Sparse meta-Gaussian information bottleneck

Detailed calculations for Lemma 3.2

A straightforward derivation leads to

$$H_{12} = (\Phi_{11} + a_1)(\Psi_{22} + a_2) - (\Phi_{22} + a_2)(\Psi_{11} + a_1).$$

Setting H_{12} to zero gives:

$$a_2 = \left(\frac{\Psi_{22} - \Phi_{22}}{\Psi_{11} - \Phi_{11}}\right) a_1 + \frac{\Phi_{11}\Psi_{22} - \Phi_{22}\Psi_{11}}{\Psi_{11} - \Phi_{11}},$$

from which we can identify

$$c_1 = \frac{\Psi_{22} - \Phi_{22}}{\Psi_{11} - \Phi_{11}}, \quad c_0 = \frac{\Phi_{11}\Psi_{22} - \Phi_{22}\Psi_{11}}{\Psi_{11} - \Phi_{11}}. \qquad (1)$$

Noting that $\Psi_{11} = \Psi_{22} = |\Psi|$, we find the final expressions for c_0 and c_1.

We further set H_0 to zero and replace a_2 by $c_1 a_1 + c_0$ to obtain:

$$\begin{aligned}
&|P_x||\Psi + A| = |\Psi|^{-1}|\Psi + A| = e^{\kappa},\\
&(\Psi_{11} + a_1)(\Psi_{22} + c_1 a_1 + c_0) - \Psi_{12}^2 = |\Psi| e^{\kappa},\\
&a_1^2 + a_1 \frac{|\Psi|(1 + c_1) + c_0}{c_1} + \frac{|\Psi|(c_0 + 1 - e^{\kappa})}{c_1} = 0 \qquad (2)
\end{aligned}$$

We can recognise in (2) a quadratic equation of the form $a_1^2 + r a_1 + s = 0$. Since we assumed that dimension 2 is the most informative we have that $c_0 > 0$ and the solution for a_1 is then given by $a_1 = -\frac{r}{2} + \left(\frac{r^2}{4} - s\right)^{0.5}$ leading directly to equation (11) of Lemma 3.2. Finally, we can compute the critical value κ_1^c by noting that $a_1 = 0$ is equivalent to $|\Phi|(c_0 + 1 - e^{\kappa})/c_1 = 0$ which implies that $c_0 + 1 - e^{\kappa_1^c} = 0$.

Additional check for Algorithm 1

As an additional check of the path of stationary points obtained with Algorithm 1, we verify that this path does not have any bifurcations. We thereby insures that no other path connecting stationary points rejoins or diverges from the obtained path. A classical way to study bifurcations in 1-dimensional manifolds is provided by the *Implicit function theorem*. We first need to derive a set of equations which characterise the set of stationary points. For a stationary point a^* with strictly positive components, the non-negativity constraints are inactive and $\epsilon_j = 0, \ \forall j$. Stationary points are characterised by a vanishing Lagrangian gradient $\nabla \mathcal{L} = 0$, meaning that $\nabla f(a^*) = \lambda \nabla g(a^*)$. This proportionality condition can be translated into an orthogonality condition which eliminates λ: $\nabla f(a^*)$ must be orthogonal to the $(p-1)$-dimensional hyperplan orthogonal to $\nabla g(a^*)$. Constructing a basis $(g_\perp^1, \ldots, g_\perp^{p-1})$ of this hyperplan we obtain $p-1$ orthogonality conditions: $\nabla f \cdot g_\perp^i = 0, i = 1, \ldots, p-1$. Adding the constraint $g(a) = \kappa$ leads to a set of p equations in $p+1$ variables (a and κ). In the following we denote the partial derivatives of a real function f of a by $\frac{\partial f}{\partial a_i}(a) = f_{a_i}$, and the matrix of partial derivatives for a vector-valued function $\mathcal{F}$ by $J_a \mathcal{F}$. We further assume that $P_x, P_{x|y}$ have full rank and write $\Phi := P_{x|y}^{-1}, \Psi := P_x^{-1}$.

