to the production of several scientific articles. Among them, two papers and a discussion were published in international peer-reviewed journals:

- **Discussion on the Paper "A Bayesian Information Criterion for Singular Models" by Drton and Plummer**, *Journal of the Royal Statistical Society: Series B*, vol. 79, pp. 370–371 (2017)
- **Multiplying a Gaussian Matrix by a Gaussian Vector**, *Statistics & Probability Letters*, vol. 128, pp. 67–70 (2017)
- **Combining a Relaxed EM Algorithm with Occam's Razor for Bayesian Variable Selection in High-Dimensional Regression** (with Charles Bouveyron, Julien Chiquet, and Pierre Latouche), *Journal of Multivariate Analysis*, vol. 146, pp. 177–190 (2016)

One article was published in the proceedings of an international peer-reviewed conference:

- **Globally Sparse Probabilistic PCA** (with Charles Bouveyron and Pierre Latouche), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, vol. 51, pp. 976–984 (2016)

And two preprints have been submitted to journals:

- **Exact Dimensionality Selection for Bayesian PCA** (with Charles Bouveyron and Pierre Latouche), Preprint HAL-01484099, Université Paris Descartes (2017).
- **Bayesian Variable Selection for Globally Sparse Probabilistic PCA** (with Charles Bouveyron and Pierre Latouche), Preprint HAL-01310409, Université Paris Descartes (2016).

6.2 Work in progress

We describe several ongoing projects related to high-dimensional classification or clustering.

6.2.1 Mixtures of Globally Sparse Probabilistic PCA

In classification problems, it is of paramount importance to be able to interpret the distribution of the classes. Class-wise variable selection provides a useful mean of class interpretation, by stating that a specific set of variables describes a given class best. This *explicative* perspective differs from the more common *discriminative* variable selection approach that looks for variables that offer the greatest discriminative power between classes. GSPPCA, which provides a scalable way of selecting relevant variables in high-dimensional contexts, can be applied to class-wise variable selection for both supervised and unsupervised classification. We give here a few examples of ongoing work in that direction.

6.2.1.a High-dimensional discriminant analysis

In discriminant analysis, it is assumed that the classes follow simple distributions, such as the Gaussian distribution (Fraley and Raftery, 2002). In high-dimensional settings, Bouveyron et al. (2007a) suggested, roughtly speaking, to use the PPCA model for each class. Using GSPPCA instead leads to an efficient way of performing class-wise variable selection. We implemented this framework motivated by a request from Dr. Gildas Bertho's team (Plateforme RMN, Université Paris Descartes). Dr. Bertho's team is interested in chronic kidney disease (CKD) diagnosis using nuclear magnetic resonance (NMR). Specifically, we considered urine samples coming from an anonymized cohort of 110 patients who were referred to the Nephrology department of Hôpital Européen Georges Pompidou in Paris for a kidney biopsy between 2013 and 2014 (see Luck et al., 2016, for more details). The four different classes correspond to different levels of CKD severity, and are defined according to creatinine rates. Beyond automated diagnosis, the goal would be to isolate some urinary metabolites as early-stage markers of the disease.

Our approach was to fit a GSPPCA model for each class, and then use the results to derive a decision rule, as in Bouveyron et al. (2007a). This resulted in finding relevant parts of NMR spectra specific to each class, as illustrated in Figure 6.1.

6.2.1.b High-dimensional clustering

In cluster analysis, the PPCA model has also been a popular way of modeling classes (Tipping and Bishop, 1999; Bouveyron et al., 2007b). Again, using GSPPCA instead would lead to an increased interpretability of the classes. In such unsupervised scenarios, this way of interpreting the found classes appears even more important, as there is often no ground truth to assess the relevance of the clustering. As an early approach, we implemented an inference strategy for a mixture of GSPPCAs model (MGSPPCA) using the classification expectation-maximization algorithm of Celeux and Govaert (1992) together with our relaxed variational model. We illustrate the behavior of this prototype using the *mnist-back-rand* data set described in Chapter 4. We built a two-class data set using 500 handwritten sevens and 500 handwritten threes. We showed the clustering performance of several approaches in Table 6.1. MGSPPCA produces the closest partition to the true one. More importantly, it leads to a better understanding of the results. Indeed, while k-means exhibits good cluster-

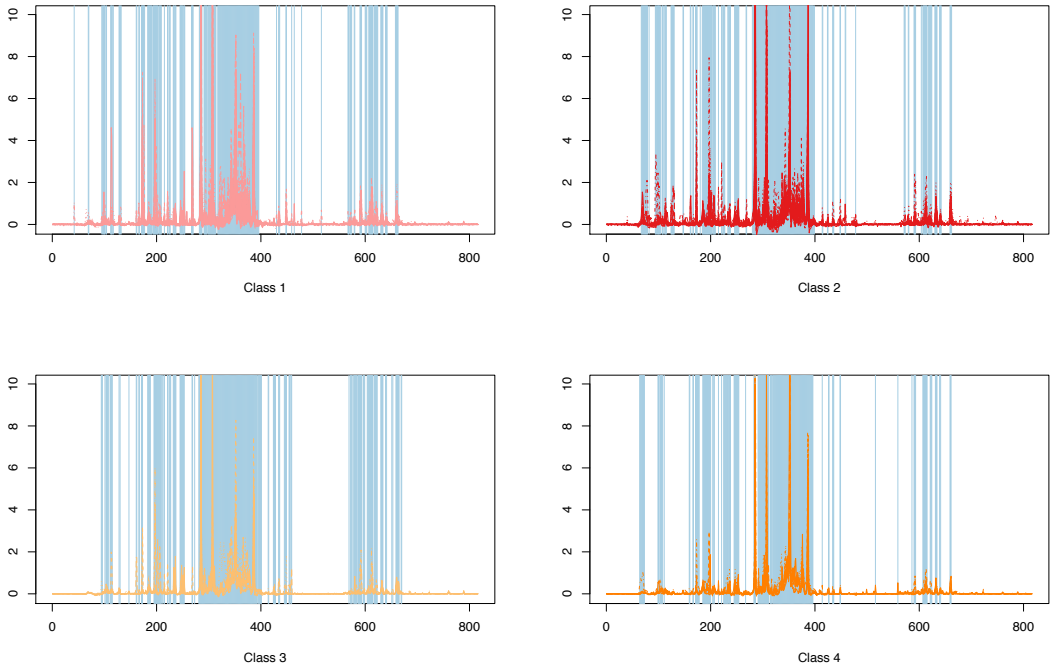


Figure 6.1 – Variables selected by applying GSPPCA to the four classes of the CKD data set. Although a large part of the spectrum is common to all four classes, some variables are class-specific, and could lead to finding relevant urinary metabolites.

	k-means	sparse k-means	MICL	MGSPPCA
ARI	80.3	75.0	77.4	**84.6**

Table 6.1 – Clustering performance – measured using the adjusted rand index (ARI, Hubert and Arabie, 1985) – of several clustering methods for the 500 threes and 500 sevens from the *mnist-back-rand* data set: k-means and its sparse version (Witten and Tibshirani, 2010), the maximum integrated complete likelihood (MICL) approach of Marbac and Sedki (2017), and mixtures of GSPPCAs (MGSPPCA).

ing performance for this data set, the cluster means (displayed on Figure 6.2) appear difficult to interpret. In contrast, displaying the sparsity patterns of the two GSPPCA components allows to interpret the two classes by clearly recognizing a three and a seven (see Figure 6.2).

Several technical points of the MGSPPCA algorithm remain to be chosen. In particular, an efficient model selection technique that allows to select both the number of clusters and their sparsity levels has to be found. Using the exact marginal likelihood expression of GSPPCA, we could derive an exact integrated complete likelihood (ICL) criterion using a technique similar to the one of Bertoletti et al. (2015).

6.2.2 Variable screening for high-dimensional multiclass discriminant analysis

6.2.2.a Improving the scalability of variable selection for Gaussian discriminant analysis

In recent years, several greedy algorithms have been proposed for variable selection in Gaussian discriminant analysis. While these algorithms exhibit extremely good performances, their scalability to high-dimensional data sets is limited by their greedy nature (Murphy et al., 2010; Maugis et al., 2011). During a two-month research visit to the Insight Center of University College Dublin funded by the Fondation Sciences Mathématiques de Paris, Pr. Bouveyron and I collaborated with Pr. Brendan Murphy and his Ph.D. student Michael Fop to enhance the scalability of these techniques. In particular, we were motivated by a food authenticity data set that involves near infrared spectra coming from five different meats: beef, chicken, lamb, pork and turkey – this data set was briefly introduced in Chapter 1, see also McElhinney et al. (1999). For this challenging data set, some classes are much more difficult to discriminate than others (for example, discriminating between chicken and turkey is much harder than between chicken and beef), and classical machine learning algorithms such as support vector machines (SVMs, Cortes and Vapnik, 1995) or random forests (Breiman, 2001) are vastly outperformed by Gaussian discriminant analysis with greedy variable selection (Murphy et al., 2010). However, discriminant analysis comes at a much more expensive computational price that we would try to reduce.

6.2.2.b Screening variables with Bayes factors and class partitions

We also developed a screening technique that allows to perform a crude variable preselection in order to limit the greedy search to this smaller preselected subset of variables. Following the seminal work of Fan and Lv (2008), marginal screening methods for supervised learning

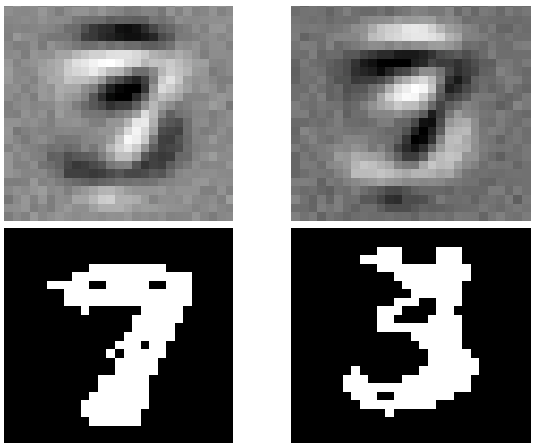


Figure 6.2 – *Top row:* Centroids of the two clusters found by k-means. It appears difficult to clearly interpret the classes. *Bottom row:* Variables selected by MGSPPCA for the two found clusters. The two digits (3 and 7) are clearly recognizable and interpretation is easier.

have been growingly popular. The key idea is to rank variables using scores that are solely based on marginal distributions. However, in multiclass classification, using a single ranking is likely to focus only on the easiest binary classification problem (white meat versus red meat in the food authenticity data set). We therefore decided to build one ranking for each possible partition of the classes. We first remind the definition of a partition of a set:

Definition 6.1. *Let $\mathcal{A}$ be a set of cardinal $n \in \mathbb{N}$. A partition of $\mathcal{A}$ is a set of nonempty subsets of $\mathcal{A}$ such that every element of $\mathcal{A}$ is exactly in one of these subsets. The number of partitions of $\mathcal{A}$ is called the nth Bell number and is denoted B_n. The partition $\{\mathcal{A}\}$, which is the only partition of cardinal 1, is called the trivial partition.*

Bell numbers can be computed recursively using a simple recurrence relation (see e.g. Rota, 1964). The first Bell numbers are

$$B_0 = B_1 = 1,\ B_2 = 2,\ B_3 = 5,\ B_4 = 15,\ B_5 = 52,\ B_6 = 203. \tag{6.1}$$

Let $\mathcal{C}$ the set of all C possibles classes. We will build *one ranking of variables for each nontrivial partition of* $\mathcal{C}$ using Bayesian model uncertainty.

