Sparse meta-Gaussian information bottleneck

Mélanie Rey MELANIE.REY@UNIBAS.CH
University of Basel, Basel, Switzerland

Thomas J. Fuchs FUCHS@CALTECH.EDU
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA

Volker Roth VOLKER.ROTH@UNIBAS.CH
University of Basel, Basel, Switzerland

Abstract

We present a new sparse compression technique based on the information bottleneck (IB) principle, which takes into account side information. This is achieved by introducing a sparse variant of IB which preserves the information in only a few selected dimensions of the original data through compression. By assuming a Gaussian copula we can capture arbitrary non-Gaussian margins, continuous or discrete. We apply our model to select a sparse number of biomarkers relevant to the evolution of malignant melanoma and show that our sparse selection provides reliable predictors.

1. Introduction

Dimensionality reduction is an important domain of research for which a large variety of techniques have been developed. Given observations of a multivariate random variable X, the aim is to construct a new representation of the data with reduced dimensionality. Amongst the most prominent methods, Principal Component Analysis (PCA) and Canonical Component Analysis (CCA) have been extensively studied and extended. Other classical techniques include Compressed Sensing, Factor Analysis and methods for manifold modeling. Every dimensionality reduction technique first needs to define what is important in the data and should therefore be preserved, e.g. PCA aims at preserving the variance. In this respect, dimensionality reduction is related to the data compression problem where the question of defining what is relevant is implicitly answered by the choice of a distortion function, i.e. a function evaluating the loss incurred by compression. An interesting alternative to distortion functions is given by the Information bottleneck method (IB) (Tishby et al., 1999). Instead of evaluating the distortion between the compression T and original data X, IB introduces a relevance variable Y and the aim becomes to compress the data while retaining most information about Y. An important advantage of IB is that compression and information are solely expressed in information-theoretic quantities. We seek T such that $I(X;T)$ is small, meaning high compression, and $I(Y;T)$ is large, meaning high informativeness. While some choices of distortion function or dependence measure (e.g. correlation) are not suitable for certain types of data, mutual information is a very general and theoretically well-founded dependence measure (Joe, 1989). In this paper, we present a new sparse compression technique based on the information bottleneck principle, i.e. we perform feature selection with relevance information. This is achieved by introducing a sparse variant of IB in which T is built using only a few selected dimensions of the original data. Efficient IB algorithms were limited to discrete data before the introduction of Gaussian IB (GIB) (Chechik et al., 2005). GIB considers the case of Gaussian variables for which it provides an insightful and practical analytical solution. Recently, meta-Gaussian IB (MGIB) (Rey & Roth, 2012) has generalised GIB to the continuous meta-Gaussian distributions i.e. distributions with a Gaussian copula and arbitrary continuous margins. However, the compression achieved by existing IB methods is usually not sparse and, therefore cannot be used for feature selection. Our model is an extension of MGIB, we impose sparsity on the projection obtained through MGIB and thereby select a sparse set of features. Our method shares some similarities with sparse regression techniques like Lasso (Tibshirani, 1996) but is based on different, less restrictive assumptions and achieves sparsity without imposing any norm penalty. To our knowledge, there exists no IB method to accommodate mixed distributions, i.e. distributions with continuous and

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

discrete margins, without resorting to discretisation of the continuous dimensions. Besides introducing sparsity, another contribution of this paper is to extend MGIB to mixed data by considering discrete margins as the result of a discretisation process of hidden continuous variables. This extension is motivated by the high prevalence of mixed data in many domains where feature selection is sought, especially in the medical and biological fields (de Leon & Chough, 2013).

2. Information Bottleneck (IB)

General IB. Consider two random vectors $X = (X_1, \dots, X_p)$ and $Y = (Y_1, \dots, Y_q)$ with joint distribution p_{XY}. To obtain a compression T of X that is most informative about Y, we solve the following variational problem

$$\min_{p_{T|X}} \mathcal{L} \mid \mathcal{L} \equiv I(X;T) - \beta I(T;Y), \tag{1}$$

where the Lagrange parameter $\beta > 0$ determines the trade-off between compression of X and preservation of information about Y. Since T is conditionally independent of Y given X, it is fully characterised by its joint distribution with X, denoted by p_{XT}. No analytical solution is available for the general problem defined by (1). When X and Y are discrete, p_{XT} can be obtained iteratively using the generalised Blahut-Arimoto algorithm, and T is a discrete variable defining soft clusters of X.

Gaussian IB. GIB considers the special case of jointly Gaussian variables (X, Y):

$$(X,Y) \sim \mathcal{N}\left(0_{p+q}, \Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{yx}^T \\ \Sigma_{yx} & \Sigma_y \end{pmatrix}\right), \tag{2}$$

where 0_{p+q} is the zero vector of length $p+q$. As shown in Chechik et al. (2005) an optimal compression T is obtained with a noisy linear transformation of X: $T = AX + \xi$, where $A \in \mathbb{R}^{p \times p}, \xi \sim \mathcal{N}(0_p, \Sigma_\xi)$. The noise term ξ is independent of X, and as such will not carry any information about X. We can therefore assume that the noise components are independent from each other and, without prior information, that these components have equal variance. As shown in Chechik et al. (2005), Σ_ξ can be set to the identity matrix I w.l.o.g, and the minimisation problem (1) is thus reduced to:

$$\min_{A} \mathcal{L} \mid \mathcal{L} \equiv I(X; AX + \xi) - \beta I(AX + \xi; Y). \tag{3}$$

The optimisation problem (3) has the major advantage that an analytical solution can be calculated: the optimal compression is a Gaussian variable $T \sim \mathcal{N}(0_p, \Sigma_t)$ with $\Sigma_t = A\Sigma_x A^T + I$. The projection matrix A has a particular structure where the number of non-zero rows depends on β:

$$A = \begin{cases} \left[0^T; \dots; 0^T\right] & \text{if } 0 \le \beta \le \beta_1^c \\ \left[\alpha_1 v_1^T; 0^T; \dots, 0^T\right] & \text{if } \beta_1^c \le \beta \le \beta_2^c \\ \vdots \end{cases}, \tag{4}$$

where 0^T is a p-dimensional row vector and semicolons separate rows of A. Here $v_1^T, \dots, v_p^T$ are the left eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ sorted by their corresponding increasing eigenvalues $\lambda_1, \dots, \lambda_p$. There are p critical β values $\beta_i^c = (1 - \lambda_i)^{-1}$, and the α_i coefficients are defined by $\alpha_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$ with $r_i = v_i^T \Sigma_x v_i$. In summary, the projection matrix A is a combination of weighted eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ and the number of selected eigenvectors depends on the parameter β.