In the p-dimensional case, the hyperplan orthogonal to ∇g is $(p-1)$-dimensional and a basis for it is given by $g_\perp^1, \ldots, g_\perp^{p-1}$, where the vectors $g_\perp^i$ have $-g_{a_{i+1}}$ at position i, g_{a_i} at position $i+1$ and 0 otherwise. The set of stationary points is then implicitly defined by the equation $H = 0$, where $H : \mathbb{R}^{p+1} \to \mathbb{R}^p$ is defined by

$$H(a, \kappa) = \begin{pmatrix} H_1(a, \kappa) \\ \vdots \\ H_{p-1}(a, \kappa) \\ H_p(a, \kappa) \end{pmatrix} = \begin{pmatrix} \nabla f(a) \cdot g_\perp^1(a) \\ \vdots \\ \nabla f(a) \cdot g_\perp^{p-1}(a) \\ g(a) - \kappa \end{pmatrix}. \qquad (3)$$

By the Implicit function theorem we know that if $|J_a H(a^*)| \neq 0$ for some point $a^* \in S$, then in a neighbourhood of a^* the solution path S has no bifurcation. While running Algorithm 1 we therefore regularly check that this determinant remains non-zero: the algorithm proceeds by successive optimisation steps with decreasing κ values $\{\kappa_0 > \cdots > \kappa_m\}$ and for each value obtains an optimum $a^*(\kappa)$, for every such optimum we can then verify that $|J_a H(a^*)| \neq 0$. This operation can be efficiently conducted since the computation of all partial derivatives $\partial H_i / \partial a_j$ requires only two matrix inversions. Indeed, for $i = 1, \ldots, p-1$ we have

$$\begin{aligned}
\frac{\partial H_i}{\partial a_j}(a) &= f_{a_{i+1}, a_j} g_{a_i} + f_{a_{i+1}} g_{a_i, a_j}\\
&\quad - f_{a_i, a_j} g_{a_{i+1}} - f_{a_i} g_{a_{i+1}, a_j}, \qquad (4)\\
f_{a_i} &= (\Phi + A)^{-1}_{ii}, \quad g_{a_i} = (\Psi + A)^{-1}_{ii},\\
f_{a_i, a_j} &= (-1)^{i+j}(\Phi + A)^{-1}_{ij} - (\Phi + A)^{-1}_{ii}(\Phi + A)^{-1}_{jj},\\
g_{a_i, a_j} &= (-1)^{i+j}(\Psi + A)^{-1}_{ij} - (\Psi + A)^{-1}_{ii}(\Psi + A)^{-1}_{jj},
\end{aligned}$$

where $f_{a_i,a_j} = \frac{\partial^2 f_i}{\partial a_i \partial a_j}(a)$ and $g_{a_i,a_j} = \frac{\partial^2 g_i}{\partial a_i \partial a_j}(a)$. The remaining elements of the Jacobian are given by $\partial H_p / \partial a_j(a) = g_{a_j} = -((\Psi + A)^{-1}_{jj})^2$ for $j = 1, \ldots, p$.

Imputation of P and $\bar{Z}$

Aim: sample from the posterior distribution $p(P|\bar{Z} \in \mathcal{D}) \propto p(P)\, p(\bar{Z} \in \mathcal{D}|P)$.
For the sampling we introduce a new variable, a precision matrix B such that P is the correlation matrix obtained by scaling B^{-1}, i.e. P has elements $P_{ij} = \frac{B^{-1}_{ij}}{\sqrt{B^{-1}_{ii} B^{-1}_{jj}}}$. The prior distribution of B is Wishart: $p(B) \sim \text{Wishart}(\nu, B_0)$. We will use the notation $P(B)$ to emphasize that P is calculated as a function of B. The Bayesian inference procedure described below and implemented as in Algorithm 1 uses Gibbs sampling. This method is analogue to the algorithm presented in (Hoff, 2007) but is parametrized using a precision matrix B instead of a covariance matrix, we thereby avoid repetitive matrix inversions. Sampling is made of three steps:

1. Sample $\bar{Z}|B, \bar{Z} \in \mathcal{D}$.
2. Sample $B|\bar{Z} \sim \text{Wishart}\left(\nu + n, \left(B_0 + \bar{Z}^T \bar{Z}\right)^{-1}\right)$, where n is the number of observations.
3. Compute $P(B)$ such that $P_{ij} = \frac{B^{-1}_{ij}}{\sqrt{B^{-1}_{ii} B^{-1}_{jj}}}$.

In step 1, we sample $\bar{Z}$ iteratively over observations $i = 1, \ldots, n$ and dimensions $j = 1, \ldots, p+q$ from a truncated Gaussian as follows:

$$\bar{Z}_{ij}|B, \bar{Z} \in \mathcal{D}, \bar{Z}_{-i,-j} \sim \mathcal{TN}\left(\mu_{ij}, \sigma_j^2, r_l, r_u\right),$$

where $\mu_{ij} = \bar{z}_{i,-j} B_{-j,j} / (-B_{jj})$ and $\sigma_j^2 = 1/B_{jj}$. Here $\bar{z}_{i,-j}$ denotes the i^{th} observation from the previous sweep from which dimension j has been removed, similarly $B_{-j,j}$ denotes the j^{th} column of B from which the row j has been removed. The truncation boundaries are determined by the condition $\bar{Z} \in \mathcal{D}$: the lower bound r_l is $\max\{\bar{z}_{ij}|z_{ij} < r\}$ and the upper bound r_u is $\min\{\bar{z}_{ij}|z_{ij} > r\}$, where the optima are taken over i for each dimension j separately. Note that here the samples of $\bar{Z}$ do not have unit variance. Samples with unit variance can be obtained by scaling.

Algorithm 1 Imputation of $\bar{Z}$ and P.

0. The prior distribution of B is $\text{Wishart}(\nu, B_0)$;
1. Update $\bar{Z}$:
for $j = 1, \ldots, p+q$ **do**
  Set $\sigma_j := 1/B_{jj}$;
  for $r \in \text{unique}\{z_{1j}, \ldots, z_{nj}\}$ **do**
    set lower bound to $r_l := \max\{\bar{z}_{ij}|z_{ij} < r\}$;
    set upper bound to $r_u := \min\{\bar{z}_{ij}|z_{ij} > r\}$;
    for $i \in \{1, \ldots, n\}|z_{ij} = r$ **do**
      compute $\mu_{ij} := z_{i,-j} B_{-j,j} / (-B_{jj})$;
      sample $\bar{z}_{ij} \sim \mathcal{TN}(\mu_{ij}, \sigma_j^2, r_l, r_u)$ from a truncated Gaussian;
    end for
  end for
end for
2. Sample $B \sim \text{Wishart}\left(\nu + n, \left(B_0 + \bar{Z}^T \bar{Z}\right)^{-1}\right)$.
3. Compute P: $P_{ij} = (B^{-1})_{ij} / \sqrt{(B^{-1})_{ii} (B^{-1})_{jj}}$.

Experimental details

We work with the data sets described in (Meyer et al., 2012). The data is composed of a primary cohort (training set) with 364 patients and an independent secondary cohort (test set) with 221 patients. We use for $X = (X_1, \ldots, X_{70})$ the 70 different biomarkers expression measurements available in the training set. For $Y = (Y_1, \ldots, Y_9)$ we use 9 different clinical observations available again for the training set. Theses 9 variables are:

1. last known stage of the tumor
2. T score tumor staging
3. clark level
4. tumor thickness
5. recurrence free survival time
6. event status for recurrence free survival
7. overall survival time
8. event status for overall survival
9. event status for overall survival (disease specific)

References

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

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.