Given a nontrivial partition $\rho = \{\rho_1, ..., \rho_K\}$ of cardinal $K \in \{2, ..., C\}$ and a variable $j \in \{1, ..., p\}$, we wish to measure the usefulness of variable j to discriminate the classes induced by ρ. To this end, we build on the Bayes factors framework, which allows to compute the statistical evidence in favor of a Bayesian model (Chapter 2). We denote by $\mathcal{M}_\rho^j$ the model where the marginal distribution of variable j is a mixture of K Gaussians (each Gaussian component corresponds to one of the classes induced by ρ). Given some parameters $\boldsymbol{\tau} \in \Delta^C$, $\mu_1, ..., \mu_K \in \mathbb{R}$ and $\sigma_1, ..., \sigma_K \in \mathbb{R}^+$, this is written

$$\mathcal{M}_\rho^j : \left\{ \begin{array}{l} z \sim \text{Cat}(\boldsymbol{\tau}) \\ x_j | \{z \in \rho_k\} \sim \mathcal{N}(\mu_k, \sigma_k). \end{array} \right. \tag{6.2}$$

Model $\mathcal{M}_\rho^j$ assumes that variable j is relevant to discriminate the classes induced by ρ. To assess the usefulness of variable j, this model will be compared to a model named $\mathcal{M}_0^j$ which assumes that variable j is marginally Gaussian (and therefore poorly discriminative). Specifically, for $\boldsymbol{\tau} \in \Delta^C$, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$, we define

$$\mathcal{M}_0^j : \begin{cases} z \sim \text{Cat}(\boldsymbol{\tau}) \\ x_j \sim \mathcal{N}(\mu, \sigma). \end{cases} \tag{6.3}$$

Some prior distributions $p(.|\mathcal{M}_0^j)$ and $p(.|\mathcal{M}_\rho^j)$ are also chosen. For mathematical convenience, we choose conjugate normal-inverse-gamma (NIG) priors (Murphy, 2007). Indeed, this choice will lead to a closed-form expression of our Bayes factors-based score. Specifically, we consider

$$p(\boldsymbol{\tau}, \mu_1, ..., \mu_K, \sigma_1, ..., \sigma_K | \mathcal{M}_\rho^j) = p(\boldsymbol{\tau}|\mathcal{M}_\rho^j) \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_{\rho,k}^j, \sigma_k V_{\rho,k}^j) \text{IG}(\sigma_k^2 | a_{\rho,k}^j, b_{\rho,k}^j), \tag{6.4}$$

and

$$p(\boldsymbol{\tau}, \mu, \sigma | \mathcal{M}_0^j) = p(\boldsymbol{\tau}|\mathcal{M}_0^j) \mathcal{N}(\mu | m^j, \sigma V^j) \text{IG}(\sigma^2 | a^j, b^j), \tag{6.5}$$

given some hyperparameters that we specified using the unit information prior rationale (Kass and Wasserman, 1995; Raftery, 1995). Note that we did not specify the prior for the class proportions parameter $\boldsymbol{\tau}$. Indeed, as long as we reasonably assume that $p(\boldsymbol{\tau}|\mathcal{M}_0^j) = p(\boldsymbol{\tau}|\mathcal{M}_\rho^j)$, this prior will have no effect on the Bayes factors-based score that we will use.

To measure the usefulness of variable j, we consider the score

$$\log \text{BF}_{\mathcal{M}_\rho^j / \mathcal{M}_0^j} = \log p(\mathbf{x}_j, \mathbf{z} | \mathcal{M}_\rho^j) - \log p(\mathbf{x}_j, \mathbf{z} | \mathcal{M}_0^j), \tag{6.6}$$

which is exactly the weight of evidence in favor of $\mathcal{M}_\rho^j$ (see Chapter 2). Using our specific prior structure, this score can be computed exactly, leading to the fast computation of all variable rankings. Once all rankings are established, the screening algorithm works as follows:

1. The practitioner chooses a maximum number of variables $q_{\max}$ to be retained by the screening algorithm.

2. The top-$k(q_{\max})$ variables in each partition-specific ranking are retained, where $k(q_{\max})$ is the largest integer such that keeping the top-$k(q_{\max})$ variables in each ranking eventually leads to selecting no more than $q_{\max}$ variables.

6.2.3 Deep adversarial clustering

This thesis was mainly concerned with model-based approaches to high-dimensional learning. These approaches assume that the data comes from a specific set of parametric models. The fact that this assumption is essentially false in practice should not be a foundational concern: recall Box's famous quote from Chapter 2

> *There is no need to ask the question "Is the model true?"*

Nevertheless, we would want our models to be as true as possible – and, in a sense, this is what model uncertainty is all about. Together with his Ph.D. student Warith Harchaoui (Université Paris Descartes & Oscaro.com), Pr. Bouveyron and I used a increasingly popular deep learning technique called *adversarial training* to perform Gaussian model-based clustering in high-dimensional settings, with some guarantees about the relevance of the Gaussian assumption.

6.2.3.a Nonlinear dimensionality reduction for model-based clustering

Directly performing model-based clustering in high-dimensional spaces is challenging, because of the peculiar nature of high-dimensional probability distributions (Chapter 1). A way of tackling this problem is to reduce the dimensionality of the data and perform the clustering in a smaller, more well-behaved space. This kind of search for a suitable transformation of the data that would simplify a statistical problem is called *representation learning*. The importance of finding a suitable representation for clustering was first highlighted by Chang (1983), who showed that embeddings based on principal component analysis were often unfit for clustering purposes. Accordingly, several proposals of clustering-aware representations have been proposed. In the context of linear embeddings, the main approach was to combine linear discriminant analysis with the k-means algorithm (De la Torre and Kanade, 2006) or more generally with Gaussian mixtures (Bouveyron and Brunet, 2012). We propose to perform non-linear dimensionality reduction using an autoencoder, which can be seen as a nonlinear generalization of PCA (Goodfellow et al., 2016, Chapter 14). Specifically, denoting by $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ the data, an autoencoder minimizes the quantity

$$\mathcal{L}_{\text{AE}}(\boldsymbol{\theta}_{\mathcal{E}}, \boldsymbol{\theta}_{\mathcal{D}}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathcal{D}_{\boldsymbol{\theta}_{\mathcal{D}}}(\mathcal{E}_{\boldsymbol{\theta}_{\mathcal{E}}}(\mathbf{x}_i))\|_2^2,$$

where $\mathcal{E}_{\boldsymbol{\theta}_{\mathcal{E}}}$ and $\mathcal{D}_{\boldsymbol{\theta}_{\mathcal{D}}}$ are functions parametrized as deep neural nets called the *encoder* (indexed by $\boldsymbol{\theta}_{\mathcal{E}}$) and the *decoder* (indexed by $\boldsymbol{\theta}_{\mathcal{D}}$). The encoder maps the data to a low-dimensional subspace $\mathbb{R}^d$ (with $d \ll p$), and the decoder allows to build data using these low-dimensional coordinates. If both functions are linear (using neural nets terminology, this would be called a shallow autoencoder with linear activations), then the autoencoder exactly reduces to PCA. By contrast, using deep neural networks for both functions (usually together with good regularization schemes) allows to learn powerful non-linear embeddings of the data. However, there are no guarantees that such embeddings may be suitable for clustering. To provide such guarantees, we propose to regularize the traditional autoencoder objective using adversarial training (Goodfellow et al., 2014). The key idea is to make sure that the distribution of the low-dimensional coordinates $\mathcal{E}_{\boldsymbol{\theta}_{\mathcal{E}}}(\mathbf{x}_1), ..., \mathcal{E}_{\boldsymbol{\theta}_{\mathcal{E}}}(\mathbf{x}_n) \in \mathbb{R}^d$ is close to a true Gaussian mixture model (GMM). To this end, we will use the fact that, given a value of $\boldsymbol{\theta}_{\mathcal{E}}$ and the parameters of a Gaussian mixture with K components $\boldsymbol{\theta}_{\mathcal{M}} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K)$, it is easy to simulate from both the Gaussian mixture and the low-dimensional coordinates. Adversarial training provides a computational framework for minimizing distances between two distributions that one can merely simulate from. The adversarial rationale involves training a neural net classifier called the discriminator $\mathcal{A}_{\boldsymbol{\theta}_{\mathcal{A}}}$ (parametrized by $\boldsymbol{\theta}_{\mathcal{A}}$) to discriminate between samples from the two distributions: when the

	k-means	GMM	AE+GMM	DEC	DAC
Clustering accuracy	53.47	53.73	82.56	84.30	**96.50**

Table 6.2 – Clustering accuracy results (%, the higher, the better) for the MNIST data set of handwritten digits. DEC corresponds to another clustering based on deep neural nets (Xie et al., 2016). Both DAC, AE+GMM and DEC use the same deep architecture – in particular, the dimension of the low-dimensional embedding is 10.

discriminator is unable to discern the differences between the two kinds of samples, it means that the distributions are close. Specifically, as in Goodfellow et al. (2014), we consider the minimax optimization problem $\min_{\theta_\mathcal{A}} \max_{\theta_\mathcal{E}}$:

$$\mathcal{L}_\text{A}(\boldsymbol{\theta}_\mathcal{E}, \boldsymbol{\theta}_\mathcal{M}) = -\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_\text{data}}[\log \mathcal{A}_{\boldsymbol{\theta}_\mathcal{A}}(\mathcal{E}(\mathbf{x}))] - \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim \text{Mixt}(\boldsymbol{\theta}_\mathcal{M})}[\log(1 - \mathcal{A}_{\boldsymbol{\theta}_\mathcal{A}}(\mathbf{z}))], \tag{6.7}$$

where p_data is the data generating distribution and $\text{Mixt}(\boldsymbol{\theta}_\mathcal{M})$ the Gaussian mixture with parameters $\boldsymbol{\theta}_\mathcal{M}$. Solving this minimax problem can be seen as way to minimize the Jensen-Shannon divergence between the Gaussian mixture and the distribution of $\mathcal{E}_{\boldsymbol{\theta}_\mathcal{E}}(\mathbf{x}_1), ..., \mathcal{E}_{\boldsymbol{\theta}_\mathcal{E}}(\mathbf{x}_n)$. Inspired by Makhzani et al. (2015), we alternatively optimize $\mathcal{L}_\text{AE}$ and $\mathcal{L}_\text{A}$, using stochastic gradient descent. This objective insures that the learnt representation approximatively follows a Gaussian mixture and is able to reconstruct the data well. We call this clustering algorithm DAC (*deep adversarial clustering*).