Meta-Gaussian IB. Consider random vectors X, Y with a Gaussian copula and arbitrary margins, i.e. their cumulative distribution function (cdf) $F_{XY}(x, y)$ has the form

$$C_P(F_{X_1}(x_1), \dots, F_{X_p}(x_p), F_{Y_1}(y_1), \dots, F_{Y_q}(y_q)), \tag{5}$$

where F_{X_i}, F_{Y_i} are the marginal cdfs of X, Y and C_P is a Gaussian copula parametrized by a correlation matrix $P = \begin{pmatrix} P_x & P_{yx}^T \\ P_{yx} & P_y \end{pmatrix}$. In Rey & Roth (2012) it was pointed out that since the mutual information between continuous variables depends only on their copula, the IB problem is actually independent of the marginal distributions. The main idea in MGIB is that the IB problem for meta-Gaussian data can be solved by applying GIB in the space of the normal scores $\tilde{X} = (\Phi^{-1} \circ F_{X_1}(X_1), \dots, \Phi^{-1} \circ F_{X_p}(X_p))$ and $\tilde{Y}$, where Φ denotes the standard univariate Gaussian cdf. An optimal projection T is obtained by first calculating $(\tilde{X}, \tilde{Y})$ and then computing $T = A\tilde{X} + \xi$ as in GIB. A is entirely determined by the covariance matrix of $(\tilde{X}, \tilde{Y})$ which also equals the correlation matrix P parametrizing the Gaussian copula C_P (the normal scores have unit variance by definition). In summary, we can apply GIB to meta-Gaussian data as long as we can make inference on two elements: P and the underlying Gaussian variables, which in the continuous case are simply the normal scores $(\tilde{X}, \tilde{Y})$.

3. Sparse IB

We first assume that the underlying Gaussian variables of our copula model and P are known. Inference will be discussed at the end of section 3. To achieve sparse compression we consider the MGIB model but restrict A to the class of diagonal matrices. If we denote by $a = (a_1, \dots, a_p) \in \mathbb{R}^p$ the vector of the diagonal entries of A, the resulting projection is the vector $AX = (a_1 X_1, \dots, a_p X_p)$. By varying the Lagrange parameter of the minimisation problem, the

vector a becomes sparse and thereby selects only a few dimensions of X. As for the case of a full projection matrix we can assume that the noise components are uncorrelated with unit variance i.e. $\Sigma_\xi = I$. Our optimisation problem is simplified by two convenient properties:

1. For any symmetric positive definite matrix B and diagonal matrix D, $\log|DBD + I| = \log|BD^2 + I|$ depends only on the squared entries D_{ii}^2. It is therefore sufficient to consider the space of positive diagonal matrices in $\mathbb{R}^p$, denoted by $\mathbb{D}_p^+$. In the following we use the notation $A := D^2$.
2. $\log|BA+I|$ is *concave* in A and strictly monotone increasing in every component A_{ii}. Concavity directly follows from the concavity of $\log|B|$ and the fact that $A \mapsto BA + I$ is an affine function.

The minimisation problem (3) can be rewritten for some $\kappa' \geq 0$ as (see also (Chechik et al., 2005)):

$$\min_{A:A\in\mathbb{D}_p^+} \underbrace{\log|P_xA+I|}_{2I(X;T)} \quad \text{s.t.} \qquad (6)$$
$$\underbrace{\log|P_xA+I| - \log|QA+I|}_{2I(T;Y)} \geq \kappa',$$

where $Q = P_x - P_{xy}P_y^{-1}P_{xy}^T$ is the conditional covariance matrix of X given Y. The corresponding Lagrangian is

$$\mathcal{L}'(A,\beta) = \log|P_xA+I| - \sum_{j=1}^{p}\eta_jA_{jj} \qquad (7)$$
$$-\beta(\log|P_xA+I| - \log|QA+I| - \kappa'),$$

where η_j is the Lagrange parameter for the j-th non-negativity constraint. Nontrivial solutions for the original IB with full matrix A exist only for $\beta > 1$, see (Chechik et al., 2005). We transform the IB problem into an equivalent form that exchanges objective function and constraint. Introducing a new Lagrange parameter $\lambda = (\beta-1)/\beta$, $0 < \lambda < 1$, and dividing (7) by β, we find the Lagrangian

$$\mathcal{L}(A,\lambda) = \log|QA+I| - \sum_{j=1}^{p}\epsilon_jA_{jj} + \lambda(\kappa - \log|P_xA+I|),$$

with $\epsilon_j = \frac{\eta_j}{\beta}, \kappa \geq 0$. The corresponding minimisation problem can then be rewritten as

$$\min_{A:A\in\mathbb{D}_p^+} \underbrace{\log|QA+I|}_{:=f(a)} \quad \text{s.t.} \quad \underbrace{\log|P_xA+I|}_{:=g(a)} \geq \kappa, \qquad (8)$$

which amounts to minimising a concave function $f(a)$ over a convex set $\{b \in \mathbb{R}^p | g(b) \geq \kappa\}$. Thus, the global minimum is attained at the boundary $g(a) = \kappa$. Note that for $\kappa = 0$ the constraint is always satisfied and there is a unique minimum at $a = 0$. While we cannot characterise all stationary points of the Lagrangian problem as formulated in (7), the minimisation problem (8) has a particular form (concave function over a convex set) for which algorithms with guaranteed convergence to a globally optimal solution exist (Benson & Horst, 1991). A Matlab code for this method is available online[1]. However, this algorithm is not efficient in higher dimensions and we therefore propose a log barrier interior point method detailed later in Algorithm 1.

