Journal of Biomedical Informatics: X 4 (2019) 100059

Contents lists available at ScienceDirect

Journal of Biomedical Informatics: X

journal homepage: www.journals.elsevier.com/journal-of-biomedical-informatics-x

Bayesian multitask learning regression for heterogeneous patient cohorts

Andre Goncalves[a,*], Priyadip Ray[a], Braden Soper[a], David Widemann[a], Mari Nygård[b], Jan F. Nygård[b], Ana Paula Sales[a]

[a] Lawrence Livermore National Laboratory, Livermore, CA, USA
[b] Cancer Registry of Norway, Oslo, Norway

ARTICLE INFO

Keywords:
Multitask learning
Bayesian modeling
Structured learning
Uncertainty quantification
Alzheimer's disease progression
Biomedical application

ABSTRACT

Multitask learning (MTL) leverages commonalities across related tasks with the aim of improving individual task performance. A key modeling choice in designing MTL models is the structure of the tasks' relatedness, which may not be known. Here we propose a Bayesian multitask learning model that is able to infer the task relationship structure directly from the data. We present two variations of the model in terms of *a priori* information of task relatedness. First, a diffuse Wishart prior is placed on a task precision matrix so that all tasks are assumed to be equally related *a priori*. Second, a Bayesian graphical LASSO prior is used on the task precision matrix to impose sparsity in the task relatedness. Motivated by machine learning applications in the biomedical domain, we emphasize interpretability and uncertainty quantification in our models. To encourage model interpretability, linear mappings from the shared input spaces to task-dependent output spaces are used. To encourage uncertainty quantification, conjugate priors are used so that full posterior inference is possible. Using synthetic data, we show that our model is able to recover the underlying task relationships as well as features jointly relevant for all tasks. We demonstrate the utility of our model on three distinct biomedical applications: Alzheimer's disease progression, Parkinson's disease assessment, and cervical cancer screening compliance. We show that our model outperforms Single Task (STL) models in terms of predictive performance, and performs better than existing MTL methods for the majority of the scenarios.

1. Introduction

Multitask learning (MTL) is the sub-field of machine learning in which individual models for performing potentially related tasks are learned jointly [1]. The need for this type of modeling arises across a wide variety of real-world applications where different datasets 1) cannot be treated as independent and identically distributed, but 2) are similar enough that independent models (a.k.a, single-task learning) are sub-optimal.

The goal of MTL is to share statistical information across task-specific models in such a way that the overall performance of related tasks is improved, without negatively impacting the performance of unrelated tasks. By jointly modeling the different tasks, MTL enables the borrowing of statistical strength across tasks, effectively increasing sample sizes, which is particularly advantageous for applications where the number of variables is large compared to the number of samples.

MTL has been an active area of research over the past two decades, with applications ranging from object detection in computer vision, integration of different types of genomic data and sentiment analysis in social media [2–5]. A wide variety of MTL approaches have been proposed (for a recent survey of the field see [6]), and common themes of information sharing have emerged, such as sharing of model parameters (such as Gaussian process parameters in [7] and neural networks layers in [1,8]), task clustering (e.g., [9,10]) and common priors in hierarchical Bayesian models (e.g., [11,12,10]). Typically, the input spaces corresponding to the different tasks are common, but the mappings from the input space to the output variables are different, with both linear and nonlinear mappings from the shared input space to task-dependent output having been proposed [13,11,14–16].

In this paper, we present a novel hierarchical Bayesian MTL approach, referred to as **B***ayesian* **M***ultitask with* **S***tructure* **L***earning* (BMSL), motivated by different biomedical machine learning applications, such as Alzheimer's and Parkinson's disease, and cervical cancer screening. In these types of applications, as is generally true across many other scientific domains, both model performance and model interpretability are critical for real-world model deployment.

Our model emphasizes interpretability in two important ways. Firstly, it imposes sparsity on the input features to enable learning from

* Corresponding author.
E-mail address: goncalves1@llnl.gov (A. Goncalves).

https://doi.org/10.1016/j.yjbinx.2019.100059

Available online 18 October 2019

smaller feature sets and consequently removing unnecessary input variables. Also, to infer the conditional dependence structure across tasks, we impose sparsity on the inverse covariance matrix using the Bayesian graphical LASSO as proposed in [17]. Secondly, our model uses linear mappings from the shared input space to the task-dependent outputs. While non-linear mappings provide for more flexible models and potentially improved predictive performance, they are generally difficult to interpret, which is unacceptable in many biomedical applications. Lack of interpretability is widely recognized as a major drawback of these types of models and is currently an active area of research. While progress is being made on this front [18,19], the reality remains that due to the complex sequence of nonlinear transformations in such models, interpretability is still a major challenge. Moreover, in a number of medical applications, linear models have been shown to perform just as well as, if not better, than more complex non-linear models [20].

In addition to interpretability, scientific problems often require the ability to incorporate model uncertainty and domain knowledge in a structured framework. All of this is naturally accomplished via the Bayesian framework in our model. Finally, in contrast to existing Bayesian MTL approaches, our proposed model has closed-form full conditional distributions so that posterior inference can be obtained via Gibbs sampling, without the need to resort to approximate inference techniques.

Using synthetic data we demonstrate the efficacy of our model and show that it can recover the underlying task relation structure. Finally, we present results on three biomedical applications: Alzheimer's disease progression, Parkinson's disease assessment, and cervical cancer screening compliance. These datasets vary in the number of observations and the number of tasks, with the number of observations ranging from dozens of samples to thousands, and the number of tasks ranging from 5 to 42. Results showed that BMSL is capable of discovering an underlying task dependence structure and identifying features which are non-relevant across all tasks. Prediction performance is shown to be superior to other existing multitask learning formulations for the majority of the tasks analyzed.

2. The BMSL model

In this section we present the Bayesian Multitask with Structure Learning (BMSL) model and propose a Gibbs sampler for inference. We first introduce notation that will be used throughout the paper.

We use bold upper case letters for matrices, bold lower case letters for vectors, and no bold lower case for scalars. Let $\mathbf{X}_t \in \mathbb{R}^{N_t \times d}$ and $\boldsymbol{y}_t \in \mathbb{R}^{N_t}$ be the design matrix and observation vector for the t-th task, d the problem dimension (assumed to be the same for all tasks), N_t the number of samples in the t-th task, T the total number of tasks, $\mathbf{W} \in \mathbb{R}^{d \times T}$ the coefficient matrix for all tasks, where $\mathbf{w}_t \in \mathbb{R}^d$ and $\mathbf{w}_{(j)} \in \mathbb{R}^T$ with $t = 1, ..., T$ and $j = 1, ..., d$ are the columns and rows of $\mathbf{W}$, respectively. The Hadamard (element-wise) product of vectors $\mathbf{a}$ and $\mathbf{b}$ is denoted by $\mathbf{a} \odot \mathbf{b}$. Let $\mathbf{I}_n$ be the $n \times n$ identity matrix. Let $x \sim \text{Normal}(\mu, \tau)$ indicate that x is a normal random variable with mean μ and precision τ, and let $\text{Normal}(x|\mu, \tau)$ denote the density function evaluated at x. Similar notation applies to other standard distributions.

2.1. BMSL with Wishart prior on precision matrix (BMSL-W)

As discussed earlier, for scientific machine learning applications linear models are often preferred for their interpretability. As an added regularization step and to further improve interpretability and explainability, we impose sparsity on $\mathbf{w}_t$ (task specific regressors) via a beta-Bernoulli hierarchy. *A priori* this will push some of the rows of $\mathbf{W}$ to be exactly zero, implying that some of the variables are irrelevant to all tasks. Next, to share statistical strengths across tasks, the rows of $\mathbf{W}$, denoted by $\mathbf{w}_{(j)}$, are drawn from a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$. We do not assume knowledge of the task dependencies, rather we infer the covariance structure from the data. The detailed model follows.

$$
\begin{aligned}
&y_t^i \sim \text{Normal}(\mathbf{x}_t^i(\mathbf{w}_t \odot \boldsymbol{\beta}), \tau_t), \qquad i = 1, ..., N_t; \quad t = 1, ..., T \\
&\tau_t \sim \text{Gamma}(a_\tau, b_\tau), \qquad t = 1, ..., T \\
&\mathbf{w}_{(j)} \sim \text{Normal}(0, \boldsymbol{\Sigma}^{-1}), \qquad j = 1, ..., d \\
&\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(\mathbf{V}, \nu) \\
&\beta_j \sim \text{Bernoulli}(\theta), \qquad j = 1, ..., d \\
&\theta \sim \text{Beta}(a_\theta, b_\theta)
\end{aligned}
$$

The hyperparameters $a_\tau > 0$, $b_\tau > 0$, $\mathbf{V} > 0$, $\nu \geqslant T$, $a_\theta > 0$ and $b_\theta > 0$ should be chosen by the researcher for the specific application. More details are given in Section 4 on choosing these values. We refer to this BMSL model with a Wishart prior on the precision matrix as the BMSL-W model.