Let us give an example using the MNIST data set which contains 70.000 images of handwritten digits with ten classes (from 0 to 9). Fitting a traditional Gaussian mixture or using k-means leads to wrongly classifying half the data set. However, as shown on Table 6.2, DAC achieves a clustering accuracy of 96.5%. Simply using a regular deep autoencoder followed by fitting a Gaussian mixture with the low-dimensional embedding leads to an accuracy of 82.6%. This means that, albeit valuable, the embedding learnt by a regular deep autoencoder is much less suitable for clustering than the one learnt with our adversarially regularized autoencoder. To assess visually the effectiveness of this regularization, we display in Figure 6.3 a 3-dimensional PCA representation of the 10-dimensional embedding learnt by DAC. Beyond its very good clustering results, a promising feature of DAC is that it is able to generate new (fake) data from the inferred classes. Indeed, for a given class k, we can generate low-dimensional coordinates $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and pass them through the decoder to create some a fake observation $\mathcal{D}_{\boldsymbol{\theta}_\mathcal{D}}(\mathbf{z})$ from class k. Such visualizations are displayed on Figure 6.4, and allow to correctly interpret the clusters discovered by DAC.

6.3 Perspectives

Eventually, we outline several perspectives about extensions of the work described in this thesis.

6.3.1 Generalized linear models and model averaging

A natural extension of the SpinyReg algorithm would be to go beyond the linear regression model. Several tractability issues associated with Bayesian generalized linear models would

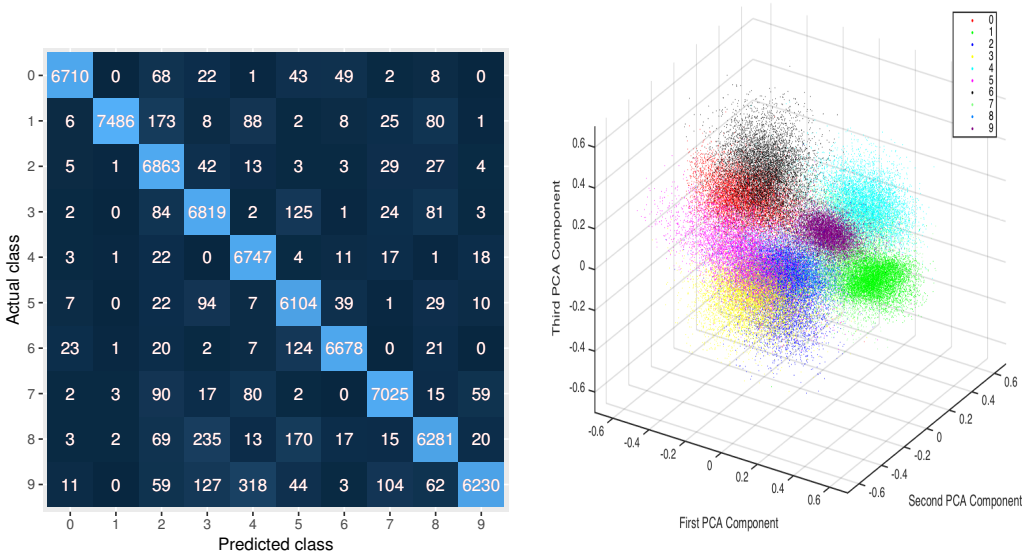


Figure 6.3 – *Left:* Confusion matrix of the DAC clustering for the MNIST data set of handwritten digits. *Right:* PCA rendering of the 10-dimensional nonlinear embedding learnt by DAC for MNIST. Due to adversarial regularization, the true classes look nearly Gaussian, which explains the extremely good clustering results of DAC.

Figure 6.4 – Fake digits generated from the ten classes inferred by DAC. From top to bottom, we go further and further away, in the 10-dimensional embedding, from the cluster means. The first row images correspond to the decoded cluster means.

have to be overcome. The exact expectation-maximization approach could be replaced by a variational approximation. Following Kucukelbir et al. (2017), using stochastic gradient ascent together with variational inference could lead to an online and scalable solution.

Another simple step forward is to take model uncertainty into account by performing Bayesian model averaging. While the model space is too large to be visited thoroughly, a greedy approach similar to Occam's window (Madigan and Raftery, 1994) might be designable.

6.3.2 Hierarchical and anisotropic extensions of GSPPCA

GSPPCA is essentially an empirical Bayes algorithm. However, a fully Bayesian approach could be considered in the future. First, it would be straightforward to add a prior distribution to the noise variance – an advantage of this is that it would solve the problem of having to estimate this parameter. Another level of hierarchy could also be added regarding the precision α of the prior on the loading coefficients. Surprisingly, early investigations suggest that, by carefully choosing the prior distribution over α, a closed-form expression of the marginal likelihood can still be derived.

As an other simple extension, we could relax the parametric assumption of the noise component: for example, a Gaussian with anisotropic covariance could be considered, as in factor analysis.

6.3.3 Consistency of NGPPCA

While NGPPCA was mainly developed motivated by nonasymptotic scenarios, it could be possible to study the asymptotic properties of the exact marginal likelihood criterion. An interesting application would be the improvement of the heuristic used for hyperparameter tuning. Indeed, we could limit our grid search to a grid of hyperparameter values that are known to lead to model selection consistency. This idea to use asymptotics to find reasonable ranges of hyperparameters was for example applied by Liang et al. (2008) in a linear regression context.

A

Multiplying a Gaussian by a Gaussian vector

A.1 Introduction

Wishart and Bartlett (1932) proved that the inner product of two independent bidimensional standard Gaussian vectors follows a Laplace distribution. This result is deeply linked to the fact that the Laplace distribution can be represented as an infinite scale mixture of Gaussians with gamma mixing distribution. Specifically, if σ^2 follows a Gamma$(1, 1/2)$ distribution and $x|\sigma \sim \mathcal{N}(0, \sigma^2)$, then x follows a standard Laplace distribution[1]. This representation – which was recently named the *Gauss-Laplace representation* by Ding and Blitzstein (2017) following a blog post by Christian P. Robert[2] – is particularly useful if one wants to simulate a Laplace random variable: its use constitutes for example the cornerstone of the Gibbs sampling scheme for the Bayesian lasso of Park and Casella (2008).

In this short appendix, we are interested in studying links between multivariate counterparts of these two characterizations. More specifically, we give a new simple characterization of the *multivariate generalized Laplace distribution* of Kotz et al. (2001). As a corollary, we

[1] The shape-rate parametrization of the gamma distribution is used through this appendix. Note also that a standard Laplace distribution is centered with variance two.

[2] https://xianblog.wordpress.com/2015/10/14/gauss-to-laplace-transmutation/

show that the product of a zero-mean Gaussian matrix with in independent and identically distributed (i.i.d.) columns and a zero-mean isotropic Gaussian vector follows a symmetric multivariate generalized Laplace distribution, a result that has useful applications for Bayesian principal component analysis (see Chapters 4 and 5).

A.2 The multivariate generalized Laplace distribution

While the definition of the univariate Laplace distribution is widely undisputed, there exist several different generalizations of this distribution to higher dimensions – a comprehensive review of such generalizations can be found in the monograph of Kotz et al. (2001). In particular, McGraw and Wagner (1968) introduced a zero-mean elliptically contoured bidimensional Laplace distribution with univariate Laplace marginals. This distribution was later generalized to the p-dimensional setting by Anderson (1992), considering characteristic functions of the form

$$\forall \mathbf{u} \in \mathbb{R}^p,\ \phi(\mathbf{u}) = \frac{1}{1 + \frac{1}{2}\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}},$$

where $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$. This distribution was notably promoted by Eltoft et al. (2006) and is arguably the most popular multivariate generalization of the Laplace distribution (Kotz et al., 2001, p. 229). Among its advantages, this distribution can be slightly generalized to model skewness, by building on characteristic functions of the form

$$\forall \mathbf{u} \in \mathbb{R}^p,\ \phi(\mathbf{u}) = \frac{1}{1 + \frac{1}{2}\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} - i\boldsymbol{\mu}^T \mathbf{u}},$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ accounts for asymmetry. Similarly to the univariate Laplace distribution, this asymmetric multivariate generalization is infinitely divisible (Kotz et al., 2001, p. 256). Therefore, it can be associated with a specific Lévy process (Kyprianou, 2014, p. 5), whose increments will follow yet another generalization of the Laplace distribution, the *multivariate generalized asymmetric Laplace distribution*. This distribution, introduced by Kotz et al. (2001, p. 257) and further studied by Kozubowski et al. (2013), will be the cornerstone of our analysis of multivariate characterizations of Laplace and Gaussian distributions.

Definition A.1. *A random variable $\mathbf{z} \in \mathbb{R}^p$ is said to have a* ***multivariate generalized asymmetric Laplace distribution*** *with parameters $s > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$ if its characteristic function is*

$$\forall \mathbf{u} \in \mathbb{R}^p,\ \phi_{\mathrm{GAL}_p(\boldsymbol{\Sigma},\boldsymbol{\mu},s)}(\mathbf{u}) = \left(\frac{1}{1 + \frac{1}{2}\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} - i\boldsymbol{\mu}^T \mathbf{u}}\right)^s.$$

It is denoted by $\mathbf{z} \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s)$.

General properties of the generalized asymmetric Laplace distribution are distribution are discussed by Kozubowski et al. (2013). We list here a few useful ones.

Proposition A.1. *Let $s > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$. If $\mathbf{z} \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s)$, we have $\mathbb{E}(\mathbf{z}) = s\boldsymbol{\mu}$ and $\mathbb{C}\mathrm{ov}(\mathbf{z}) = s(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)$. Moreover, if $\boldsymbol{\Sigma}$ is positive definite, the density of $\mathbf{z}$*

is given by

$$\forall \mathbf{x} \in \mathbb{R}^p,\ f_{\mathbf{z}}(\mathbf{x}) = \frac{2e^{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}}{(2\pi)^{p/2}\Gamma(s)\sqrt{\det\boldsymbol{\Sigma}}}\left(\frac{Q_{\boldsymbol{\Sigma}}(\mathbf{x})}{C(\boldsymbol{\Sigma},\boldsymbol{\mu})}\right)^{s-p/2} K_{s-p/2}\left(Q_{\boldsymbol{\Sigma}}(\mathbf{x})C(\boldsymbol{\Sigma},\boldsymbol{\mu})\right),$$

where $Q_{\boldsymbol{\Sigma}}(\mathbf{x}) = \sqrt{\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}$, $C(\boldsymbol{\Sigma},\boldsymbol{\mu}) = \sqrt{2+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$ *and* $K_{s-p/2}$ *is the modified Bessel function of the second kind of order* $s-p/2$.