The solution set S of optimisation problem (8) is defined as the set of points a^* in the non-negative orthant of $\mathbb{R}^p$ which are global minima for a certain value of κ, i.e. $S = \{a^* \in \mathbb{R}_+^p \text{ s.t. } a^* \text{ is a solution of (8) for some } \kappa \geq 0\}$. In the following theorem, we show that for an interval $[0, \kappa_2^c]$, S is a curve parametrized by κ, meaning that to every κ in this interval corresponds a unique point in S. In the following we assume that P_x and P_y are random matrices of maximal rank and write $\Phi := Q^{-1}$, $\Psi := P_x^{-1}$. We denote points in S by $a^* = (a_1^*, \ldots, a_p^*)$.

Theorem 3.1. *With probability one, there exist $\kappa_2^c > \kappa_1^c > 0$ such that:*

1. *If $\kappa \in [0, \kappa_1^c]$ then*

$$a_i^* = \begin{cases} e^\kappa - 1 & \text{if } i = \operatorname{argmin}_j(Q_{jj}) =: i_f, \\ 0 & \text{else.} \end{cases}$$

2. *If $\kappa \in [\kappa_1^c, \kappa_2^c]$ then*

$$a_i^* = \begin{cases} G(\kappa;\Psi,\Phi) & \text{if } i = i_s, \\ c_1a_{i_s}^* + c_0 & \text{if } i = i_f, \\ 0 & \text{else} \end{cases}$$

where $c_1 = \frac{|\Psi|-\Phi_{22}}{|\Psi|-\Phi_{11}}$, $c_0 = \frac{|\Psi|(\Phi_{11}-\Phi_{22})}{|\Psi|-\Phi_{11}}$ and

$$G(\kappa;\Psi,\Phi) = -\frac{|\Psi|(1+c_1)+c_0}{2c_1} + \left(\frac{(|\Psi|(1+c_1)+c_0)^2}{4c_1^2} - \frac{|\Psi|(c_0+1-e^\kappa)}{c_1}\right)^{0.5},$$

The value of i_s can be determined by searching over the $p-1$ possible combinations (i_f, i) and choosing the dimension i_s which gives the minimal value of $f(a)$.

We use the term critical values to designate κ_1^c, κ_2^c. We call i_f the most informative dimension of X and i_s the second most informative dimension. Theorem 3.1 tells us that, for small enough κ values, the solution set is a curve parametrized by κ which starts at the point zero, runs along the i_f-axis until c_0 is reached, and then takes the form of a straight line with slope c_1, see Figure 1. We therefore call S the solution path. We prove this result in three steps given by Lemma 3.1, Lemma 3.2 and Lemma 3.3.

[1] http://www.mathworks.com/matlabcentral/fileexchange/36247-function-for-global-minimisation-of-a-concave-function

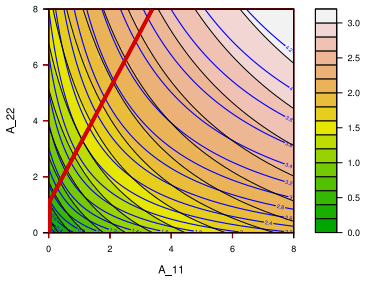

Figure 1. Objective function f (filled contours, colour coded) and constraint g (thick blue contour lines). The solution path is depicted by the red line. Above the critical constraint value (in this example the $\kappa_1^c \approx 0.8$ contour line), all solutions lie on the line $a_2^* = c_1 a_1^* + c_0$ with positive slope c_1, below the critical value, the solution path is a vertical line at the origin, i.e. solutions are sparse ($a_1^* = 0$).

Lemma 3.1 (Most informative dimension)**.** *The most informative dimension is the dimension i_f with the smallest corresponding entry in Q, i.e. $i_f = \mathrm{argmin}_i(Q_{ii})$. Moreover, when only one component i_f of a is non-zero, $a_{i_f}^* = e^\kappa - 1$.*

Proof. For every possible choice of $i = 1, \ldots, p$, the value of a_i is uniquely determined by the constraint:

$$\kappa = g(a_i) = \log((P_x)_{ii} a_i + 1) = \log(a_i + 1), \qquad (9)$$

which implies that $a_i = e^\kappa - 1$. The optimal i_f is then determined by the minimum value of $f(a_i) = \log(Q_{ii} a_i + 1) = \log(Q_{ii}(e^\kappa - 1) + 1)$. □

Lemma 3.2 bellow provides an analytical solution in the case $p = 2$. In the following we denote the partial derivatives of a real function f of a by $\frac{\partial f}{\partial a_i}(a) = f_{a_i}$.