2.1.1. BMSL with sparsity on precision matrix (BMSL-GL)

In the above construction, the task dependence structure is captured by the inverse covariance (precision) matrix, whose prior distribution is assumed to be Wishart. *A priori* this does not impose any sparsity on the precision matrix. However, sparsity of the precision matrix can be a desired feature, particularly for high-dimensional problems with limited data or for uncovering the conditional dependence structure across the variables. In our particular application, a zero entry in the inverse covariance matrix $\Sigma_{i,j}^{-1}$ indicates that tasks i and j are conditionally independent and, hence, conditioned on the remaining tasks, do not share any commonalities.

To extend the BMSL model to impose sparsity on the precision matrix $\boldsymbol{\Sigma}^{-1}$, we adopt the graphical LASSO prior [17] on the task dependence matrix. The hierarchical model is as follows.

$$
\begin{aligned}
y_t^i &\sim \text{Normal}(\mathbf{x}_t^i(\mathbf{w}_t \odot \boldsymbol{\beta}), \tau_t), \qquad i = 1, ..., N_t; \quad t = 1, ..., T \\
\tau_t &\sim \text{Gamma}(a_\tau, b_\tau), \qquad t = 1, ..., T \\
\mathbf{w}_{(j)} &\sim \text{Normal}(0, \boldsymbol{\Sigma}^{-1}), \qquad j = 1, ..., d \\
\beta_j &\sim \text{Bernoulli}(\theta), \qquad j = 1, ..., d \\
\theta &\sim \text{Beta}(a_\theta, b_\theta) \\
\boldsymbol{\Sigma}^{-1} &\sim C^{-1} \prod_{i<j} \{DE(\sigma_{ij}^{-1}|\lambda)\} \prod_{i=1}^{p} \{EXP(\sigma_{ii}^{-1}|\lambda/2)\} \mathbf{1}_{\boldsymbol{\Sigma}^{-1} \in \mathbf{M}^+} \\
\lambda &\sim \text{Gamma}(a_\lambda, b_\lambda)
\end{aligned}
$$

Here C is a normalizing constant not involving λ, $DE(x|\lambda)$ is the double exponential density function of the form $p(x) = \lambda/2\exp(-\lambda|x|)$, and $EXP(x|\lambda)$ is the exponential density function of the form $p(x) = \lambda\exp(-\lambda x)\mathbf{1}_{x>0}$. The Bayesian graphical lasso prior encourages the shrinkage of off-diagonal elements of the precision matrix towards zero as the dimensionality d increases. The hyperparameters $a_\tau > 0$, $b_\tau > 0$, $a_\theta > 0$, $b_\theta > 0$, $a_\lambda > 0$, $b_\lambda > 0$ should again be chosen by the researcher for the specific application. More details are given in Section 4 on choosing these values. We refer to this BMSL variant with a Bayesian graphical LASSO prior on the precision matrix as BMSL-GL.

The Bayesian Multitask with Structure Learning (BMSL) approach proposed above is applicable for general regression tasks. As demonstrated in Section 4.1 (Experimental results for synthetic data), the model is applicable to a wide range of scenarios, including both strong and weak coupling across tasks. However, the performance of the multitask model approaches that of a single task model, as the coupling across tasks reduces.

The proposed BMSL approach is a linear model with i.i.d. Gaussian noise (residue). This is a fairly general and widely applicable assumption. However, BMSL cannot be applied to applications where the noise is impulsive/heavy-tailed. Such scenarios are often encountered in imaging applications [21,22]. Further, the BMSL approach is directly applicable to continuous datatypes. However, the model can be easily extended to categorical and ordinal datatypes, via appropriate encoding schemes such as one-hot encoding and label encoding. For example, for the results on the CRN dataset presented in Section 4.4, we have applied

label encoding for categorical variables such as, marital status, sex, and school level.

2.2. Gibbs sampling for BMSL

We next provide the Gibbs inference steps for the BMSL-W, that is, the model without sparsity on the precision matrix. The full posterior distribution is given as follows.

$$\begin{aligned} p(\mathbf{W}, \tau, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \theta|\mathbf{X}, \mathbf{Y}) &\propto p(\mathbf{Y}|\mathbf{W}, \tau, \boldsymbol{\beta}, \mathbf{X}) p(\mathbf{W}|\boldsymbol{\Sigma}) \\ &\quad p(\boldsymbol{\tau}|a_\tau, b_\tau) p(\boldsymbol{\Sigma}^{-1}|\mathbf{V}, \nu) p(\boldsymbol{\beta}|\theta) p(\theta|a_\theta, b_\theta) \\ &= \prod_{t=1}^{T} \prod_{i=1}^{N_t} p(y_t^i|\mathbf{w}_t, \boldsymbol{\beta}, \mathbf{x}_t^i, \tau_t) \prod_{j=1}^{d} p(\mathbf{w}_{(j)}|0, \boldsymbol{\Sigma}) \\ &\quad \prod_{t=1}^{T} p(\tau_t|a_\tau, b_\tau) p(\boldsymbol{\Sigma}^{-1}|\mathbf{V}, \nu) \prod_{j=1}^{d} p(\beta_j|\theta) p(\theta|a_\theta, b_\theta) \end{aligned}$$

Gibbs sampling requires the following full conditional distributions.

Updates for $p(\mathbf{W}|\cdots)$: The full conditional distribution for $\mathbf{W}$ is given by

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{Y}, \tau, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto &\prod_{t=1}^{T} \prod_{i=1}^{N_t} \text{Normal}(y_t^i|\mathbf{x}_t^i(\mathbf{w}_t \odot \boldsymbol{\beta}), \tau_t) \\ &\prod_{j=1}^{d} \text{Normal}(\mathbf{w}_{(j)}|0, \boldsymbol{\Sigma}^{-1}). \end{aligned}$$

Because priors are placed on rows of $\mathbf{W}$, while the likelihood depends on columns of $\mathbf{W}$, standard conjugacy results for multivariate normal distributions do not directly apply to either $\mathbf{w}_{(j)}$ or $\mathbf{w}_t$. However, updating $\mathbf{W}$ element-wise results in a conjugate model so that the full conditional of each element of $\mathbf{W}$ is normal.

We first introduce some notation. For all $j \in \{1, ..., d\}$ and $k \in \{1, ..., T\}$ let w_{jk} be the j, k element of the matrix $\mathbf{W}$, i.e. w_{jk} is the k^{th} element of $\mathbf{w}_{(j)}$. Denote the vector $\mathbf{w}_{(j)}$ with the k^{th} element removed by $\mathbf{w}_{(j),-k}$. Let $\boldsymbol{\Sigma}_{k,k}$ be the k^{th} diagonal element of $\boldsymbol{\Sigma}$, let $\boldsymbol{\Sigma}_{-k,-k}$ be the submatrix defined by removing the k^{th} column and row from $\boldsymbol{\Sigma}$, let $\boldsymbol{\Sigma}_{k:}$ be the k^{th} row of $\boldsymbol{\Sigma}$ with the k^{th} element removed, and let $\boldsymbol{\Sigma}_{:k}$ be the k^{th} column of $\boldsymbol{\Sigma}$ with the k^{th} element removed. Finally, we define the following parameters:

$$\begin{aligned} \alpha_{ijk} &= \sum_{l \neq j} w_{lk} \beta_l x_{kl}^i \text{ for } i = 1, ..., N_k \\ \mu_{jk} &= \boldsymbol{\Sigma}_{k:}[\boldsymbol{\Sigma}_{-k,-k}]^{-1} \mathbf{w}_{(j),-k} \\ \boldsymbol{\Sigma}_k^2 &= \boldsymbol{\Sigma}_{k,k} - \boldsymbol{\Sigma}_{k:}[\boldsymbol{\Sigma}_{-k,-k}]^{-1} \boldsymbol{\Sigma}_{:k} \\ a_{jk} &= \frac{\mu_{jk}\sigma_k^{-2} + \beta_j \tau_k \sum_i (y_k^i - \alpha_{ijk}) x_{kj}^i}{\sigma_k^{-2} + \beta_j \tau_k \sum_i (x_{kj}^i)^2} \\ b_{jk} &= \sigma_k^{-2} + \beta_j \tau_k \sum_i (x_{kj}^i)^2. \end{aligned}$$

The full conditional for w_{jk} can then be written as

$$w_{jk}|\cdots \sim \text{Normal}(a_{jk}, b_{jk}) \tag{1}$$

Derivation: Fix $j \in \{1, ..., d\}$ and $k \in \{1, ..., T\}$. Because the prior on $\mathbf{w}_{(j)}$ is a multivariate normal, the prior distribution of w_{jk} conditioned on $\mathbf{w}_{(j),-k}$ is given as follows:

$$w_{jk}|\mathbf{w}_{(j),-k} \sim \text{Normal}(\mu_{jk}, \sigma_k^{-2}).$$

With α_{ijk} as above we have

$$p(w_{jk}|\cdots) \propto \text{Normal}(w_{jk}|\mu_{jk}, \sigma_k^{-2}) \prod_{i=1}^{N_k} \text{Normal}(y_k^i | \alpha_{ijk} + w_{jk}\beta_j x_{kj}^i, \tau_k).$$

First consider the case that $\beta_j x_{kj}^i \neq 0$ for all $i = 1, ..., N_k$. To simplify notation we define $z_{ijk} = (y_k^i - \alpha_{ijk})/\beta_j x_{kj}^i$ and $\tau_{ijk} = \tau_k(\beta_j x_{kj}^i)^2$. The posterior full conditional distribution can then be written as

$$p(w_{jk}|\cdots) \propto \text{Normal}(w_{jk}|\mu_{jk}, \sigma_k^{-2}) \prod_{i=1}^{N_k} \text{Normal}(z_{ijk} | w_{jk}, \tau_{ijk}). \tag{2}$$

Standard conjugacy results (see for example [23]) now give us

$$w_{jk}|\cdots \sim \text{Normal}\left(\frac{\mu_{jk}\sigma_k^{-2} + \sum_i z_{ijk}\tau_{ijk}}{\sigma_k^{-2} + \sum_i \tau_{ijk}}, \sigma_k^{-2} + \sum_i \tau_{ijk} \right).$$

Replacing z_{ijk} and τ_{ijk} then gives us (1).