Note that the $\mathrm{GAL}_1(2b^2, 0, 1)$ case corresponds to a centered univariate Laplace distribution with scale parameter $b > 0$. In the symmetric case ($\boldsymbol{\mu} = 0$) and when $s = 1$, we recover the multivariate generalization of the Laplace distribution of Anderson (1992).

An appealing property of the multivariate generalized Laplace distribution is that it is also endowed with a multivariate counterpart of the Gauss-Laplace representation.

Theorem A.1 (Generalized Gauss-Laplace representation)**.** *Let* $s > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ *and* $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$. *If* $u \sim \mathrm{Gamma}(s, 1)$ *and* $\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ *is independent of* u, *we have*

$$u\boldsymbol{\mu} + \sqrt{u}\mathbf{x} \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s). \tag{A.1}$$

A proof of this result can be found in Kotz et al. (2001, chap. 6). This representation explains why the multivariate generalized Laplace distribution can also be seen as a multivariate generalization of the *variance-gamma distribution* which is widely used in the field of quantitative finance (Madan et al., 1998). Infinite mixtures similar to (A.1) are called *variance-mean mixtures* (Barndorff-Nielsen et al., 1982) and are discussed for example by Yu (2017).

Another useful property of the multivariate generalized Laplace distribution is that, under some conditions, it is closed under convolution.

Proposition A.2. *Let* $s_1, s_2 > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ *and* $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$. *If* $\mathbf{z}_1 \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s_1)$ *and* $\mathbf{z}_2 \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s_2)$ *are independant random variables, then*

$$\mathbf{z}_1 + \mathbf{z}_2 \sim \mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s_1 + s_2). \tag{A.2}$$

Proof. Since $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent, we have for all $\mathbf{u} \in \mathbb{R}^p$,

$$\phi_{\mathbf{z}_1+\mathbf{z}_2}(\mathbf{u}) = \phi_{\mathrm{GAL}_p(\boldsymbol{\Sigma},\boldsymbol{\mu},s_1)}(\mathbf{u})\phi_{\mathrm{GAL}_p(\boldsymbol{\Sigma},\boldsymbol{\mu},s_2)}(\mathbf{u}) = \left(\frac{1}{1+\frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u} - i\boldsymbol{\mu}^T\mathbf{u}}\right)^{s_1+s_2}$$

which is the characteristic function of the $\mathrm{GAL}_p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, s_1 + s_2)$ distribution. □

A.3 A new characterization involving a product between a Gaussian matrix and a Gaussian vector

We now state our main theorem, which gives a new characterization of multivariate generalized Laplace distributions with half-integer shape parameters.

Theorem A.2. *Let* $\mathbf{W}$ *be a* $p \times d$ *random matrix with i.i.d. columns following a* $\mathcal{N}(0, \mathbf{\Sigma})$ *distribution,* $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_d)$ *be a Gaussian vector independent from* $\mathbf{W}$ *and let* $\boldsymbol{\mu} \in \mathbb{R}^p$*. We have*

$$\mathbf{W}\mathbf{y} + ||\mathbf{y}||_2^2 \boldsymbol{\mu} \sim \mathrm{GAL}_p(2\mathbf{\Sigma}, 2\boldsymbol{\mu}, d/2). \tag{A.3}$$

Proof. For each $k \in \{1, ..., d\}$ let $\mathbf{w}_k$ be the k-th column of $\mathbf{W}$, $u_k = y_k^2$ and $\boldsymbol{\xi}_k = y_k \mathbf{w}_k + y_k^2 \boldsymbol{\mu}$. To prove the theorem, we will prove that $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_d$ follow a GAL distribution and use the decomposition

$$\mathbf{W}\mathbf{y} + ||\mathbf{y}||_2^2 \boldsymbol{\mu} = \sum_{k=1}^{d} \boldsymbol{\xi}_k.$$

Let $k \in \{1, ..., d\}$. Since $\mathbf{y}$ is standard Gaussian, $u_k = y_k^2$ follows a $\chi^2(1)$ distribution, or equivalently a Gamma$(1/2, 1/2)$ distribution. Therefore, $u_k/2 \sim$ Gamma$(1/2, 1)$. Moreover, note that $\sqrt{u_k}\mathbf{w}_k = |y_k|\mathbf{w}_k = y_k \mathrm{sign}(y_k)\mathbf{w}_k \overset{d}{=} y_k \mathbf{w}_k$ since $|y_k|$ and $\mathrm{sign}(y_k)$ are independent and $\mathrm{sign}(y_k)\mathbf{w}_k \overset{d}{=} \mathbf{w}_k$. Therefore, according to the generalized Gauss-Laplace representation, we have

$$\boldsymbol{\xi}_k \overset{d}{=} \sqrt{\frac{u_k}{2}}\sqrt{2}\mathbf{w}_k + \frac{u_k}{2} 2\boldsymbol{\mu} \sim \mathrm{GAL}_p(2\mathbf{\Sigma}, 2\boldsymbol{\mu}, 1/2).$$

Since $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_d$ are i.i.d. and following a $\mathrm{GAL}_p(2\mathbf{\Sigma}, 2\boldsymbol{\mu}, 1/2)$ distribution, we can use Proposition A.2 to conclude that

$$\mathbf{W}\mathbf{y} + ||\mathbf{y}||_2^2 \boldsymbol{\mu} = \sum_{k=1}^{d} \boldsymbol{\xi}_k \sim \mathrm{GAL}_p(2\mathbf{\Sigma}, 2\boldsymbol{\mu}, d/2).$$

□

In the symmetric case ($\boldsymbol{\mu} = 0$), this result gives the distribution of the product between a Gaussian matrix with i.i.d. columns and an isotropic Gaussian vector.

Corollary A.2.1. *Let* $\mathbf{W}$ *be a* $p \times d$ *random matrix with i.i.d. columns following a* $\mathcal{N}(0, \mathbf{\Sigma})$ *distribution and let* $\mathbf{y} \sim \mathcal{N}(0, \alpha \mathbf{I}_d)$ *be a Gaussian vector independent from* $\mathbf{W}$*. Then*

$$\mathbf{W}\mathbf{y} \sim \mathrm{GAL}_p(2\alpha\mathbf{\Sigma}, 0, d/2). \tag{A.4}$$

Moreover, if u *is a standard Gamma variable with shape parameter* $d/2$ *and if* $\mathbf{x} \sim \mathcal{N}(0, 2\alpha\mathbf{\Sigma})$ *is a Gaussian vector independent of* u*, then*

$$\mathbf{W}\mathbf{y} \overset{d}{=} \sqrt{u}\mathbf{x}. \tag{A.5}$$

Less general versions of Theorem A.2 have been proven in the past, dating back to the derivation of the inner product of two i.i.d. standard Gaussian vectors by Wishart and Bartlett (1932). In particular, the unidimensional case ($p = 1$) was recently proven by Gaunt (2014) in order to obtain bounds for the convergence rate of random sums involving Gaussian products.

A.4 Perspectives

The new characterization presented in this appendix may notably prove useful in two contexts.

First, it indicates a new way of handling situations involving the product of a Gaussian matrix and a Gaussian vector. An important instance is the Bayesian factor analysis model (Lopes and West, 2004), of which principal component analysis is a particular case. In this framework, the marginal distribution of the data, which is essential for model selection purposes, can be derived using representation (A.5) together with the Gauss-Laplace representation (see Chapters 4 and 5).

Moreover, our characterization offers a means to get around problems encountered when dealing with distributions related to the GAL distribution. For example, representation (A.3) might lead to alternative estimation strategies for some problems related to portfolio allocation (Mencía and Sentana, 2009; Breymann and Lüthi, 2013) or cluster analysis (McNicholas et al., 2013; Franczak et al., 2014).

B

Benchmark Study for Sparse Linear Regression

This appendix presents the entire benchmark study for sparse linear regression algorithms described in Chapter 3.

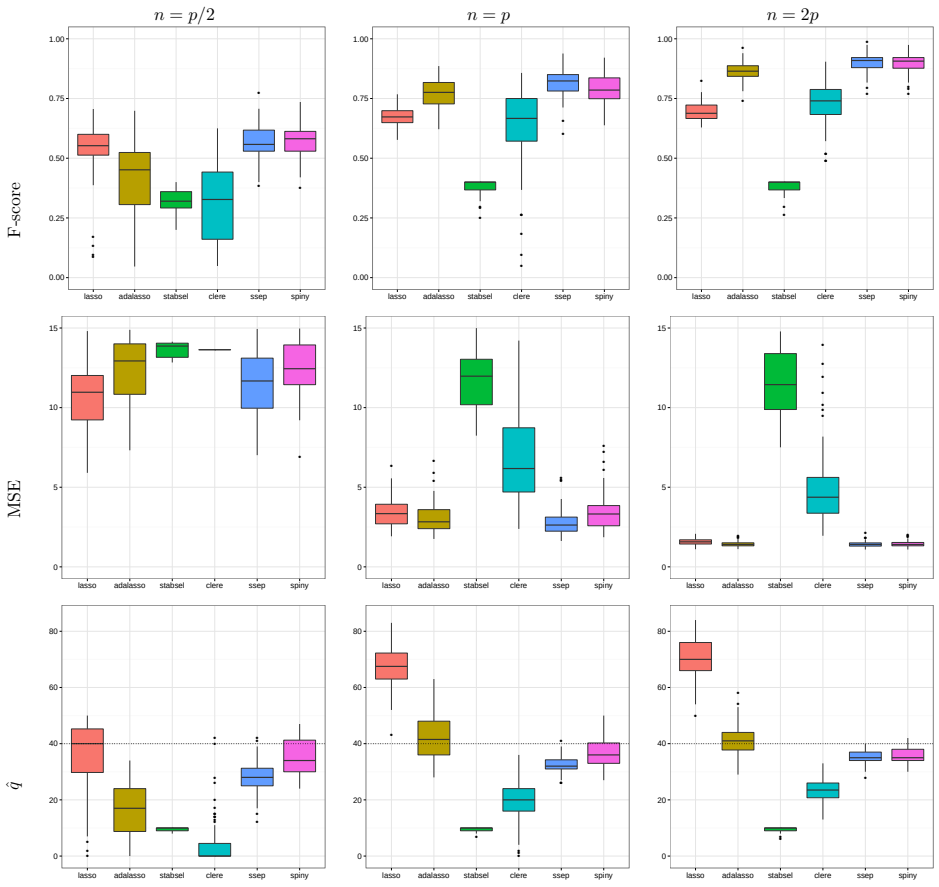


Figure 1: Scenario "blockwise" with $\rho = 0.25$.