Lemma 3.2 (2-dimensional case)**.** *Assume $p = 2$ and, w.l.o.g. that the most informative dimension is $i_f = 2$. Let $c_1, c_0, G(\kappa; \Psi, \Phi)$ be defined as in Theorem 3.1. The first critical value is $\kappa_1^c = \log(c_0 + 1)$, and with probability one, $\kappa_1^c > 0$. Moreover, the optimal a^* is given*

1. *for every $\kappa \in [\kappa_1^c, \infty)$ by*

$$a_2^* = c_1 a_1^* + c_0, \qquad (10)$$
$$a_1^* = G(\kappa; \Psi, \Phi), \qquad (11)$$

2. *for every $\kappa \in [0, \kappa_1^c]$ by $a_2^* = e^\kappa - 1, a_1^* = 0$.*

Proof. We first determine the set of stationary points with strictly positive components. When $a_i > 0, i = 1, 2$, the non-negativity constraints are inactive and $\epsilon_i = 0, i = 1, 2$. Stationary points are characterised by a vanishing Lagrangian gradient $\nabla \mathcal{L} = 0$, which here means that $\nabla f(a) = \lambda \nabla g(a)$. When $p = 2$, this proportionality condition is equivalent to the orthogonality condition $-g_{a_2} f_{a_1} + g_{a_1} f_{a_2} = 0$. Adding the constraint $g(a) = \kappa$ leads to a system of 2 equations in 3 variables (a_1, a_2 and κ) which implicitly defines the set of all stationary points with strictly positive components:

$$\begin{cases} \tilde{H}_{12} & : \quad f_{a_2} g_{a_1} - f_{a_1} g_{a_2} = 0 \\ \tilde{H}_0 & : \quad g(a) - \kappa = 0 \end{cases} \qquad (12)$$

To solve the above system we first need to to compute the partial derivatives f_{a_i}, g_{a_i}. Rewriting $f(a) = \log(|Q||Q^{-1} + A|) = \log|Q| + \log|\Phi + A|$ and $g(a) = \log|P_x| + \log|\Psi + A|$, we directly obtain

$$f_{a_j} = \frac{|\Phi + A|^{[-j]}}{|\Phi + A|}, \quad g_{a_j} = \frac{|\Psi + A|^{[-j]}}{|\Psi + A|}. \qquad (13)$$

From expressions 13 we can see that $\tilde{H}_{12}$ can advantageously be replaced by

$$H_{12} : |\Phi + A||\Psi + A|(f_{a_2} g_{a_1} - f_{a_1} g_{a_2}) = 0. \qquad (14)$$

Further, $\tilde{H}_0$ can be replaced by $H_0 : |P_x||\Psi + A| - e^\kappa = 0$. We solve system (12) in two steps. First, using H_{12} we obtain an expression for a_2 as a function of a_1, leading to equation (10). Second, we replace a_2 in H_0 by expression (10), leading to equation (11). Finally, we can compute the critical value κ_1^c by solving $a_1^* = G(\kappa; \Psi, \Phi) = 0$ for κ. Calculations are straightforward and can be found in the supplementary material. Solutions for $\kappa \leq \kappa_1^c$ are given by Lemma 3.1. The probability of c_0 being equal to zero is null, i.e. $\mathbb{P}\{c_0 = 0\} = 0$, since $\mathbb{P}\{Q_{11} = Q_{22}\} = 0$ for random correlation matrices P_x and P_y. This implies that $\kappa_1^c > 0$ with probability one and there always exist values of κ for which the 1-dimensional solution is optimal. □

We come back to the general p-dimensional case with the following result.

Lemma 3.3 (General p-dimensional case)**.** *In the p-dimensional problem, with probability one, the following statements hold:*

1. *There is an non-empty interval $[\kappa_1^c, \kappa_3^c]$ for which the global optimum is attained with two active dimensions.*
2. *There is an non-empty interval $[\kappa_1^c, \kappa_2^c]$ for which the global optimum is attained in a fixed 2-dimensional subspace.*

Proof. Assume w.l.o.g. that the most informative dimension is $i_f = 1$. As seen in the proof of Lemma 3.2, stationary points with strictly positive components are characterised by $\nabla f(a) = \lambda \nabla g(a)$. A new dimension $j \neq 1$ becomes active when $f_{a_j} = \lambda g_{a_j}$.

We prove the first statement by contradiction. Assume that there exists no global optimum having exactly two non-zero components, this means that when varying κ, solutions

"jump" from one to (at least) three non-zero components[2]. In particular, at the jumping point $a^* = (a_1^*, 0, \dots, 0)$ two equations of the form

$$\begin{cases} H_{1m} & : \quad f_{a_m} g_{a_1} - f_{a_1} g_{a_m} = 0 \\ H_{1s} & : \quad f_{a_s} g_{a_1} - f_{a_1} g_{a_s} = 0 \end{cases} \tag{15}$$

for $m, s \neq 1$ with $m \neq s$ must be fulfilled. When only the component a_1^* is non-zero, both equations in (15) are linear in a_1^*. We can therefore eliminate a_1^* from the above system and are left with one equation in P_x and Q only.

$$\begin{aligned} &(P_x)_{1m}^2 Q_{11}(Q_{11} - Q_{ss}) + (P_x)_{1s}^2 Q_{11}(-Q_{11} + Q_{mm}) + \\ &[Q_{11}(Q_{1s}^2 - Q_{11}Q_{ss}) + Q_{11}(-Q_{1m}^2 - Q_{mm} + Q_{11}Q_{mm}) \\ &+ (Q_{1m}^2 Q_{ss} - Q_{1s}^2 Q_{mm})] = 0 \end{aligned} \tag{16}$$

However, since P_x, P_y are random matrices, the probability that equation (16) holds is null. We can therefore conclude that with probability one there is only one other dimension $m \neq 1$ for which $f_{a_m} = \lambda g_{a_m}$. This implies that there exist $\kappa_3^c > \kappa_1^c \geq 0$ such that exactly two dimensions are active.