If $\beta_j = 0$ then the likelihood does not depend on the parameter w_{jk}, so the full conditional posterior reduces to the conditional prior distribution $\text{Normal}(w_{jk}|\mu_{jk}, \sigma_k^{-2})$ in (2). Notice also that a_{jk} and b_{jk} reduce to μ_{jk} and σ_k^{-2}, respectively, so that (1) still holds in this case. Similarly, if $\beta_j = 1$ and $x_{ikj} = 0$ for at least one i, the corresponding likelihood components can be dropped from the right-hand side of (2) as these data do not depend on the parameter w_{jk}. Following the same arguments as above, we see that (1) holds in this case as well.

Updates for $p(\tau|\cdots)$: The full conditional distribution for τ is given by

$$p(\tau|\mathbf{Y}, \mathbf{X}, \mathbf{W}, \boldsymbol{\beta}) \propto \prod_{t=1}^{T} \prod_{i=1}^{N_t} \text{Normal}(y_t^i|\mathbf{x}_t^i(\mathbf{w}_t \odot \boldsymbol{\beta}), \tau_t) \text{Gamma}(\tau_t|a_\tau, b_\tau).$$

As the precision values are independent for each task, we can update τ_t independently and apply standard conjugacy results [23]. For $t = 1, ..., T$ we have

$$p(\tau_t|\mathbf{y}_t, \mathbf{X}_t, \mathbf{w}_t, \boldsymbol{\beta}) \propto \text{Gamma}\left(a + \frac{N_t}{2}, b + \frac{1}{2} \sum_{i=1}^{N_t} (y_t^i - \mathbf{x}_t^i(\mathbf{w}_t \odot \boldsymbol{\beta}))^2 \right).$$

Updates for $p(\boldsymbol{\Sigma}^{-1}|\cdots)$: The full conditional distribution for $\boldsymbol{\Sigma}^{-1}$ is given by

$$p(\boldsymbol{\Sigma}^{-1}|\mathbf{W}, \mathbf{V}, \nu) \propto \prod_{j=1}^{d} \text{Normal}(\mathbf{w}_{(j)}|0, \boldsymbol{\Sigma}^{-1}) \text{Wishart}(\boldsymbol{\Sigma}^{-1}|\mathbf{V}, \nu)$$

which is of a conjugate form. It follows that

$$p(\boldsymbol{\Sigma}^{-1}|\mathbf{W}, \mathbf{V}, \nu) \propto \text{Wishart}(\mathbf{V}_{N_t}, \nu_{N_t})$$

where $\nu_{N_t} = \nu + d$ and

$$\mathbf{V}_{N_t} = \left[\mathbf{V}^{-1} + \sum_{j=1}^{d} ((\mathbf{w}_{(j)} - \bar{\mathbf{w}})(\mathbf{w}_{(j)} - \bar{\mathbf{w}})^T) \right]^{-1}.$$

Updates for $p(\boldsymbol{\beta}|\cdots)$: The full conditional distribution for $\boldsymbol{\beta}$ is given by

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \mathbf{W}, \tau, \theta) \propto &\prod_{t=1}^{T} \prod_{i=1}^{N_t} \text{Normal}(y_t^i|\mathbf{x}_t^i(\mathbf{w}_t \odot \boldsymbol{\beta}), \tau_t) \\ &\prod_{j=1}^{d} \text{Bernoulli}(\beta_j|\theta). \end{aligned}$$

The Bernoulli distribution is a conjugate prior on the β_j parameter for each j. Thus we update $\boldsymbol{\beta}$ element-wise. Denote by $\boldsymbol{\beta}_{-j} \in \mathbb{R}^{d-1}$ the vector $\boldsymbol{\beta}$ with the jth element removed. Denote by $\boldsymbol{\beta}_j^z \in \mathbb{R}^d$ the vector $\boldsymbol{\beta}$ with the jth element set to $z \in \{0, 1\}$. To simplify the notation we define $g_j^z = \prod_{t=1}^{T} \prod_{i=1}^{N_t} \text{Normal}(y_t^i|\mathbf{x}_t^i(\mathbf{w}_t \odot \boldsymbol{\beta}_j^z), \tau_t)$ for each $j \in \{1, ..., d\}$ and $z \in \{0, 1\}$. Note that g_j^z does *not* depend on β_j for fixed z. We can then write

$$p(\beta_j|\mathbf{Y}, \mathbf{X}, \mathbf{W}, \tau, \theta) \propto \text{Bernoulli}\left(\frac{\theta g_j^1}{\theta g_j^1 + (1 - \theta) g_j^0} \right).$$

Updates for $p(\theta|\cdots)$: The full conditional distribution for θ is given by

$$p(\theta|\boldsymbol{\beta}, a_\theta, b_\theta) \propto \prod_{j=1}^{d} \text{Bernoulli}(\beta_j|\theta)\text{Beta}(\theta|a_\theta, b_\theta)$$
$$\propto \text{Beta}(a_\theta + n_0, b_\theta + n_1)$$

where $n_0 = \sum_{i=1}^{d} \beta_j$ and $n_1 = d - \sum_{i=1}^{d} \beta_j$, which follows from standard conjugacy results.

Inference for the BMSL model with sparsity on the precision matrix follows along similar lines, except for the updates for $\boldsymbol{\Sigma}^{-1}$ and λ. $\boldsymbol{\Sigma}^{-1}$ and λ are updated based on an efficient block Gibbs sampler proposed by [17]. Hence the inference method for BMSL with Bayesian graphical LASSO (BMSL-GL) is still fully Gibbs.

A Python implementation of both BMSL-W and BMSL-GL is available to the community on Github[1].

3. Related work

Many MTL approaches are posed in terms of estimating a task dependency matrix and the mappings from input to output space. In this context, a key modeling choice is whether structure or sparsity should be imposed on the tasks' coefficients and dependency matrix. Different applications may benefit from either choice.

A number of models assume that all tasks are related to each other. For instance, in [24–27] all tasks are taken to be related, sharing a set of latent basis tasks. However, this assumption may be inappropriate for certain applications, where different tasks may exhibit different degrees of relatedness. For these types of scenarios, several models have focused on approaches that learn task groupings (e.g., [10]) or that aim to accommodate and identify potential outlier tasks (e.g.,[28]). Similar to [24,29] proposes a model where task parameters are linear combinations of a small subset of latent basis tasks. However, in their model, not all tasks need be related to each other, and groupings are identified via similar selections of latent bases. In [10], a Dirichlet process is used to infer clusters across tasks. In [12], the authors propose a matrix-variate normal penalty with sparse inverse covariances to couple multiple tasks. In [30], the authors propose a Gaussian matrix generalized inverse Gaussian model for low-rank approximation to the task covariance matrix.

Here we propose two variants of the BMSL model. In BMSL-W, we impose a non-informative Wishart prior on the task covariance matrix, such that no structure is imposed on the tasks' relations *a priori*. However, because of the adopted Bayesian framework, if structure does exist in the data, it will be reflected in the posterior distribution of the covariance matrix due to the influence of the likelihood term. The BMSL-GL version of our model leverages a more informative prior, which assumes that there is sparsity in the task precision matrix, indicating a bias towards conditional independence across tasks.

Our models are similar in spirit to the work of [31,13], in which both the linear regression coefficients of the tasks, $\mathbf{W}$, as well as the structure dependence, $\boldsymbol{\Sigma}$, are learned from the data. Contrary to our model, their estimation is done via optimization with no prior distributions placed on model parameters. Many approaches proposed in the literature only allow for point estimates of model parameters. Due to model complexity only local optima are reached via optimization (e.g., [13,29,32,31]) or need to resort to approximate inference techniques or Expectation Maximization (EM) based solutions (e.g., [12,30]). In contrast the BMSL models proposed here have closed-form full conditional distributions for all parameters, so that full posterior inference can be obtained via Gibbs sampling.

More recently MTL has been applied in the medical domain. In both [33,34] MTL approaches are devised in order to deal with heterogeneous patient cohorts in medical data. The goal is to develop more personalized predictive models, so tasks are defined by homogeneous sub-populations that are learned through an unsupervised pre-processing step. Both models are non-linear and parameters are learned via non-convex optimization, thus rely on locally optimal point estimates. In contrast to our proposed approach, uncertainty quantification and interpretability remain a challenge for such models.