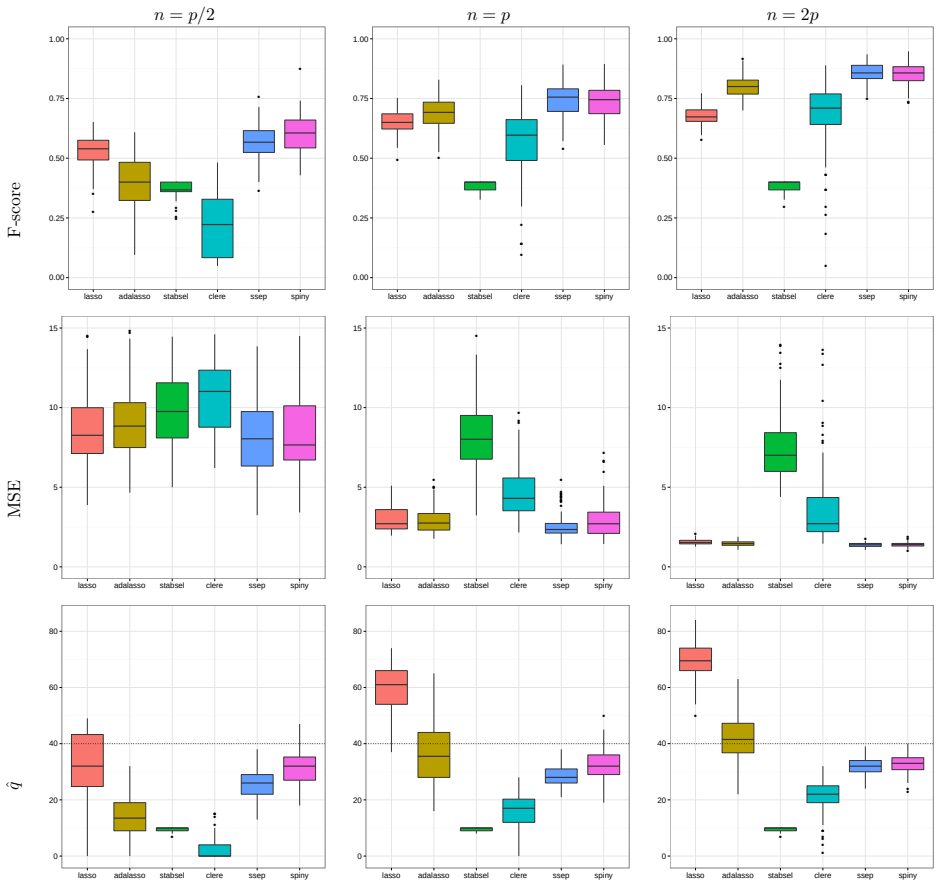


Figure 2: Scenario "blockwise" with $\rho = 0.75$.

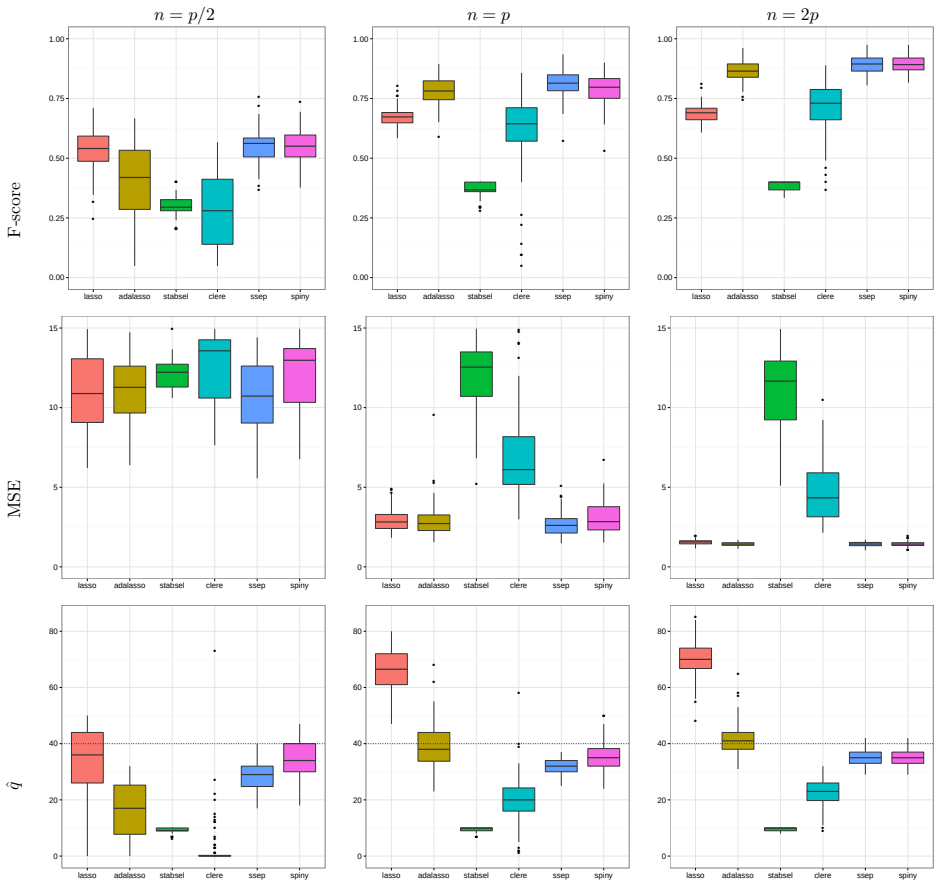


Figure 3: Scenario "uniform" with $\rho = 0.25$.

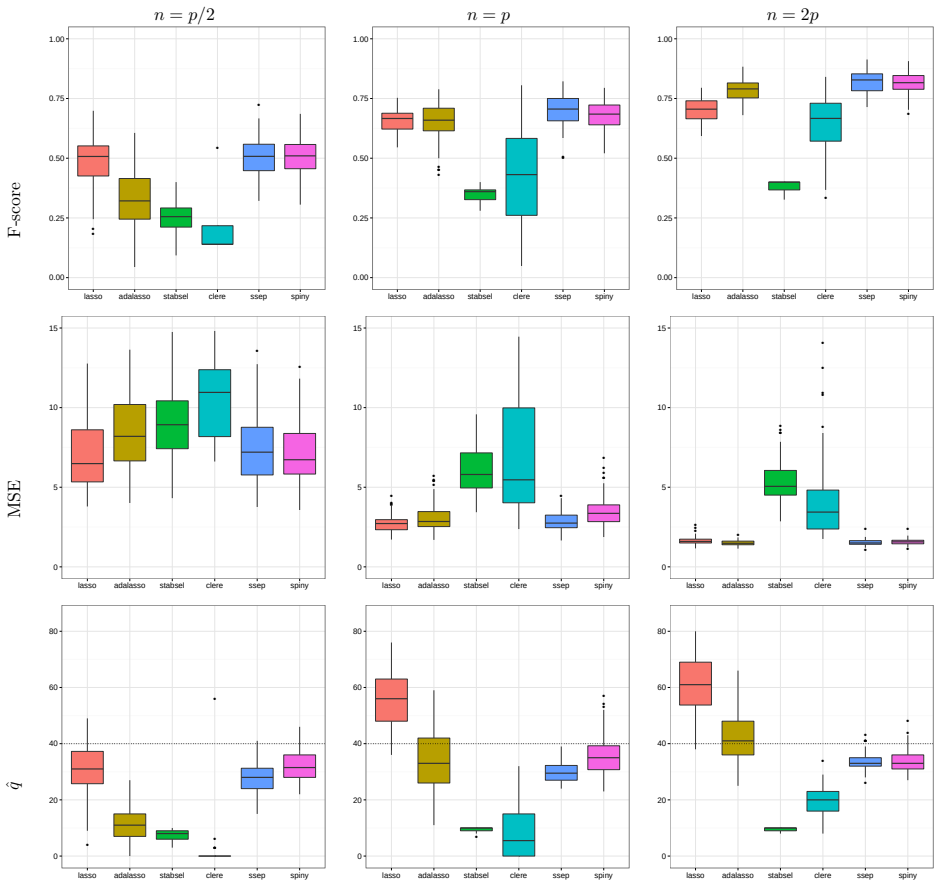


Figure 4: Scenario "uniform" with $\rho = 0.75$.

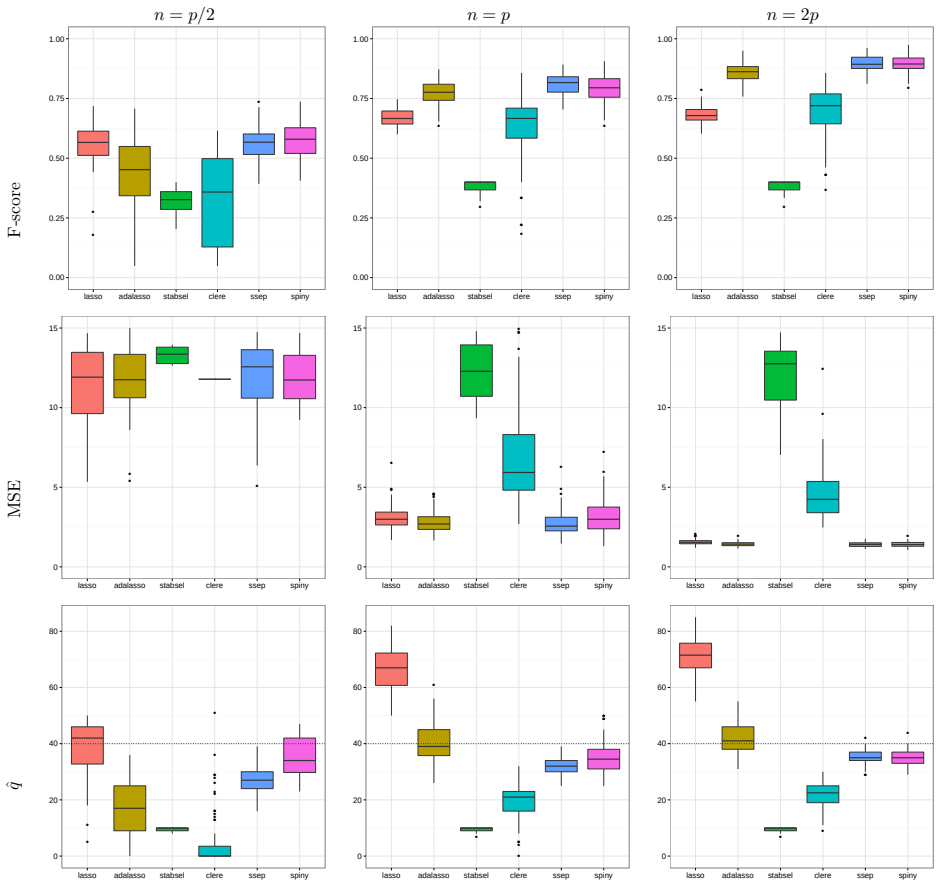


Figure 5: Scenario "Toeplitz" with $\rho = 0.25$.