We now prove the second statement. We restrict ourselves w.l.o.g to the interval $[0, \kappa_3^c]$ where the global optimum is attained only with one or two dimensions. The most informative dimension i_f being fixed, there are $p-1$ different possible 2-dimensional subspaces. We can apply Lemma 3.2 to each subspace separately to obtain $p-1$ different values of the first critical κ: $0 < \kappa_{1,1}^c < \dots < \kappa_{1,p}^c$. Since each $\kappa_{1,i}^c$ is a function of different entries in Ψ and Φ (see Lemma 3.2), and recalling that P_x, P_y are random matrices, we can see that all values $\kappa_{1,1}^c, \dots, \kappa_{1,p}^c$ are indeed distinct. The solution path S leaves the i_f axis when the first solution with two non-zero components becomes optimal, i.e. when $\kappa = \kappa_{1,1}^c$ and we can finally set $\kappa_1^c := \kappa_{1,1}^c$. In the interval $[\kappa_{1,1}^c, \kappa_{1,2}^c[$ only one 2-dimensional subspace can contain global optima, namely the subspace having first critical value $\kappa_{1,1}^c$. We can therefore set $\kappa_2^c := \kappa_{1,2}^c$.

□

We solve optimisation problem (8) using a log barrier interior point method detailed in Algorithm 1. Algorithm 1 starts by minimising f for a large value of the constraint κ such that all dimensions are selected, and then successively decreases κ until finally a unique dimension remains active. Even if we cannot prove that the solution path obtained with Algorithm 1 connects only global minima, we can verify that it reaches the globally optimal 2-dimensional subspace. This was indeed in the case in all simulations conducted. We also propose in the supplementary material an additional method to check that while running Algorithm 1 the solution path does not have any bifurcation.

[2]The case of a jump from $a = (0, \dots, 0)$ to a least three active dimensions can similarly be excluded.

Algorithm 1 Optimisation of sparse MGIB

1. Denote the entries of A by $a \in \mathbb{R}^p$ and fix the set of κ values: $\kappa \in \{\kappa_0 > \kappa_1 > \dots > \kappa_m\}$;
2. Initialisation step with $\kappa = \kappa_0$:
   compute $a^{\max} \in \mathbb{R}^p$: $a_j^{\max} = (e^\kappa - 1)/(P_x)_{jj}$;
   set $a^{\mathrm{m}} := 1/\sum_j (a_j^{\max})^{-1}$;
   compute $\lambda_1, \dots, \lambda_p$ the eigenvalues of P_x;
   compute $c = \operatorname{argmin}_{[0, a^{\mathrm{m}}]} f_1(v), v \in \mathbb{R}$ with
   $f_1(v) = \left[\sum_j \log(\lambda_j) + \sum_j \log(\lambda_j^{-1} + v) - \kappa\right]^2$;
   set $a = (c + \delta, \dots, c + \delta)$ for a small $\delta > 0$;
3. Optimisation for $\kappa \geq \kappa_0$:

for $\kappa \in \{\kappa_0, \dots, \kappa_m\}$ **do**
  for $\epsilon \to 0$ **do**
    Set $a^* = \operatorname{argmin} f_2(w)$ where $w \in \mathbb{R}^p$, W is the diagonal matrix with elements w and
    $f_2(w) = \log|P_{x|y} W + I| -$
    $\epsilon \log\left[\log|P_x W + I| - \kappa\right] - \epsilon \sum_j \log(w_j)$;
  end for
  Exclude the dimensions corresponding to zero elements in a^* from the minimisation;
end for

Inference for mixed continuous-discrete data. To make our model applicable to mixed data we use Gaussian hidden variables and, as in MGIB, apply GIB to these underlying variables. This approach is based on the fact that every ordinal categorical variable can be assumed to be a quantised version of an underlying continuous one (Joe, 1989). Our model can be applied to data with any combination of continuous and ordinal discrete margins. Copula models have mainly been used with continuous margins since for any continuous multivariate cdf $F = (F_1, \dots, F_n)$ Sklar's theorem (Sklar, 1959) ensures the existence and the uniqueness of the copula C such that $F(v_1, \dots, v_n) = C(F_1(v_1), \dots, F_n(v_n))$. However, as explained in Genest & Nešlehová (2007), the copula construction $C(F_1(.), \dots, F_n(.))$ stills leads to a valid cdf when all or some of the margins are discrete. The main difficulty in copula modeling with discrete margins arises from the fact that uniqueness of the copula is guaranteed only on the range of the margins and traditional estimation techniques face an unidentifiability problem (Genest & Nešlehová, 2007). Despite this unidentifiability issue, efficient methods for copula estimation have recently been developed (Pitt et al., 2006), (Smith & Khaled, 2012). We follow the Bayesian semiparametric approach for Gaussian copula introduced in Hoff (2007) and consider the following model:

$$(\bar{X}_1, \dots, \bar{X}_p, \bar{Y}_1, \dots, \bar{Y}_q) \sim \mathcal{N}(0, P), \tag{17}$$

$$X_j = F_{X_j}^{-1}(\Phi(\bar{X}_j)), \quad Y_l = F_{Y_l}^{-1}(\Phi(\bar{Y}_l)), \tag{18}$$

for $j = 1, \ldots, p, \quad l = 1, \ldots, q$. where $F_{X_j}^{-1}$, $F_{Y_l}^{-1}$ are the generalised inverse of arbitrary, continuous or discrete, cdfs. We assume a parametric form for the copula, namely Gaussian, but not for the margins which are treated non-parametrically. Equation (18) imply that $X_j \sim F_{X_j}$, $Y_l \sim F_{Y_l}$. Unlike in the continuous case, the underlying Gaussian variables $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_p)$, $\bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_q)$ are not of the form $(\Phi^{-1} \circ F_{X_1}(X_1), \ldots, \Phi^{-1} \circ F_{X_p}(X_p))$. When some margins are discrete, estimates based on the empirical marginal distributions (Liu et al., 2009) or on the rank correlation cannot be used. To make inference on P and $(\bar{X}, \bar{Y})$ we use the marginal likelihood method introduced in Hoff (2007). Denote the observed data matrix by $Z = (z_{ij})$, where the i^{th} row $z_{i*} = (x_{i1}, \ldots, x_{ip}, y_{i1}, \ldots, y_{iq})$ is the i^{th} observation of (X, Y), and the corresponding unobserved realisations of $(\bar{X}, \bar{Y})$ by $\bar{Z}$. Even without assuming any knowledge about the margins, the observed data Z provide some information about $\bar{Z}$. Since the marginal cdfs are non-decreasing, the following inequality holds: $z_{mj} < z_{nj} \Rightarrow \bar{z}_{mj} < \bar{z}_{nj}, \forall m, n, j$, and $\bar{Z}$ must lie in the set $\mathcal{D}$:

$$\bar{Z} \in \Big\{ G = (g_{ij}) \in \mathbb{R}^{n \times (p+q)} \,|\, \max\{g_{kj} : z_{kj} < z_{ij}\} < g_{ij} < \min\{g_{kj} : z_{ij} < z_{kj}\} \Big\} =: \mathcal{D}. \tag{19}$$

The probability of the observed data can then be written as:

$$\begin{aligned} p(Z|P, F_X, F_Y) &= p(Z, \bar{Z} \in \mathcal{D}|P, F_X, F_Y) \\ &= \Pr(\bar{Z} \in \mathcal{D}|P)\; p(Z|\bar{Z} \in \mathcal{D}, P, F_X, F_Y), \end{aligned}$$

where $F_X = \{F_{X_1}, \ldots, F_{X_p}\}, F_Y = \{F_{Y_1}, \ldots, F_{Y_q}\}$. The marginal likelihood approach estimates P using $\Pr(\bar{Z} \in \mathcal{D}|P)$ only, treating F_X and F_Y as nuisance parameters. Bayesian inference conducted using Gibbs sampling gives samples of the posterior distribution, $p(P|\bar{Z} \in \mathcal{D}) \propto p(P)\; p(\bar{Z} \in \mathcal{D}|P)$, can handle missing values and rank-deficient correlation matrices when $n < p + q$. Our algorithm (available in the supplementary material along with our R implementation) follows Hoff (2007) but is reparametrized in terms of precision matrix to gain efficiency.

4. Experiments

Simulation: Comparison between different IB methods. We generate training samples with $n = 1000$ observations $(x_i, y_i), i = 1, \ldots, n$ and dimensions fixed to $p = 15$, $q = 15$. The samples are drawn from a meta-Gaussian distribution (i.e. in the form of (5)) in two steps. First, we generate the Gaussian hidden variables $(\bar{X}, \bar{Y})$, then using the margin transformations (18) we obtain samples of (X, Y). P is obtained by scaling a covariance matrix drawn from a Wishart distribution. We use a Wishart distribution centered at a correlation matrix P_0 populated with some high correlation values to ensure some dependency between X and Y. In our first experiment we compare:

1. *MGIB bound*: MGIB is applied to observations of the continuous hidden variables $\bar{X}, \bar{Y}$. These hidden variables are not observable in practice and MGIB bound provides an upper bound on achievable information curves.
2. *MGIB*: We apply MGIB to observations $\{(x_i, y_i), i = 1, \ldots, n\}$ of the mixed variables without adjustment for mixed data i.e. using the model introduced in Rey & Roth (2012).
3. *Sparse MGIB*: We apply sparse MGIB (no adjustment for mixed data) to $\{(x_i, y_i)\}$.
4. *MMGIB*: Mixed Meta-Gaussian IB is MGIB for mixed data applied to $\{(x_i, y_i)\}$.
5. *Sparse MMGIB*: Sparse version of MMGIB applied to $\{(x_i, y_i)\}$.

We assess the efficiency of the different compression matrices A obtained by the above methods on a test set with 5000 observations. The compression T is calculated using the projection matrix A obtained on the training data and the mutual information $I(\bar{X}; T), I(\bar{Y}; T)$ are calculated using the formula for meta-Gaussian variables given in Rey & Roth (2012). Simulations are conducted with mixed margins for X and Y. For each dimension we use one of the following distributions: Student t_4, $\mathrm{Binomial}(2, 0.5)$ or $\mathrm{Binomial}(10, 0.5)$. By varying the parameter κ between 0.1 and 80 we can represent $I(Y; T)$ as a function of $I(X; T)$ and obtain the information curves. Each experiment is repeated to obtain the 50 curves for each method (shown in the top panel of Figure 2). We can see that some information was lost during the discretisation process $I(X; Y) < I(\bar{X}, \bar{Y})$, and therefore the most effective compression (green curves) is obtained when applying MGIB to $\bar{X}, \bar{Y}$ (not observable in practice). MMGIB (blue curves) performs clearly better than MGIB (red curves) on mixed data and achieves a compression rate closer to the MGIB bound. The top panel of Figure 2 also illustrates the difference between sparse and traditional IB. The information curves for sparse IB lie slightly below the traditional IB curves since less information can be captured when A is restricted to a diagonal matrix. However, sparse and traditional IB curves tend to the same value $I(Y; T)$ as $I(X; T)$ increases.

Simulation: Feature selection. To test the efficiency of sparse MMGIB in selecting relevant dimensions of X, we generate data such that only some dimensions of X were informative about Y. This is achieved by using a correlation matrix P_0 with only a few non-zero entries and a Wishart distribution with high degrees of freedom. The margins of X and Y were again mixed, following either a $\mathrm{Beta}(0.5, 0.5)$ or a $\mathrm{Binomial}(10, 0.5)$ distribution. The bottom panel of Figure 2 shows results obtained with the following choice for P_0: three dimensions of X are strongly informative with corresponding values of 0.8 in

P_0, three other dimensions have entries 0.6, three more dimensions 0.4 and the six remaining dimensions are noise with zero correlation in P_0. Figure 2 shows the solution paths for the 15 entries of A, each line corresponding to one dimension of X. The information curve obtained is shown in red with the corresponding κ values. The most informative dimensions (green curves) are selected first and can be clearly identified. Then, as κ increases, the remaining informative dimensions (blue and lilac curves) are selected as well. The noise dimensions are always selected last, when the value of κ becomes higher, to allow $I(X;T)$ to reach the required level of 0.5κ.