4. Experimental analysis

In this section, we assess the predictive performance of BMSL and investigate its ability to discover the latent covariance structure across tasks. Comparative performance of BMSL and related MTL/STL approaches are presented on multiple healthcare-related real datasets.

To be able to comparatively assess BMSL performance, we use the following methods as baselines:

- **LASSO** [35]: This is a single-task approach in which a sparse linear model is fitted to each task's data independently. Hence, there is no sharing of information across tasks. This is a frequentist STL method.
- **Bayesian Regression with Joint Feature Selection** (BJFS): This is a Bayesian MTL approach where the regression coefficients across the tasks are inferred independently, but the sparsity pattern is shared and inferred jointly across all tasks. This model is similar in spirit to [36,37], however it is a Bayesian model with closed-form full conditionals. It may be viewed as a simplification of the BMSL approach proposed in this paper, with a diagonal covariance matrix on the task regressors (instead of a Wishart or BGL), and a shared beta-Bernoulli prior on the feature selection matrix.
- **Across-tasks feature selection (ATFS)** [32]: This is a MTL approach that uses L2,1 penalization to impose shared sparsity across tasks. Conceptually, it is similar to the Bayesian regression method just mentioned, but based on the minimization of a regularized cost function (frequentist approach).
- **MTRL** [31]: The MultiTask Relationship Learning (MTRL) estimates task dependence along with task specific coefficients via a probabilistic model. A convex relaxation of the penalized joint log-likelihood function is proposed to ease optimization. This is also a frequentist approach.
- **AMTL** [38]: The Asymmetric MultiTask Learning (AMTL) also attempts to model task dependence, but unlike MTRL, it considers an asymmetric task dependence matrix. Therefore, the amount of information transferred from A to B can be different than B to A. Additionally, AMTL encourages tasks with lower residual on the training set to share more than tasks with larger residuals. The idea is that easier tasks regularize the learning of more difficult ones, but no vice versa. This is also a frequentist approach. We used the code made available by the authors on Github[2].
- **rMTFL** [39]: robust MultiTask Feature Learning (rMTFL) models task coefficient matrix $\mathbf{W}$ as the sum of two group lasso-penalized latent matrices $\mathbf{P}$ and $\mathbf{Q}$, which are responsible to capture shared features across tasks and automatically identify outlier tasks, respectively. No explicit tasks dependence structure is learned. This is also a frequentist model. We used the rMTFL code available in the MALSAR package[3].
- **MSSL** [13]: This MTL method jointly learns the task coefficients $\mathbf{W}$ and the precision matrix that captures the dependence across tasks. MSSL is similar to BMSL in the sense that it explicitly learns a precision matrix encoding the tasks' relatedness, but unlike BMSL it only obtains a point estimation. Also, BMSL imposes shared feature sparsity across all tasks, which is not present in the MSSL. This is also a frequentist approach. We used the code made available by the authors on Github[4].

[1] https://github.com/LLNL/bmsl.

[2] https://github.com/BlasterL/AMTL.

[3] https://github.com/jiayuzhou/MALSAR.

[4] https://github.com/andreric/mssl-python.

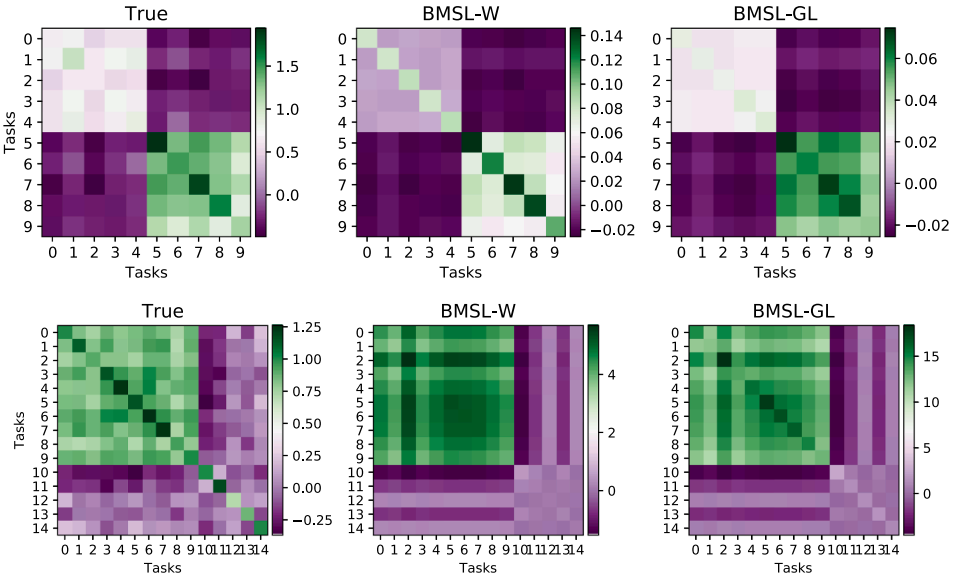

Fig. 1. True and average of posterior samples of covariance matrix, $\mathbf{\Sigma}$, for BMSL-W and BMSL-GL. **Top row**: Synthetic dataset 1: tasks dependence, including the two task groupings, is correctly inferred by both models. **Bottom row**: Synthetic dataset 2: single group of tasks is correctly identified, while decoupled tasks are left out, thus helping to avoid sharing information across unrelated tasks. [Best viewed in color].

In all experiments the hyper-parameters for non-Bayesian models were chosen by cross-validation, where 20% of the training data was used as validation. For BJFS, BMSL-W and BMSL-GL we used the same hyper-parameter values for all datasets. We used flat non-informative prior ($a_\tau = 0.01$, $b_\tau = 100$) for the precision of the model residue. The results are fairly insensitive to these hyperparameters, as long as the prior is flat and non-informative. We used an uniform, non-informative prior on θ ($a_\theta = 1$, $b_\theta = 1$, i.e., no sparsity *a priori* on the regression weights). It was observed that the model is able to infer the correct sparsity on the synthetic datasets, with these choice of hyperparameters. We choose $a_\lambda = 1$ and $a_\lambda = 0.1$, to encourage sparsity on the inverse-covariance matrix across task. The results were fairly insensitive to the choice of these hyperparameters.

Empirically, we observed that among all the BMSL hyperparameters, the results are most sensitive to the choice of the hyperparameter ν (degrees of freedom) for the Wishart prior. The choice of this hyperparameter can lead to numerical instability when computing the updated parameters for the posterior distribution, especially due to the necessary matrix inversions. We observed that for choices of ν close to the number of tasks (dimensionality of the W matrix), the model performance is fairly stable. Hence for all experiments we chose $\nu = T + 3$, where T is the number of tasks. The scale matrix V in the prior distribution is set to be an identity matrix, implying that independence across the tasks is assumed *a priori*.

The number of Gibbs iterations is set to 5000, and number of Gibbs burn-in samples is set to 3000. The data is standardized before the training and de-standardized for purpose of performance evaluation. The performance on each task is measured in terms of *normalized mean squared error* (nMSE), which is defined as:

$$nMSE = \frac{\|\mathbf{y}_{\text{true}} - \mathbf{y}_{\text{pred}}\|^2}{\text{std}(\mathbf{y}_{\text{true}})},$$

where the $\mathbf{y}$'s are vectors. To represent the performance on all tasks with a single metric, we use the average of nMSE over all tasks.

4.1. Synthetic data

We first create synthetic multitask datasets to examine whether BMSL can correctly recover the hidden tasks' dependence and coefficient sparsity structures. Two synthetic datasets are presented: 1) *Synth_1* which is composed of two groups of related tasks; and 2) *Synth_2* that contains one group of related tasks, while the others are completely independent. With these datasets we want to investigate if BMSL is capable of identifying groups of related tasks and, also equally important, when tasks are independent, thus helping to avoid negative transfer.

Dataset *Synth_1*:

This dataset consists of 10 regression tasks with 100 samples each. A task is a 30-dimensional regression problem in which the first 10 variables are independent of the output variable y. The 10 tasks are related in a group-wise manner: the first 5 tasks form a group and the last 5 tasks belong to another group. Tasks' coefficients in the same group are completely related to each other, while totally unrelated to tasks in a different group.

Tasks' data are generated as follows: weight vectors corresponding to tasks in group 1 (tasks 1 to 5) are $\mathbf{w}_{1:5} = \mathbf{w}_{g1} + 0.5 * \epsilon_1$, where $\epsilon_1 \sim \text{Normal}(0, 1)$ and $\mathbf{w}_{g1} \sim \text{Normal}(-0.5, 0.8 * \mathbf{I}_{20})$, which is used to make tasks in the same group related. Similarly, weight vectors for tasks in group 2 are obtained as: $\mathbf{w}_{6:10} = \mathbf{w}_{g2} + 0.8 * \epsilon_2$, where $\epsilon_2 \sim \text{Normal}(0, 1)$ and $\mathbf{w}_{g2} \sim \text{Normal}(0.5, 1.0 * \mathbf{I}_{20})$. Design and output matrices for the t-th task, $\mathbf{X}_t$ and $\mathbf{y}_t$, are generated as $\mathbf{X}_t = \text{Uniform}(0, 1)$ and $\mathbf{y}_t = \mathbf{X}_t\mathbf{w}_t + \text{Normal}(0, 1)$, for $t = 1, ..., 10$. Ten unrelated variables are then concatenated to the design matrix, $\mathbf{X}_t = [\text{Uniform}(0, 1), \mathbf{X}_t]$.