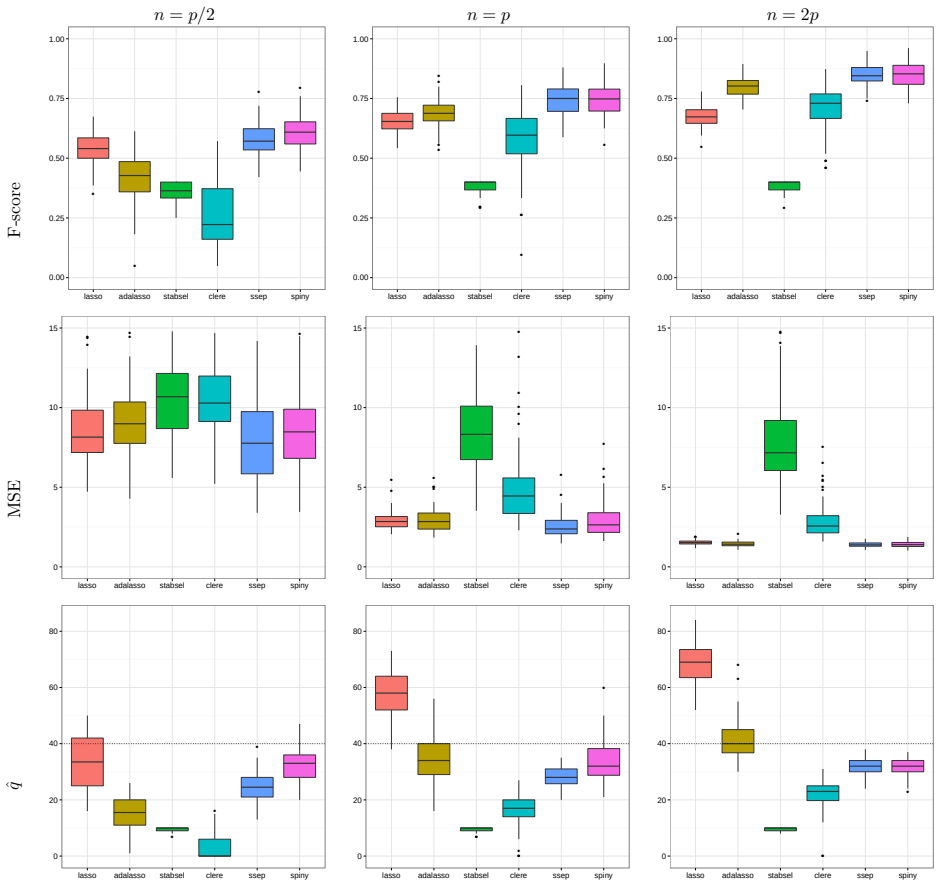


Figure 6: Scenario "Toeplitz" with $\rho = 0.75$.

Bibliography

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1965.

M. Aitkin. Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–142, 1991.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory.*, pages 267–281. Akademia Kiado. B. N. Petrov and F. Csaki, editors., 1973.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

P. Alquier and K. Lounici. PAC-bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

M. Aminghafari, N. Cheze, and J.-M. Poggi. Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis*, 50(9):2381–2398, 2006.

D. N. Anderson. A multivariate Linnik distribution. *Statistics & Probability Letters*, 14(4): 333–336, 1992.

T. Ando. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, pages 443–458, 2007.

T. Ando. Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood. *Journal of Multivariate Analysis*, 100(8):1717–1726, 2009.

T. Ando and R. Tsay. Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting*, 26(4):744–763, 2010.

C. Archambeau and F. Bach. Sparse probabilistic projections. In *Advances in neural information processing systems*, pages 73–80, 2009.

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

R. Bapat and T. Raghavan. *Nonnegative matrices and applications*, volume 64. Cambridge University Press, 1997.

M. Baragatti and D. Pommeret. A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics & Data Analysis*, 56(6):1920–1934, 2012.

R. F. Barber, M. Drton, and K. M. Tan. Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data*, pages 15–36. Springer, 2016.

O. Barndorff-Nielsen, J. Kent, and M. Sørensen. Normal variance-mean mixtures and z distributions. *International Statistical Review/Revue Internationale de Statistique*, 50(2): 145–159, 1982.

M. Bartlett. A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44(1-2):533–534, 1957.

J. Bayarri and J. Berger. Hypothesis testing and model uncertainty. In P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens, editors, *Bayesian Theory and Applications*, pages 361 – 394. Oxford university press, 2013.

R. Bellman. *Dynamic programming*. Princeton University Press, 1957.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 289–300, 1995.

J. O. Berger. *Statistical decision theory and Bayesian analysis (second edition)*. Springer: New York, 1985.

J. O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003.

J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

J. M. Bernardo and A. F. M. Smith. *Bayesian theory*. John Wiley and Sons, 1994.

M. Bertoletti, N. Friel, and R. Rastelli. Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron*, 73(2):177–199, 2015.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability theory and related fields*, 138(1):33–73, 2007.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer: New York, 2006.

C. M. Bishop. Bayesian PCA. *Advances in neural information processing systems*, pages 382–388, 1999a.

C. M. Bishop. Variational principal components. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, pages 509–514, 1999b.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (in press), 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

L. Bouranis, N. Friel, and F. Maire. Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. *arXiv preprint arXiv:1706.06344*, 2017.

C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, 22(1):301–324, 2012.

C. Bouveyron, S. Girard, and C. Schmid. High-dimensional discriminant analysis. *Communications in Statistics – Theory and Methods*, 36(14):2607–2623, 2007a.

C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007b.

C. Bouveyron, G. Celeux, and S. Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713, 2011.

C. Bouveyron, E. Côme, and J. Jacques. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4): 1726–1760, 2015.

C. Bouveyron, P. Latouche, and P.-A. Mattei. Bayesian variable selection for globally sparse probabilistic PCA. *Technical report, HAL-01310409, Université Paris Descartes*, 2016.

C. Bouveyron, P. Latouche, and P.-A. Mattei. Exact dimensionality selection for Bayesian PCA. *Technical report, HAL-01484099, Université Paris Descartes*, 2017.

G. E. P. Box. Robustness in the strategy of scientific model building. *Robustness in statistics*, 1:201–236, 1979.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman and J. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

W. Breymann and D. Lüthi. ghyp: A package on generalized hyperbolic distributions. *Manual for R Package ghyp*, 2013.

R. Bro, K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers. Cross-validation of component models: a critical look at current methods. *Analytical and bioanalytical chemistry*, 390 (5):1241–1251, 2008.

M. J. Brusco. A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics & Data Analysis*, 77:38–53, 2014.

R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *Journal on Scientific and Statistical Computing*, 16:1190–1208, 1995.

E. J. Candès. Mathematics of sparsity (and a few other things). In *Proceedings of the International Congress of Mathematicians, Seoul, South Korea*, 2014.

E. J. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

P. Carbonetto and M. Stephens. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1): 73–108, 2012.

B. P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 473–484, 1995.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37, 2016.

O. Catoni. *Pac-Bayesian Supervised Classification*, volume 56 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

R. B. Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1 (2):245–276, 1966.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

G. Celeux, M. El Anbari, J.-M. Marin, and C. P. Robert. Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(2):477–502, 2012.

J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical methods for data analysis*. Belmont, CA: Wadsworth, 1983.

T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. PCANet: A simple deep learning baseline for image classification? *Image Processing, IEEE Transactions on*, 24(12): 5017–5032, 2015.

W.-C. Chang. On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, pages 267–275, 1983.

D. Chatterjee, T. Maitra, and S. Bhattacharya. A short note on almost sure convergence of Bayes factors in the general set-up. *arXiv preprint arXiv:1703.04956*, 2017.

M.-H. Chen, Q.-M. Shao, and J. G. Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2000.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

S. Chib and T. A. Kuffner. Bayes factor consistency. *arXiv preprint arXiv:1607.00292*, 2016.

M. Clyde and E. I. George. Model uncertainty. *Statistical science*, 19:81–94, 2004.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

K. Csilléry, M. G. B. Blum, O. E. Gaggiotti, and O. François. Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.

A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A*, pages 278–292, 1984.

A. P. Dawid. Posterior model probabilities. In P. S. Bandyopadhyay and M. R. Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 607 – 630. North-Holland, Amsterdam, 2011.

F. De la Torre and T. Kanade. Discriminative cluster analysis. pages 241–248, 2006.

C.-A. Deledalle, J. Salmon, and A. S. Dalalyan. Image denoising with patch based PCA: local versus global. In *Proceedings of the British Machine Vision Conference*, pages 25.1–25.10, 2011.

P. Dellaportas, J. J. Forster, and I. Ntzoufras. Joint specification of model space and parameter space prior distributions. *Statistical Science*, pages 232–246, 2012.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

P. Ding and J. K. Blitzstein. On the Gaussian mixture representation of the Laplace distribution. *The American Statistician*, in press, 2017.

S. Drăghici. *Statistics and data analysis for microarrays using R and Bioconductor*. CRC Press, 2012.

D. Draper. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 45–97, 1995.

D. Draper. Comment on "Bayesian model averaging: A tutorial". *Statistical Science*, 14(4): 405–409, 1999.

M. Drton and M. Plummer. A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380, 2017.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

T. Eltoft, T. Kim, and T.-W. Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

A. Erkanli. Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *Journal of the American Statistical Association*, 89 (425):250–258, 1994.

A. Etz and E.-J. Wagenmakers. J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, 32(2):313–329, 2017.

N. Evangelopoulos, X. Zhang, and V. R. Prybutok. Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems*, 21(1):70–86, 2012.

A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The Reactome pathway Knowledgebase. *Nucleic acids research*, 44(D1): D481–D487, 2016.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*. Chapman and Hall, 1990.

R. A. Fisher. Presidential address. *Sankhyā: The Indian Journal of Statistics*, pages 14–17, 1938.

M. Fop and T. B. Murphy. Variable selection methods for model-based clustering. *arXiv preprint arXiv:1707.00306*, 2017.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

B. C. Franczak, R. P. Browne, and P. D. McNicholas. Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1149–1157, 2014.

N. Friel. Evidence and Bayes factor estimation for Gibbs random fields. *Journal of Computational and Graphical Statistics*, 22(3):518–532, 2013.

N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607, 2008.

N. Friel and J. Wyse. Estimating the evidence–a review. *Statistica Neerlandica*, 66(3): 288–308, 2012.

N. Friel, J. P. McKeone, C. J. Oates, and A. N. Pettitt. Investigation of the widely applicable Bayesian information criterion. *Statistics and Computing*, 27(3):833–844, 2017.

W. Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

T. Gao and V. Jojic. Degrees of freedom in deep neural networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.

R. E. Gaunt. Variance-Gamma approximation via Stein's method. *Electronic Journal of Probability*, 19(38):1–33, 2014.

M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

A. Gelman and C. R. Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis (third edition)*. Chapman & Hall, 2013.

E. I. George and D. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

I. J. Good. Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, 66(2):393–396, 1979.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

A. Gramfort, D. Strohmeier, J. Haueisen, M. S. Hämäläinen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70:410–422, 2013.