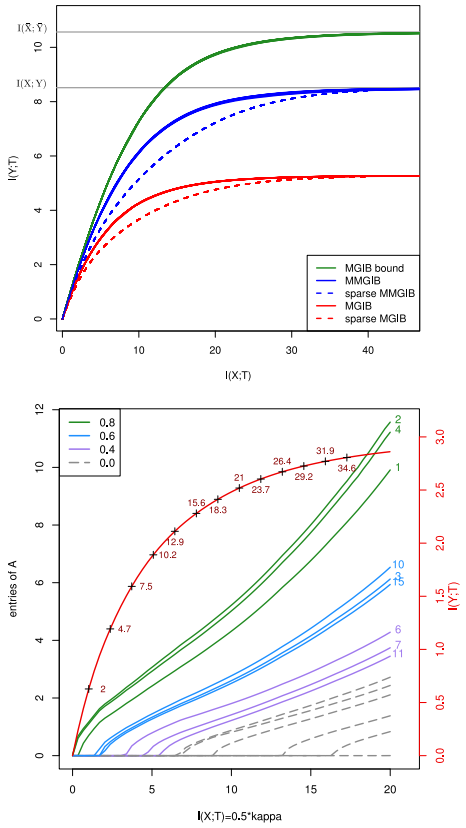

Figure 2. **Top:** The figure shows 50 (overlapping) information curves for each method. MGIB bound (green curves) provides a benchmark. MMGIB (blue curves) achieves a better compression than MGIB (red curves). Sparsity in T is achieved at the price of a small decrease in efficiency (dashed curves). **Bottom:** Solution paths for the 15 entries of A and corresponding information curve with κ values. Green curves correspond to very informative dimensions of X (correlation 0.8 in P_0), blue and lilac curves to informative dimensions (corr. of 0.6 and 0.4 in P_0, respectively), and gray dashed curves represent noise (corr. 0 in P_0).

4.1. Real data

We consider a real world problem from computational biology. A recurring and important task in medical prognosis is to identity biomarkers relevant to the disease evolution (Fuchs & Buhmann, 2011). Identification of a small number of key variables provides additional insight into the disease's mechanisms, and is crucial for cost-effective prognosis and therapy optimisation. We consider this selection problem in the context of cutaneous malignant melanoma (MM), the most common cause for fatalities in skin cancer. A first promising approach to identify biomarkers important for survival prediction was reported in Meyer et al. (2012). Data was available in the form of immunohistochemical (IHC) expressions of 70 candidate biomarkers measured for 364 patients. Additionally, 9 different clinical observations were available which reflected experts' opinion about the stage of the tumor or directly characterised the severeness of the disease in terms of survival times. Focusing exclusively on survival information (and ignoring all other clinical attributes), a 7 *marker signature* which is used to separate the patients into a low-risk and high-risk group has been proposed in Meyer et al. (2012). In particular, the signature is defined via a risk-score of the form

$$\text{score}(x) = \frac{\sum_{i=1}^{7}(\beta_i x_i)\alpha_i}{\sum_{i=1}^{7}\alpha_i}, \alpha_i = \begin{cases} 1 & \text{if } x_i \text{ is measured} \\ 0 & \text{if } x_i \text{ is missing} \end{cases}$$

where β_i are the regression coefficients of a univariate Cox model and x_i are are the IHC expression measurements of the 7 markers. A convincing statistical interpretation of the selected markers, however, remains unclear: the selection proceeded in several stages where only univariate tests have been used. Moreover, the relation of the biomarkers to established prognosis-related clinical observations like the pathological Tumor-Node-Metastasis (pTNM) staging or the Clark level was ignored in the model.

Our sparse MMGIB model is best adapted to this problem, since the data falls into two groups which nicely fit into our framework: the 70 markers constitute the candidate features X, whereas the 9 clinical observations can be used for the target variable Y. Defining a *signature* of molecular markers might be seen as finding the best sparse compression of the biomarkers' expression on a molecular level which is still informative with respect to the clinical data in the second group. Further, the technical specifications of this dataset also perfectly fit into our mixed-data Bayesian framework: most of the expression levels are represented as ordered factors in a semi-quantitative scoring system with 5 levels, but other variables like survival times are continuous in nature, and roughly 10% of all values are missing. Repeated experiments conducted to a final selection of 6 markers as explained in the left panel of Figure 3.

Interestingly, our information-theoretic analysis method which is not particularly tailored to survival prediction could nicely reproduce the survival regression results in Meyer et al. (2012): using our set of 6 markers (containing 4 markers which were already part of the original 7 markers signature) in the same "signature" formalism also

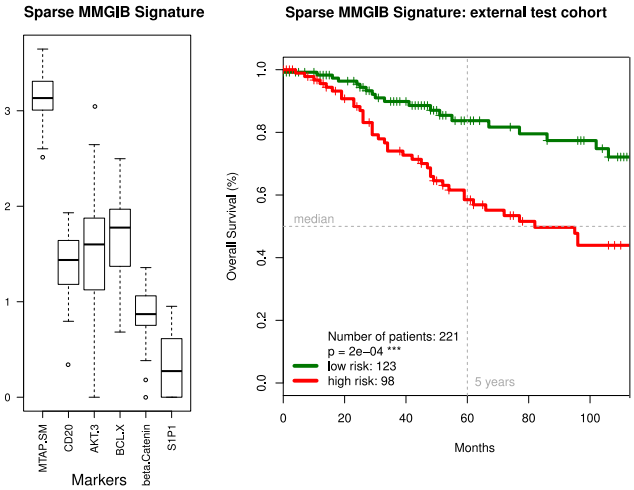

Figure 3. **Left:** boxplots of the elements A_{ii} in the diagonal projection matrix computed on the basis of 50 consecutive samples from the posterior distribution of the correlation matrix in the Gibbs sampler. We show here only the values for the 6 markers finally selected (those with a median above zero). The markers MTAP, CD20, Bcl-X and Beta-catenin were already identified in Meyer et al. (2012). **Right:** Kaplan-Meier plots of the two patient groups from the test cohort resulting from thresholding the *risk score* in eq. (4.1) computed from our 6 marker signature