Dataset *Synth_2*:

For this dataset, fifteen 30-dimensional regression tasks are present. Ten tasks are related to each other forming a group, while the remaining 5 tasks are completely independent to all other tasks. Therefore, this dataset consist of a single group of tasks and a few other decoupled tasks. Tasks' data are created as follows: weight vectors corresponding to the 10 related tasks are created as $\mathbf{w}_{1:10} = \mathbf{w}_{g1} + \epsilon_1$, where $\epsilon_1 \sim \text{Normal}(0, 1)$ and $\mathbf{w}_{g1} \sim \text{Normal}(-1, 0.8 * \mathbf{I}_{30})$, which is used to make tasks in the same group related. The remaining 5 tasks are generated as $\mathbf{w}_{11:15} = \text{Normal}(0, \mathbf{I}_{30})$. Unlike *Synth_1*, all 30 features are relevant to the problem, that is, no sparsity is present. Design and output matrices for the t-th task, $\mathbf{X}_t$ and $\mathbf{y}_t$, are generated as $\mathbf{X}_t = \text{Uniform}(0, 1)$ and $\mathbf{y}_t = \mathbf{X}_t\mathbf{w}_t + \text{Normal}(0, 1)$, for $t = 1, ..., 15$.

To obtain a visual representation of the inferred precision matrix, Fig. 1 presents the mean precision matrix computed from the posterior samples for both synthetic datasets. For *Synth_1*, we notice that the two groups of tasks are clearly inferred from the data. While for *Synth_2*, BMSL correctly identifies the existing group of 10 tasks and also the set

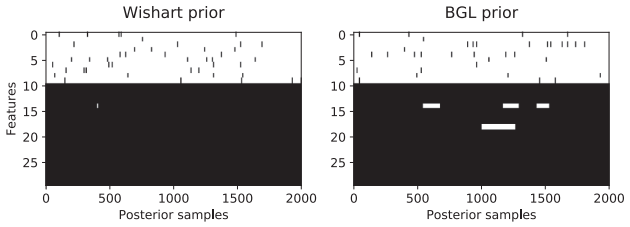

Fig. 2. Synthetic dataset 1: posterior samples of $\beta_j, j = 1, ...,d$ learned by BMSL on the synthetic dataset. A black tick mark indicates that the corresponding feature is selected. Note that the sparsity structure, with the first 10 variables being unrelated to the tasks, was correctly identified.

of decoupled tasks. Fig. 2 shows the posterior samples of $\boldsymbol{\beta}$ for each dimension in the *Synth_1* dataset case. BMSL correctly identified that the initial 10 variables are not related to the regression task.

4.2. Alzheimer's disease progression

We target the problem of assessing the progression of Alzheimer's disease (AD) from a multitask learning perspective. For this study we used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[5], which is a multi-site study that aims to improve clinical trials for the prevention and treatment of AD.

For the current study, we use structural MRI imaging data from 788 subjects, who are categorized in three diagnostic groups: cognitively normal (healthy individuals, n = 225), mild cognitive impairment (initial stages of the disease, n = 390), and Alzheimer's disease (n = 173). From the raw MRI images, a collection of 319 features were extracted by a team from the University of California at San Francisco, who performed cortical reconstruction and volumetric segmentation. For more information about the feature construction procedure, please refer to http://adni.loni.usc.edu/methods/mri-tool/ and [40].

Using the set of extracted features, we want to infer the current individual's cognition function state. Features with higher predictive power might be a potential biomarker for AD progression. Cognitive tests in the form of questionnaires and/or activities are conducted to measure memory, language, attention and other cognitive abilities which often reflect the symptoms of AD. The test provides a numerical score about individual's cognitive capability. The most common cognitive tests are: Alzheimer's Disease Assessment Scale cognitive total score (ADAS-cog), Mini Mental State Exam score (MMSE), Rey Auditory Verbal Learning Test (RAVLT) total score (TOTAL), RAVLT 30 min delay score (T30), and RAVLT recognition score (RECOG). Therefore, the goal is to infer the cognitive score, based on the structural MRI features.

From the multitask learning point of view, predicting a specific cognitive score for a set of subjects based on MRI features is considered a regression task. Hence, MTL approaches attempt to train regression models for all five cognitive scores simultaneously. For our experiments, we used the following hyper-parameter values for the non-Bayesian algorithms: LASSO ($\alpha = 0.01$), ATFS ($\rho_{L21} = 0.1$, $\rho_{L2} = 0.1$), MTRL ($\lambda_1 = 0.01$, $\lambda_2 = 0.1$), AMTL ($\mu = 1$, $\lambda = 0.01$), rMTFL ($\lambda_1 = 50$, $\lambda_2 = 300$), and MSSL ($\lambda_1 = 0.5$, $\lambda_2 = 1$). Training set consisted of 70% of the available data, and the remaining is used as a test set.

Table 1 reports the nMSE average and std for each task over 10 independent runs. The majority of the MTL methods outperform the STL approach for the problem. BMSL-GL provides significantly better predictions for ADNI than all contenders. We hypothesize that the advantage of BMSL-GL over BMSL-W may primarily be attributed to the inferred precision matrix (which is sparser than BMSL-W) and inferred feature selection matrix (which is fuller than BMSL-W).

[5] http://adni.loni.usc.edu.

Table 1
Comparison of predictive performance (mean and standard deviation of nMSE based on 10 independent runs) between the BMSL models and various STL and MTL baselines for each task of the ADNI dataset.

	ADAS	MMSE	RECOG	T30	TOTAL
LASSO	0.77 ± 0.01	0.73 ± 0.04	1.07 ± 0.02	0.86 ± 0.09	0.88 ± 0.13
ATFS	0.72 ± 0.01	0.79 ± 0.05	1.1 ± 0.05	0.9 ± 0.13	0.82 ± 0.11
MTRL	0.85 ± 0.05	1.06 ± 0.07	0.89 ± 0.03	0.76 ± 0.03	0.95 ± 0.03
AMTL	0.81 ± 0.07	1.14 ± 0.08	1.22 ± 0.11	1.03 ± 0.09	1.02 ± 0.09
rMTFL	0.91 ± 0.04	0.81 ± 0.02	1.17 ± 0.16	1.0 ± 0.10	1.10 ± 0.13
MSSL	0.61 ± 0.02	0.7 ± 0.03	0.91 ± 0.0	0.8 ± 0.07	0.74 ± 0.04
BJFS	0.58 ± 0.02	0.68 ± 0.02	0.83 ± 0.02	0.76 ± 0.08	0.7 ± 0.02
BMSL-W	0.55 ± 0.0	0.67 ± 0.01	0.82 ± 0.06	0.74 ± 0.06	0.67 ± 0.03
BMSL-GL	$\mathbf{0.52 \pm 0.01}$	$\mathbf{0.65 \pm 0.03}$	$\mathbf{0.8 \pm 0.05}$	$\mathbf{0.71 \pm 0.04}$	$\mathbf{0.64 \pm 0.03}$

Posterior samples of the variable $\boldsymbol{\beta}$ are shown in Fig. 3. Recalling that $\boldsymbol{\beta}$ captures the shared sparsity across tasks, BMSL-GL (with BGL prior) clearly led to a less sparse set of features, and consequently, a superior predictive performance. Compared to BMSL with BGL and Wishart, BJFS has an even sparser solution. It means that for the ADNI, most of the constructed MRI features are relevant for predicting the cognitive scores.

Covariance matrices for MSSL, BMSL-W, and BMSL-GL are shown in Fig. 4. Diagonal values were set to zero for ease of visualization. TOTAL cognitive score (task 4) presents a strong relation to ADAS-Cog (task 0), as it was captured by all three methods. TOTAL is also related to all other tasks to some extent. Overall, BMSL captured more dependencies across tasks in relation to MSSL. Taking into consideration that all cognitive scores (tasks) attempt to assess the latent subject's cognitive capability, it is expected that tasks show mutual dependence.

Unlike optimization-based MTL methods, such as LASSO, ATFS and MSSL, Bayesian approaches like BMSL provide uncertainty quantification about their predictions, which is essential for applications such as health-care. To exemplify this capability, Fig. 5 shows the posterior predictive distribution for a given held-out subject in the ADNI dataset. BMSL-GL distribution is reported along with point prediction obtained with three non-Bayesian models. We notice a higher uncertainty about its prediction for TOTAL and ADAS-Cog cognitive scores.

4.3. Parkinson's disease assessment

This dataset is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease (PD) recruited to a six-month trial of a tele-monitoring device for remote symptom progression monitoring [41]. For each patient, 18 features are collected: age, gender, and 16 jitter and shimmer voice measurements. For the categorical variable 'gender', we applied label encoding that converts genders into a numeric representation. Speech alteration is a common symptom of patients in more advanced stages of the disease, therefore it might be used as a marker for PD progression assessment.