Y. Grandvalet, J. Chiquet, and C. Ambroise. Sparsity by worst-case quadratic penalties. *arXiv preprint arXiv:1210.2077*, 2016.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, pages 711–732, 1995.

Q. Gu, Z. Li, and J. Han. Joint feature selection and subspace learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 22, page 1294, 2011.

Y. Guan and J. G. Dy. Sparse probabilistic principal component analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 185–192, 2009.

S. F. Gull. Bayesian inductive inference and maximum entropy. In *Maximum-entropy and Bayesian methods in science and engineering*, pages 53–74. Springer, 1988.

I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: Beyond the Bayesian/frequentist divide. *The Journal of Machine Learning Research*, 11:61–87, 2010.

J. B. S. Haldane. A note on inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 55–61. Cambridge University Press, 1932.

C. Han and B. P. Carlin. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

A. Hannachi, I. T. Jolliffe, D. B. Stephenson, and N. Trendafilov. In search of simple structures in climate: simplifying EOFs. *International journal of climatology*, 26(1):7–28, 2006.

P. Hartman and G. S. Watson. "Normal" distribution functions on spheres and the modified Bessel functions. *The Annals of Probability*, pages 593–607, 1974.

D. I. Hastie and P. J. Green. Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338, 2012.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

D. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

S. Hee, W. J. Handley, M. P. Hobson, and A. N. Lasenby. Bayesian model selection without evidences: application to the dark energy equation-of-state. *Monthly Notices of the Royal Astronomical Society*, 455(3):2461–2473, 2016.

D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *The Journal of Machine Learning Research*, 14(1):1891–1945, 2013.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, pages 382–401, 1999.

P. D. Hoff. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478):674–685, 2007.

C. C. Holmes, F. Caron, J. E. Griffin, and D. A. Stephens. Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Analysis*, 10(2):297–320, 2015.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

D. C. Hoyle. Automatic PCA dimension selection for high dimensional data and small sample sizes. *Journal of Machine Learning Research*, 9(Dec):2733–2759, 2008.

C. K. Hsiao. Approximate Bayes factors when a mode occurs on the boundary. *Journal of the American Statistical Association*, 92(438):656–663, 1997.

X. Huang, J. Wang, and F. Liang. A variational algorithm for Bayesian variable selection. *arXiv preprint arXiv:1602.07640*, 2016.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

F. Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.

H. Ishwaran and J. Rao. Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association*, 100(471):764–780, 2005a.

H. Ishwaran and J. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, pages 730–773, 2005b.

H. Ishwaran, U. Kogalur, and J. Rao. spikeslab: Prediction and variable selection using spike and slab regression. *R Journal*, 2(2), 2010.

D. A. Jackson. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214, 1993.

W. H. Jefferys and J. O. Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 80(1):64–72, 1992.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 1939.

H. Jeffreys. *Theory of Probability (third edition)*. Oxford University Press, 1961.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics*, 2009.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12(Oct):2777–2824, 2011.

V. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.

I. T. Jolliffe. Discarding variables in a principal component analysis. I: Artificial data. *Applied statistics*, pages 160–173, 1972.

I. T. Jolliffe. Discarding variables in a principal component analysis. II: Real data. *Applied Statistics*, pages 21–31, 1973.

I. T. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 374(2065), 2016.

M. I. Jordan. What are the open problems in Bayesian statistics? *The ISBA Bulletin*, 18 (1):568, 2011.

J. Josse and F. Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6): 1869–1879, 2012.

M. Journée. *Geometric algorithms for component analysis with a view to gene expression data analysis*. PhD thesis, Université de Liège, 2009.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11: 517–553, 2010.

K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau. Testing hypotheses via a mixture estimation model. *arXiv preprint arXiv:1412.2044*, 2014.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

R. E. Kass and D. Steffey. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726, 1989.

R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90 (431):928–934, 1995.

R. E. Kass, L. Tierney, and J. B. Kadane. The validity of posterior expansions based on Laplace's method. In S. Geisser, J. Hodges, S. Press, and A. Zellner, editors, *Bayesian and likelihood methods in statistics and econometrics*, pages 473–488. 1990.

J. M. Keynes. *A Treatise on Probability*. London: Macmillan, 1921.

Z. Khan, F. Shafait, and A. Mian. Joint group sparse pca for compressed hyperspectral imaging. *Image Processing, IEEE Transactions on*, 24(12):4934–4942, 2015.

R. Khanna, J. Ghosh, R. Poldrack, and O. Koyejo. Sparse submodular probabilistic PCA. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 453–461, 2015.

S. Kotz, T. J. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, exonomics, engineering, and finance*. Number 183. Springer Science & Business Media, 2001.

T. J. Kozubowski, K. Podgórski, and I. Rychlik. Multivariate generalized Laplace distribution and related random fields. *Journal of Multivariate Analysis*, 113:59–72, 2013.

N. Kraemer, J. Schaefer, and A.-L. Boulesteix. Regularized estimation of large-scale gene regulatory networks using gaussian graphical models. *BMC Bioinformatics*, 10(384), 2009.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

J. Kuha. AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229, 2004.

A. Kyprianou. *Fluctuations of Lévy processes with applications: Introductory Lectures*. Springer Science & Business Media, 2014.

H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480. ACM, 2007.

P. Latouche, E. Birmelé, and C. Ambroise. Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115, 2012.

P. Latouche, E. Birmelé, and C. Ambroise. Model selection in overlapping stochastic block models. *Electronic journal of statistics*, 8(1):762–794, 2014.

P. Latouche, P.-A. Mattei, C. Bouveyron, and J. Chiquet. Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression. *Journal of Multivariate Analysis*, 146:177–190, 2016.

M. Lavine and M. J. Schervish. Bayes factors: what they are and what they are not. *The American Statistician*, 53(2):119–122, 1999.

D. N. Lawley. A modified method of estimation in factor analysis and some large sample results. *Proceedings of the Uppsala Symposium on Psychological Factor Analysis, Uppsala, Sweden*, pages 35–42, 1953.

M. Lázaro-Gredilla and M. K. Titsias. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in neural information processing systems*, pages 2339–2347, 2011.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, 2001.

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481): 410–423, 2008.

B. Liebmann, A. Friedl, and K. Varmuza. Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Analytica Chimica Acta*, 642 (1):171–178, 2009.

S. Lin, B. Sturmfels, and Z. Xu. Marginal likelihood integrals for mixtures of independence models. *Journal of Machine Learning Research*, 10:1611–1631, 2009.

D. V. Lindley. A statistical paradox. *Biometrika*, 44(1-2):187–192, 1957.

D. V. Lindley. Some comments on Bayes factors. *Journal of Statistical Planning and Inference*, 61(1):181–189, 1997.

T.-Y. Liu, L. Trinchera, A. Tenenhaus, D. Wei, and A. O. Hero. Globally sparse PLS regression. In *New Perspectives in Partial Least Squares and Related Methods*, pages 117–127. Springer, 2013.

B. A. Logsdon, G. E. Hoffman, and J. G. Mezey. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC bioinformatics*, 11(1):58, 2010.

H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67, 2004.

L. Lorch. Inequalities for some Whittaker functions. *Archivum Mathematicum*, 3(1):1–9, 1967.

M. Luck, G. Bertho, M. Bateson, A. Karras, A. Yartseva, E. Thervet, C. Damon, and N. Pallet. Rule-mining for the early prediction of chronic kidney disease based on metabolomics and multi-source data. *PloS one*, 11(11):e0166905, 2016.

D. J. C. MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1991.

D. J. C. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992a.

D. J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992b.

D. J. C. MacKay. Bayesian methods for backpropagation networks. In *Models of neural networks III*, pages 211–254. Springer, 1994.

D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11(5):1035–1068, 1999.

D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

D. B. Madan, P. P. Carr, and E. C. Chang. The variance gamma process and option pricing. *European Finance Review*, 2(1):79–105, 1998.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

M. Mäechler. `Bessel`: *Bessel – Bessel Functions Computations and Approximations*, 2013. URL `https://CRAN.R-project.org/package=Bessel`. R package version 0.5-5.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

M. Marbac and M. Sedki. Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, 27(4):1049–1063, 2017.

J.-M. Marin and C. P. Robert. Importance sampling methods for Bayesian discrimination between embedded models. *Frontiers of Statistical Decision Making and Bayesian Analysis*, pages 513–527, 2010.

J.-M. Marin and C. P. Robert. *Bayesian essentials with R*. Springer, 2014.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, pages 1–14, 2012.

J.-M. Marin, P. Pudlo, A. Estoup, and C. P. Robert. Likelihood-free model choice. *arXiv preprint arXiv:1503.07689*, 2015.

H. Markowitz. Portfolio selection. *The journal of Finance*, 7(1):77–91, 1952.

M. Masaeli, Y. Yan, Y. Cui, G. Fung, and J. G. Dy. Convex principal feature selection. In *In SIAM International Conference on Data Mining*, pages 619–628, 2010.

P.-A. Mattei. Multiplying a Gaussian matrix by a Gaussian vector. *Statistics & Probability Letters*, 128:67–70, 2017a.

P.-A. Mattei. Discussion on "a Bayesian information criterion for singular models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):370–371, 2017b.

P.-A. Mattei, C. Bouveyron, and P. Latouche. Globally sparse probabilistic PCA. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 976–984, 2016.

C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, 102(10):1374–1387, 2011.

D. G. Mayo and A. Spanos. *Error and inference*. Cambridge University Press, Cambridge, 2009.

R. Mazumder and P. Radchenko. The discrete Dantzig selector: Estimating sparse linear models via mixed integer linear optimization. *IEEE Transactions on Information Theory*, 63(5):3053–3075, 2017.

D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234. ACM, 1998.

J. McElhinney, G. Downey, and T. Fearn. Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *Journal of Near Infrared Spectroscopy*, 7(3):145–154, 1999.

D. McGraw and J. Wagner. Elliptically symmetric distributions. *IEEE Transactions on Information Theory*, 14(1):110–120, 1968.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions. Second Edition.* John Wiley & Sons, New York, 2008.

S. M. McNicholas, P. D. McNicholas, and R. P. Browne. Mixtures of variance-gamma distributions. *arXiv preprint arXiv:1309.2695*, 2013.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 27, 2010.

J. Mencía and E. Sentana. Multivariate location–scale mixtures of normals and mean–variance–skewness portfolio allocation. *Journal of Econometrics*, 153(2):105–121, 2009.

J. A. Miller, C. Cai, P. Langfelder, D. H. Geschwind, S. M. Kurian, D. R. Salomon, and S. Horvath. Strategies for aggregating gene expression data: the collapseRows R function. *BMC bioinformatics*, 12(1):1, 2011.