led to clear separation of low-risk and high-risk patients on an independent test cohort. The right panel of Figure 3 shows Kaplan-Meier estimates of two different patient groups separated by using the risk-score defined above applied to *our* signature. The β_i coefficients were calculated on the training set described above but the 6 markers expressions were measured on an external test cohort of 221 patients. We see that the 6 markers selected using our method indeed provide highly effective prognosis predictors. Given the relatively small number of patients, the differences in p-values ($2 \cdot 10^{-4}$ for our signature vs. $2 \cdot 10^{-5}$ reported in panel A of Figure 6 in Meyer et al. (2012)) are probably not too relevant. This observation is corroborated by the fact that the effect of regular updates on the the patients' censoring status from new follow-up reports have led to even bigger differences for the individual signatures in the past. We conclude that our sparse IB model can indeed be used as a high-quality prognosis predictor, even though it was not specifically designed for survival regression. The real advantage of the IB model, however, is its capability to extract many more details about the interaction between markers and a rich set of clinical measurements. Much could be said about the interpretation of the joint (X, Y)-correlation matrix obtained within the semi-parametric copula framework. Due to space constraints, however, we focus here only on one particularly interesting aspect, namely the *reversal* of the roles of X and Y, resulting in a sparse compression of the clinical variables Y subject to a constraint of preserving information about the molecular markers X. Figure 4 shows the corresponding solutions paths. Interestingly, it turns out that the *disease specific event status for overall survival* contains by far the most information about the molecular markers. This dominance of the disease specific event status over classical prognosis-related quantities like the *T score* or the *clark level* is interesting for the following reason: it basically shows that the classical clinical indicators fail to capture all relevant disease-specific information contained in the molecular data, showing that quantitative analysis of the biomarkers' expression patterns indeed adds valuable information about prognosis and survival in addition to the expert's macroscopic scoring/staging estimates.

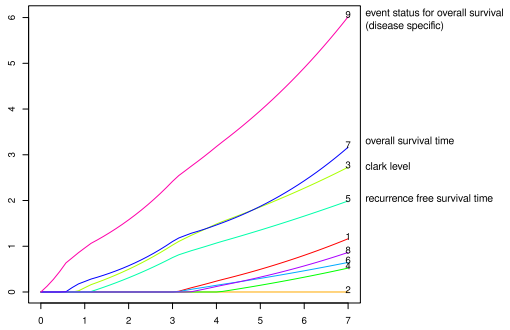

Figure 4. Solution paths for 9 diagonal entries of A when the roles of X and Y are reversed.

5. Conclusion

Sparse Meta-Gaussian IB provides a very flexible method for sparse compression with side information. By assuming a Gaussian copula, it encompasses a wide class of distributions with arbitrary non-Gaussian margins (continuous or discrete). Using relevance information, we do not need to impose any norm penalty to obtain sparsity. Our Bayesian framework for copula estimation can handle large-scale high dimensional datasets with potentially missing values. Our log barrier interior point algorithm is efficient even in high dimensions. Moreover, we prove that the two globally best input features can be found in arbitrary dimensions in an efficient way, thereby also providing an additional validation of the results obtained with our algorithm. Finally, we demonstrate in a clinical application that our model can compete with state-of-the-art survival prediction methods, while additionally allowing for an in-depth analysis of relevance- and interaction patterns between molecular markers and clinical measurements. We conclude that the proposed model is a highly flexible analysis tool which has the potential to significantly advance the field of exploratory data analysis within a well-defined information-theoretic framework.

References

Benson, H. P. and Horst, R. A branch and bound-outer approximation algorithm for concave minimization over a convex set. *Journal of Computers and Mathematics with Applications*, 21(6/7):67–76, 1991.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

de Leon, A.R. and Chough, K. Carriére. *Analysis of mixed data: methods and applications*. Chapman and Hall, 2013.

Fuchs, T.J. and Buhmann, J.M. Computational pathology: challenges and promises fortissue analysis. *Journal of Computerized Medical Imaging and Graphics*, 35(7):515–530, April 2011. ISSN 0895-6111. doi: DOI:10.1016/j.compmedimag.2011.02.006. URL http://www.sciencedirect.com/science/article/pii/S0895611111000383.

Genest, C. and Nešlehová, J. A primer on copulas for count data. *Astin Bulletin*, 37(2):475–515, 2007.

Hoff, Peter D. Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1 (1):273, 2007.

Joe, H. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164, 1989.

Liu, H., Lafferty, J., and Wasserman, L. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

Meyer, S., Fuchs, T.J., and Wild, P.J. A seven-marker signature and clinical outcome in malignant melanoma: a large-scale tissue-microarray study with two independen patient cohorts. *PLoS ONE*, 7(6), 2012.

Pitt, M., Chan, D., and Kohn, R. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

Rey, M. and Roth, V. Meta-gaussian information bottleneck. *Proceedings of the Neural Information Processing Systems (NIPS)*, 2012.

Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Smith, M.S. and Khaled, M.A. Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, 107(497):290–303, 2012.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1): 267–288, 1996.

Tishby, N., Pereira, F.C., and Bialek, W. The information bottleneck method. *The 37th annual Allerton Conference on Communication, Control, and Computing*, (29-30):368–377, 1999.