Based on the set of features, the goal is to predict the motor *Unified Parkinson's Disease Rating Scale* (UPDRS) that is used to gauge the progress of Parkinson's disease in patients. Predicting the motor UPDRS score for each patient is a task, resulting in 42 regression tasks. The number of available samples per task range from 101 to 168. To investigate the ability of the methods on low training sample size regimes, we tested the methods on portions of the data with increasing sample sizes. The following hyper-parameter values for each algorithm were selected by cross-validation (20% of training set as validation): LASSO ($\alpha = 0.01$), ATFS ($\rho_{L21} = 0.1$, $\rho_{L2} = 0.1$), MTRL ($\lambda_1 = 0.001$, $\lambda_2 = 1e\text{-}3$), AMTL ($\mu = 1$, $\lambda = 1e\text{-}2$), rMTFL ($\lambda_1 = 10$, $\lambda_2 = 10$), and MSSL ($\lambda_1 = 0.5$, $\lambda_2 = 1$).

Table 2 reports the mean and standard deviation (std) of nMSE computed over 10 independent runs of each algorithm using exactly the same train/test sets. Single-task learning model LASSO has a

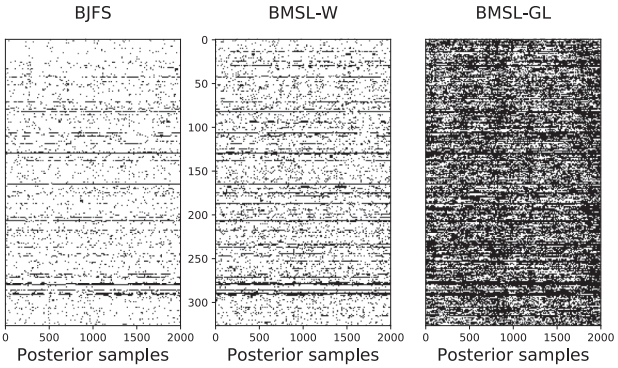

Fig. 3. Posterior samples for β learned on ADNI dataset from a single run of the algorithm. All other runs provided similar results. A black tick mark indicates that the corresponding feature is selected.

significantly higher nMSE if compared to the MTL methods. It is a clear indication that the tasks indeed have a relationship to be explored by the MTL approaches. BMSL-W and MSSL have high predictive skill across all training sample sizes.

Besides its predictive power, BMSL quantifies its uncertainty via the posterior distribution of its parameters. Fig. 7 reports the posterior distribution of the noise precision (inverse variance) τ, indicating the magnitude of the noise on each patient model. As we can see, many patients present much higher noise than others, which can be due to a noisier underlying signal, maybe related to an issue during the data collection, or meaning that a more flexible model possibly can better explain the data.

The mean covariance matrix over all posterior samples is presented in Fig. 6. The three models captured similar task dependence structure, although the absolute magnitudes vary from model to model. The block structure in all matrices suggests that there exist groups of related tasks, which in the current context implies patients with related speech-to-UPDRS mapping functions.

4.4. Cervical cancer screening compliance

Here we address the question of whether lifestyle can be predictive of women's compliance with cervical screening guidelines. This dataset comes from the Cancer Registry of Norway (CRN), and consists of data from 21,563 women of ages between 18 and 45 years. These women were randomly selected from the general female population in 2004 and 2011, and invited to fill out a questionnaire with information on education, marital status, smoking history, alcohol intake, sexual habits, contraceptive use, sexually transmitted diseases, and reproductive history. The questionnaire data were linked to the Norwegian Cervical Cancer Screening Program databases, obtaining for each person information on every cervical screening test between 1992 and 2016. For further information on the quality of the registry data, on the lifestyle survey study or the process to anonymize the data, see ([42–44])

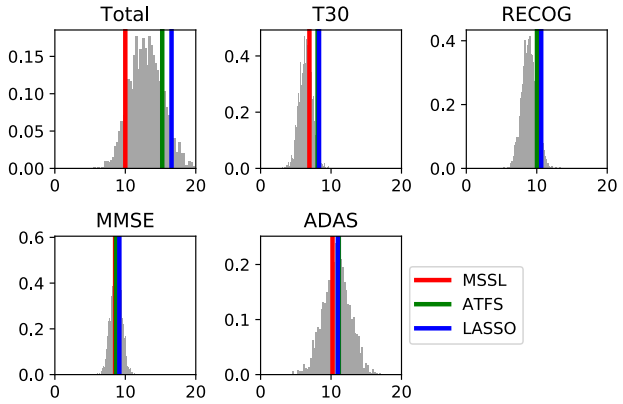

Fig. 5. BMSL-GL posterior predictive distribution for a held-out subject in the ADNI dataset. Predictions for TOTAL and ADAS-Cog have higher uncertainty. [Best viewed in color].

Table 2
Comparison of predictive performance (mean and standard deviation of nMSE based on 10 independent runs) between the BMSL models and various STL and MTL baselines for varying training set sizes of the Parkinson's Disease dataset. BMSL achieved low nMSE even for small sample sizes.

	Training sample size			
	30	40	50	60
LASSO	0.73 ± 0.1	0.63 ± 0.02	0.5 ± 0.02	0.49 ± 0.02
ATFS	0.51 ± 0.03	0.49 ± 0.03	0.45 ± 0.02	0.45 ± 0.02
MTRL	0.49 ± 0.08	0.47 ± 0.08	$\mathbf{0.44 \pm 0.06}$	$\mathbf{0.44 \pm 0.07}$
AMTL	0.51 ± 0.1	0.47 ± 0.09	0.46 ± 0.08	0.45 ± 0.08
rMTFL	0.57 ± 0.14	0.49 ± 0.11	0.47 ± 0.09	0.47 ± 0.10
MSSL	0.49 ± 0.01	0.47 ± 0.02	$\mathbf{0.44 \pm 0.01}$	$\mathbf{0.44 \pm 0.02}$
BJFS	0.54 ± 0.01	0.5 ± 0.02	0.47 ± 0.01	0.46 ± 0.02
BMSL-W	$\mathbf{0.48 \pm 0.01}$	$\mathbf{0.46 \pm 0.01}$	0.45 ± 0.01	$\mathbf{0.44 \pm 0.01}$
BMSL-GL	0.5 ± 0.03	0.48 ± 0.04	0.45 ± 0.01	0.45 ± 0.01

respectively.

In order to determine whether lifestyle can be predictive of women's compliance, we derive a *compliance score* feature, which is defined as the number of cervical screening tests taken. Because the compliance score depends on the length of time a woman has been recommended for screening, and thus her age, women are divided into groups based on their age at the time they answered the questionnaire. In the MTL context this implies that age groups (shown in Table 3) constitutes the tasks in the prediction problem.

Many variables in the CRN cancer screening dataset are categorical or ordinal. For these variables we used label encoding to create a numeric representation and so that they could be consumed by the STL and MTL methods. Label encoding assigns a integer from 0 to the number of categories in each variable. All other real-valued variables were kept as they were originally.

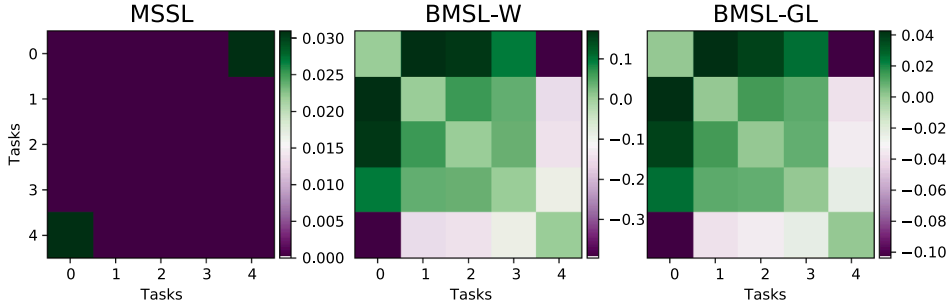

Fig. 4. Mean covariance matrix on ADNI dataset over all ten runs. White means a zero (sparse) entry. BMSL-GL and BMSL-W captured similar task dependence structure, whereas MSSL missed many dependencies. [Best viewed in color].

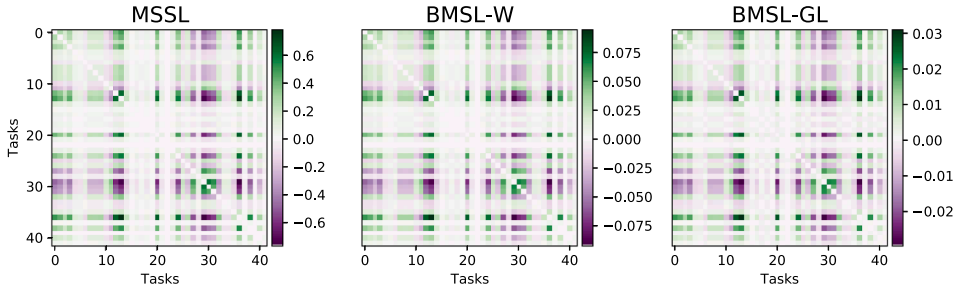

Fig. 6. Mean covariance matrix on PD dataset over all ten runs. BMSL-GL infers a sparser matrix compared to BMSL-W. [Best viewed in color].