T. P. Minka. Automatic choice of dimensionality for PCA. In *NIPS*, volume 13, pages 598–604, 2000.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

A. J. Minn, G. P. Gupta, D. Padua, P. Bos, D. X. Nguyen, D. Nuyten, B. Kreike, Y. Zhang, Y. Wang, H. Ishwaran, J. A. Foekens, M. van de Vijver, and J. Massagué. Lung metastasis genes couple breast tumor size and metastatic spread. *Proceedings of the National Academy of Sciences*, 104(16):6740–6745, 2007.

T. Mitchell and J. Beauchamp. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83:1023–1036, 1988.

B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Advances in neural information processing systems*, pages 915–922, 2005.

S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian and l1 approaches for sparse unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 751–758, 2012.

E. Moreno, J. Girón, and G. Casella. Posterior model consistency in variable selection as the model dimension grows. *Statistical Science*, 30(2):228–241, 2015.

K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *Technical report*, 2007.

T. B. Murphy, N. Dean, and A. E. Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics*, 4(1):396, 2010.

I. Murray and Z. Ghahramani. A note on the evidence and Bayesian Occam's razor. *Technical report*, 2005.

S. Nakajima, M. Sugiyama, and D. Babacan. On Bayesian PCA: Automatic dimensionality selection and analytic solution. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 497–504, 2011.

N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

B. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, New York, 1996.

R. M. Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

A. Njato Randriamanamihaga, E. Côme, L. Oukhellou, and G. Govaert. Clustering the vélib'dynamic origin/destination flows using a family of poisson mixture models. *Neurocomputing*, 2014.

H. Ogata. A numerical integration formula based on the bessel functions. *Publications of the Research Institute for Mathematical Sciences*, 41(4):949–970, 2005.

A. O'Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–138, 1995.

R. O'Hara and M. Sillanpää. A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.

A. Oppenheim. Inequalities connected with definite hermitian forms. *Journal of the London Mathematical Society*, 1(2):114–119, 1930.

T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

D. Passemier, Z. Li, and J. Yao. On estimation of the noise variance in high dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):51–67, 2017.

D. K. Pauler, J. C. Wakefield, and R. E. Kass. Bayes factors and approximations for variance component models. *Journal of the American Statistical Association*, 94(448):1242–1253, 1999.

K. Pearson. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(11):559–572, 1901.

S. Petrone, J. Rousseau, and C. Scricciolo. Bayes and empirical bayes: do they merge? *Biometrika*, 2014.

J. Piironen and A. Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2016.

M. Plummer. Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9(3): 523–539, 2008.

B. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis*, 100(9):2065–2082, 2009.

P. Pudlo, J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

Y. Qiu and J. Mei. *RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems*, 2016. URL `https://CRAN.R-project.org/package=RSpectra`. R package version 0.12-0.

A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, pages 111–163, 1995.

A. E. Raftery. Bayes factors and BIC: Comment on a critique of the Bayesian information criterion for model selection? *Sociological Methods & Research*, 27(3):411–427, 1999.

A. E. Raftery and Y. Zheng. Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*, 98(464):931–938, 2003.

A. E. Raftery, D. Madigan, and C. T. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). *Bayesian statistics*, 5:323–349, 1996.

A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

C. E. Rasmussen and Z. Ghahramani. Occam's razor. *Advances in neural information processing systems*, pages 294–300, 2001.

P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

M. Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2007.

C. Robert and G. Casella. *Monte Carlo statistical methods*, volume 319. Springer, 2004.

C. P. Robert. A note on Jeffreys-Lindley paradox. *Statistica Sinica*, 3(2):601–608, 1993.

C. P. Robert. On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2):216–232, 2014.

C. P. Robert. The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72:33–37, 2016.

C. P. Robert and D. Wraith. Computational methods for Bayesian model choice. In *The 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. AIP Conference Proceedings*, volume 1193, pages 251–262. AIP, 2009.

C. P. Robert, N. Chopin, and J. Rousseau. Harold Jeffreys's theory of probability revisited. *Statistical Science*, pages 141–172, 2009.

P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied statistics*, pages 257–265, 1976.

V. Ročková and E. George. EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, just-accepted, 2013.

G.-C. Rota. The number of partitions of a set. *The American Mathematical Monthly*, 71 (5):498–504, 1964.

S. Roweis. EM algorithms for PCA and SPCA. *Advances in neural information processing systems*, pages 626–632, 1998.

L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

R. Schaback and Z. Wu. Operators on radial functions. *Journal of computational and applied mathematics*, 73(1):257–270, 1996.

T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, and T. L. Casavant. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

M. Schroeder, B. Haibe-Kains, A. Culhane, C. Sotiriou, G. Bontempi, and J. Quackenbush. *breastCancerVDX: Gene expression datasets published by Wang et al. [2005] and Minn et al. [2007] (VDX)*, 2011. URL http://compbio.dfci.harvard.edu/. R package version 1.8.0.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

D. W. Scott and J. R. Thompson. Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proceedings of the fifteenth symposium on the interface*, volume 528, pages 173–179. North-Holland, Amsterdam, 1983.

J. Scott and J. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

M. Seeger. *Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations*. PhD thesis, University of Edinburgh, 2003.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9. ACM, 1997.

H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008.

C. D. Sigg and J. M. Buhmann. Expectation-maximization for sparse and non-negative PCA. In *Proceedings of the 25th international conference on Machine learning*, pages 960–967. ACM, 2008.

B. Silverman and J. Ramsay. *Functional Data Analysis*. Springer, 2005.

L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, 1987.

S. A. Sisson. Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, 100(471):1077–1089, 2005.

T. Skeggs. Special report, visitor figures 2013. *The Art Newspaper*, 23(256), April 2014.

J. Skilling. Nested sampling for general Bayesian computation. *Bayesian analysis*, 1(4): 833–859, 2006.

P. Sobczyk, M. Bogdan, and J. Josse. Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood. *Journal of Computational and Graphical Statistics*, in press, 2017.

A. Spanos. Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science*, 80(1):73–93, 2013.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. The deviance information criterion: 12 years on (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493, 2014.

J. Stoehr. A review on statistical inference methods for discrete Markov random fields. *arXiv preprint arXiv:1704.03331*, 2017.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

A. E. Teschendorff, M. Journée, P. A. Absil, R. Sepulchre, and C. Caldas. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput Biol*, 3(8):e161, 2007.

C. M. Theobald. An inequality with application to multivariate analysis. *Biometrika*, 62 (2):461–466, 1975. ISSN 00063444.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)*, 58(1):267–288, 1996.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

D. Tran, R. Ranganath, and D. M. Blei. Deep and hierarchical implicit models. *arXiv preprint arXiv:1702.08896*, 2017a.

M.-N. Tran, D. J. Nott, and R. Kohn. Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, (in press), 2017b.

M. O. Ulfarsson and V. Solo. Sparse variable PCA using geodesic steepest descent. *Signal Processing, IEEE Transactions on*, 56(12):5823–5832, 2008a.

M. O. Ulfarsson and V. Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Transactions on Signal Processing*, 56(12):5804–5816, 2008b.

M. O. Ulfarsson and V. Solo. Vector l0 sparse variable PCA. *Signal Processing, IEEE Transactions on*, 59(5):1949–1958, 2011.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

A. Vehtari and J. Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

C. Villa and S. Walker. On the mathematics of the Jeffreys–Lindley paradox. *Communications in Statistics-Theory and Methods*, (in press), 2017.

V. Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

S. G. Walker. Modern Bayesian asymptotics. *Statistical Science*, pages 111–117, 2004.

Y. Wang, J. G. M. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, M. Talantov, D.and Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. J. J. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

S. Watanabe. Algebraic analysis for singular statistical estimation. In *International Conference on Algorithmic Learning Theory*, pages 39–50. Springer, 1999.

S. Watanabe. *Algebraic geometry and statistical learning theory*. Cambridge University Press, 2009.

S. Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897, 2013.

S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda. Application of variational Bayesian approach to speech recognition. In *Advances in Neural Information Processing Systems*, pages 1261–1268, 2003.

D. L. Weakliem. A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3):359–397, 1999.

M. D. Weinberg. Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution. *Bayesian Analysis*, 7(3):737–770, 2012.

S. Weisberg. *Applied Linear Regression*. Wiley, New York, 1980.

D. Wipf and S. Nagarajan. A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632, 2008.

D. Wipf and S. Nagarajan. A unified bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966, 2009.

D. P. Wipf, B. D. Rao, and S. Nagarajan. Latent variable Bayesian models for promoting sparsity. *Information Theory, IEEE Transactions on*, 57(9):6236–6255, 2011.

J. Wishart and M. S. Bartlett. The distribution of second order moment statistics in a normal system. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(4): 455–459, 1932.

D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

S. Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978.

D. Wrinch and H. Jeffreys. On some aspects of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 38(228):715–731, 1919.

D. Wrinch and H. Jeffreys. On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 42(249): 369–390, 1921.

D. Wrinch and H. Jeffreys. On certain fundamental principles of scientific inquiry (second paper). *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 45(266):368–374, 1923.

C. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

S. Xiaoshuang, L. Zhihui, G. Zhenhua, W. Minghua, Z. Cairong, and K. Heng. Sparse principal component analysis via joint $L_{2,1}$-norm penalty. In *AI 2013: Advances in Artificial Intelligence*, pages 148–159. Springer, 2013.

J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487, 2016.

L. Xu and M. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

Y. Yang, M. J. Wainwright, and M. I. Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.

T.-J. Yen. A majorization-minimization approach to variable selection using spike and slab priors. *The Annals of Statistics*, pages 1748–1775, 2011.

L. Yengo, J. Jacques, and C. Biernacki. Variable clustering in high dimensional linear regression models. *Journal de la Société Française de Statistique*, 155(2):38–56, 2014.

L. Yengo, J. Jacques, C. Biernacki, and M. Canouil. Variable clustering in high-dimensional linear regression: The R package clere. *The R Journal*, 8(1):92–106, 2016.

G. Yu and Q.-Y. He. Reactomepa: an R/Bioconductor package for Reactome pathway analysis and visualization. *Molecular BioSystems*, 2016.

L. Yu, R. R. Snapp, T. Ruiz, and M. Radermacher. Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. *Journal of structural biology*, 171(1):18–30, 2010.

Y. Yu. On normal variance–mean mixtures. *Statistics & Probability Letters*, 121:45–50, 2017.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986.

A. Zellner. Keep it sophisticatedly simple. In A. Zellner, H. A. Keuzenkamp, and M. McAleer, editors, *Simplicity, inference and modelling: Keeping it sophisticatedly simple*, chapter 14, pages 242–262. Cambridge University Press, 2001.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of the fifth International Conference on Learning Representations*, 2017.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Y. Zhang and L. El Ghaoui. Large-scale sparse principal component analysis with application to text data. In *Advances in Neural Information Processing Systems*, pages 532–539, 2011.

Y. Zhang, A. d'Aspremont, and L. El Ghaoui. Sparse PCA: Convex relaxations, algorithms and applications. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 915–940. Springer, 2012.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.