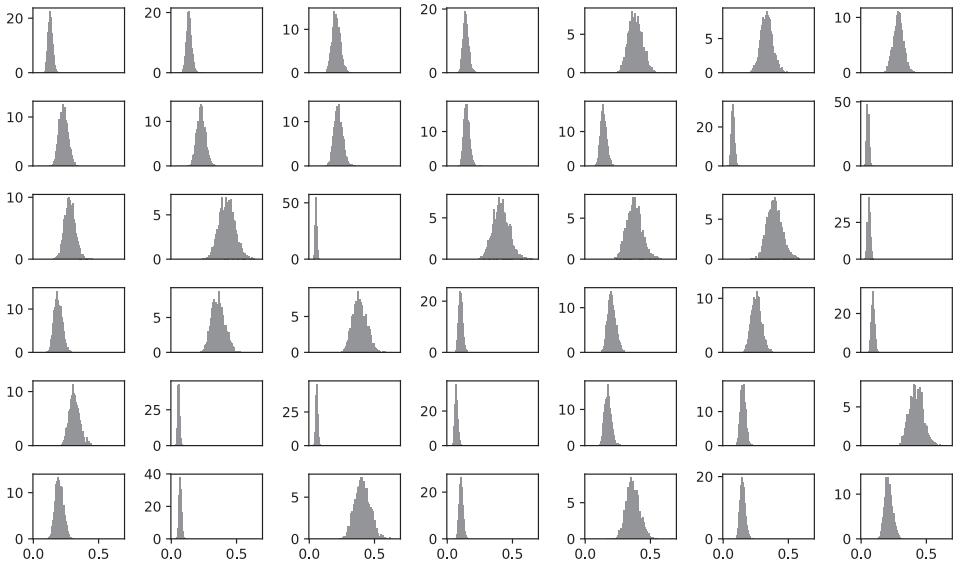

Fig. 7. Histograms of τ posterior samples for each patient in the PD dataset. Patients 1 to 42 from top-left to bottom-right.

Table 3
Groups of women categorized by age on the CRN dataset. Predicting cervical cancer screening compliance for each age group is a task.

Group/Task	1	2	3	4	5	6
Age	up to 24	25–29	30–34	35–39	40–44	45
# of samples	4169	4663	5698	4525	2254	254

Table 4
Comparison of predictive performance (mean and standard deviation of nMSE based on 10 independent runs) between the BMSL models and various STL and MTL baselines for varying training set sizes of the CRN dataset.

	Training sample size				
	50	60	70	80	90
LASSO	1.56 ± 0.34	1.4 ± 0.46	1.22 ± 0.12	1.24 ± 0.19	1.33 ± 0.48
ATFS	1.6 ± 0.38	1.43 ± 0.48	1.24 ± 0.13	1.27 ± 0.19	1.36 ± 0.49
MTRL	1.13 ± 0.11	1.09 ± 0.09	1.05 ± 0.07	1.05 ± 0.06	1.02 ± 0.06
AMTL	1.5 ± 0.20	1.42 ± 0.20	1.35 ± 0.20	1.29 ± 0.14	1.23 ± 0.10
rMTFL	1.02 ± 0.05	1.03 ± 0.04	1.02 ± 0.06	1.00 ± 0.04	1.02 ± 0.04
MSSL	1.05 ± 0.07	1.03 ± 0.05	$\mathbf{1.0 \pm 0.04}$	$\mathbf{1.0 \pm 0.05}$	1.01 ± 0.06
BJFS	1.02 ± 0.02	$\mathbf{1.02 \pm 0.03}$	1.01 ± 0.02	1.01 ± 0.02	1.0 ± 0.03
BMSL-W	$\mathbf{1.01 \pm 0.03}$	$\mathbf{1.02 \pm 0.03}$	$\mathbf{1.0 \pm 0.02}$	$\mathbf{1.0 \pm 0.02}$	$\mathbf{0.99 \pm 0.04}$
BMSL-GL	1.02 ± 0.03	1.05 ± 0.13	$\mathbf{1.0 \pm 0.02}$	$\mathbf{1.0 \pm 0.02}$	$\mathbf{0.99 \pm 0.04}$

In the current experiment, we used the following hyper-parameter values for each algorithm: LASSO ($\alpha = 0.01$), ATFS ($\rho_{L21} = 0.1$, $\rho_{L2} = 0.1$), MTRL ($\lambda_1 = 1e\text{-}3$, $\lambda_2 = 1e\text{-}2$), AMTL ($\mu = 1$, $\lambda = 1e\text{-}3$), rMTFL ($\lambda_1 = 100$, $\lambda_2 = 300$), and MSSL ($\lambda_1 = 0.5$, $\lambda_2 = 0.01$).

As shown in Table 4, BMSL shows superior prediction skill for small training sample sizes. All the MTL models, except for ATFS, achieve lower nMSE even for a relatively small sample size, compared to the number of parameters to be estimated.

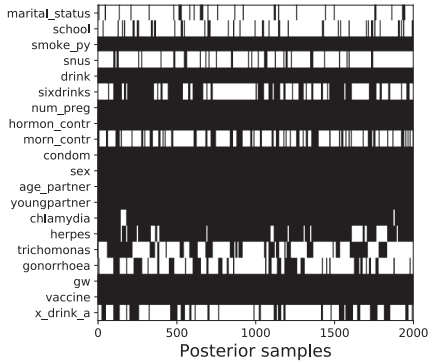

Fig. 8. Posterior samples for β learned on CRN dataset from a single run of the BMSL-GL. All other runs provided similar results. A black tick mark indicates that the corresponding feature is selected.

Fig. 8 shows the β posterior samples for all features, giving an indication of what features are relevant for predicting screening guidelines compliance, and Fig. 9 shows the posterior distribution of the weights of some of the features. Note that 'vaccine' (a binary feature indicating whether a woman has received the HPV vaccine) is included in the model virtually in 100% of the posterior samples. However, as show in Fig. 9, its coefficient is null in all but the younger age group. This is not surprising given that typically only younger women have been given the HPV vaccine. Note that 'school' (an ordinal variable that indicates the number of schooling years) is not a strong predictor of

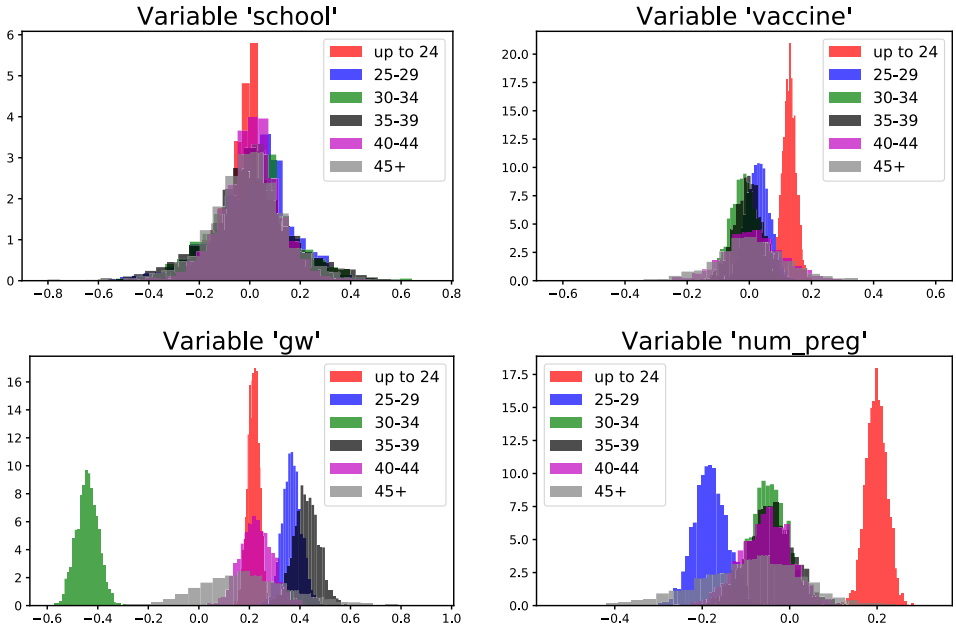

Fig. 9. Posterior distribution $\mathbf{w}$ posterior samples for 'school' (number of school years), 'vaccine' (binary, HPV vaccine), 'gw' (binary, presence of genital warts), and 'num_preg' (number of pregnancies), generated by BMSL-GL for the CRN dataset. [Best viewed in color].

compliance, which is understandable given that in this cohort more than 70% of the women have had more than 12 years of schooling. We hypothesize that in societies with lower levels of formal education, this feature would probably play a stronger role in predicting screening compliance.

Other variables, such as 'num_preg' (number of pregnancies) and 'gw' (having had genital warts), have posterior distributions that vary across groups, and interpretation is more challenging. The variable 'gw' indicates the presence of genital warts, which are lesions caused by the same type of virus (HPV, human papillomavirus) that causes virtually 100% of cervical cancers. The posterior distribution for 'gw' for all but one group has a positive mean. One might expect that women at a higher risk of sexually transmitted infections would seek their GP/gynecologist more often, and thereby be screened more often than women at lower risk.

Overall, in applications such as the current one, in which interpretability is important, inspection of the posteriors of the $\boldsymbol{\beta}$'s and $\mathbf{w}$'s is particularly useful.

5. Conclusion

Uncertainty quantification is always a desired element in any statistical/machine learning model, particularly in scientific applications. Bayesian approaches are suitable for this goal, naturally providing uncertainty quantification via posterior probabilities. In the multitask learning (MTL) setup, in which multiple tasks are jointly learned so that possible commonalities among the tasks can be explored by the model, uncertainty quantification is critical, particularly for healthcare applications.

We presented a fully Bayesian multitask learning model with shared sparsity pattern which also explicitly models the tasks relationship via Bayesian Graphical LASSO and Wishart distributions, referred to BMSL. The model uses conjugate priors and has closed-form full conditionals, thus an efficient Gibbs sampler can be derived. This is a significant benefit compared to other existing Bayesian MTL models that are non-conjugate and rely on approximate inference methods. In comparison with the vast majority of the MTL methods in the literature that are based on the optimization of a regularized cost function, BMSL provides predictive distributions rather than point predictions.

An extensive experimental evaluation was conducted on synthetic and real datasets from three distinct medical applications: Parkinson's disease assessment, Alzheimer's disease progression, and cervical cancer screening guideline compliance. In terms of nMSE, results showed that BMSL outperforms state-of-the-art MTL methods on the Alzheimers dataset and is competitive or slightly better on the Parkinsons and CRN datasets. It is to be noted that for clinical applications, interpretability and uncertainty quantification of the learned model is as relevant as the predictive performance. This is also true for many other scientific applications. It can help experts to better comprehend and also come up with new hypothesis about the underlying process, which can further be the topic future research. For example, Figures such as 7, 8 and 9 provide interesting insights (such as relative importance and relevance of the features for task prediction and uncertainty quantification of the regressor weights), which are not provided by many of the baseline models considered in this paper. However, compared to the frequentist models which perform point estimation, the proposed approach performs sampling to infer the posterior distributions, and is computationally much more expensive than the frequentist models.

Future research includes the extension of the current model to classification tasks. This will pose additional challenges due to lack of conjugate priors resulting from non-linear link functions. Development of efficient inference methods will be a challenge.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Caruana, Multitask learning, Machine Learn. 28 (1) (1997) 41–75, https://doi.org/10.1023/A:1007379606734.
[2] C. Widmer, G. Rätsch, Multitask learning in computational biology, ICML Workshop on Unsupervised and Transfer Learning, 2012, pp. 207–216.
[3] X. Wang, C. Zhang, Z. Zhang, Boosted multi-task learning for face verification with applications to web image and video search, IEEE Conference on Computer Vision and Pattern Recognition (2009) 142–149.
[4] P. Ray, L. Zheng, J. Lucas, L. Carin, Bayesian joint analysis of heterogeneous

genomics data, Bioinformatics 30 (10) (2014) 1370–1376.
[5] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: A deep learning approach, International Conference on Machine Learning (ICML), 2011, pp. 513–520.
[6] Y. Zhang, Q. Yang, A survey on multi-task learning, CoRR abs/1707.08114, 2017, pp. 1–20.
[7] E.V. Bonilla, K.M. Chai, C. Williams, Multi-task Gaussian process prediction, Advances in Neural Information Processing Systems, 2008, pp. 153–160.
[8] S. Ruder, An overview of multi-task learning in deep neural networks, CoRR abs/1706.05098, 2017, pp. 1 – 14.
[9] L. Jacob, J.-P. Vert, F.R. Bach, Clustered multi-task learning: A convex formulation, Advances in Neural Information Processing Systems, 2009, pp. 745–752.
[10] Y. Xue, X. Liao, L. Carin, B. Krishnapuram, Multi-task learning for classification with Dirichlet process priors, J. Mach. Learn. Res. 8 (Jan) (2007) 35–63.
[11] S. Guo, O. Zoeter, C. Archambeau, Sparse Bayesian multi-task learning, Advances in Neural Information Processing Systems, 2011, pp. 1755–1763.
[12] Y. Zhang, J. Schneider, Learning multiple tasks with a sparse matrix-normal penalty, Advances in Neural Information Processing Systems, 2010, pp. 2550–2558.
[13] A.R. Gonçalves, F.J.V. Zuben, A. Banerjee, Multi-task sparse structure learning with Gaussian copula models, J. Machine Learn. Res. 17 (33) (2016) 1–30.
[14] A. Agarwal, S. Gerber, H. Daume III, Learning multiple tasks using manifold regularization, Advances in Neural Information Processing Systems, 2010, pp. 46–54.
[15] X.J. Hunt, S. Emrani, I.K. Kabul, J. Silva, Multi-task learning with incomplete data for healthcare, arXiv preprint arXiv:1807.02442.
[16] L. Nie, L. Zhang, L. Meng, X. Song, X. Chang, X. Li, Modeling disease progression via multisource multitask learners: A case study with Alzheimers disease, IEEE Trans. Neural Networks Learn. Syst. 28 (7) (2017) 1508–1519.
[17] H. Wang, Bayesian graphical lasso models and efficient posterior computation, Bayesian Anal. 7 (4) (2012) 867–886.
[18] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.
[19] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
[20] C.M. Lynch, B. Abdollahi, J.D. Fuqua, A.R. de Carlo, J.A. Bartholomai, R.N. Balgemann, V.H. van Berkel, H.B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Informatics 108 (2017) 1–8, https://doi.org/10.1016/j.ijmedinf.2017.09.013.
[21] P.S. Windyga, Fast impulsive noise removal, IEEE Trans. Image Process. 10 (1) (2001) 173–179.
[22] L. Bar, A. Brook, N. Sochen, N. Kiryati, Deblurring of color images corrupted by impulsive noise, IEEE Trans. Image Process. 16 (4) (2007) 1101–1111.
[23] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, 2nd ed., Chapman and Hall/CRC, 2004.
[24] P. Rai, H. Daume III, Infinite predictor subspace models for multitask learning, International Conference on Artificial Intelligence and Statistics (AISTATS), 2010, pp. 613–620.
[25] G. Obozinski, B. Taskar, M.I. Jordan, Joint covariate selection and joint subspace selection for multiple classification problems, Stat. Comput. 20 (2) (2010) 231–252.
[26] Z. Huo, D. Shen, H. Huang, New multi-task learning model to predict Alzheimer's disease cognitive assessment, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016, pp. 317–325, , https://doi.org/10.1007/978-3-319-46720-7_37.
[27] F. Nie, Z. Hu, X. Li, Calibrated multi-task learning, Knowledge Discovery and Data Mining (KDD), 2018, pp. 2012–2021, , https://doi.org/10.1145/3219819.3219951.
[28] S. Yu, V. Tresp, K. Yu, Robust multi-task learning with t-processes, International Conference on Machine Learning, 2007, pp. 1103–1110, , https://doi.org/10.1145/1273496.1273635.
[29] H. Daumé III, A. Kumar, Learning task grouping and overlap in multi-task learning, International Conference on Machine Learning, 2013, pp. 1723–1730.
[30] M. Yang, Y. Li, Z. Zhang, Multi-task learning with Gaussian matrix generalized inverse Gaussian model, International Conference on Machine Learning, 2013, pp. 423–431.
[31] Y. Zhang, D.-Y. Yeung, A convex formulation for learning task relationships in multi-task learning, Conference on Uncertainty in Artificial Intelligence, 2010, pp. 733–742.
[32] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient l2,1-norm minimization, Conference on Uncertainty in Artificial Intelligence, 2009, pp. 339–348.
[33] H. Suresh, J.J. Gong, J.V. Guttag, Learning tasks for multitask learning: Heterogenous patient populations in the icu, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, ACM, New York, NY, USA, 2018, pp. 802–810, , https://doi.org/10.1145/3219819.3219930.
[34] J. Xu, J. Zhou, P.-N. Tan, FORMULA: FactORized MUlti-task LeArning for task discovery in personalized medical models, SIAM International Conference on Data Mining, 2015, pp. 496–504.
[35] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. Ser. B (Methodological) 58 (1) (1996) 267–288.
[36] T. Xiong, J. Bi, B. Rao, V. Cherkassky, Probabilistic joint feature selection for multi-task learning, SIAM International Conference on Data Mining, 2007, pp. 332–342.
[37] Y. Zhang, D.-Y. Yeung, Q. Xu, Probabilistic multi-task feature selection, Advances in Neural Information Processing Systems, 2010, pp. 2559–2567.
[38] G. Lee, E. Yang, et al., Asymmetric multi-task learning based on task relatedness and confidence, International Conference on Machine Learning, 2016, pp. 230–238.
[39] P. Gong, J. Ye, C. Zhang, Robust multi-task feature learning, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 895–903.
[40] X. Liu, A.R. Gonçalves, P. Cao, D. Zhao, A. Banerjee, Modeling Alzheimer's disease cognitive scores using multi-task sparse group lasso, Comput. Med. Imaging Graph. 66 (2018) 100–114, https://doi.org/10.1016/j.compmedimag.2017.11.001.
[41] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests, IEEE Trans. Biomed. Eng. 57 (4) (2010) 884–893.
[42] I.K. Larsen, M. Småstuen, T.B. Johannesen, F. Langmark, D.M. Parkin, F. Bray, B. Møller, Data quality at the cancer registry of Norway: an overview of comparability, completeness, validity and timeliness, Eur. J. Cancer 45 (7) (2009) 1218–1231.
[43] M.K. Leinonen, S.A. Hansen, G.B. Skare, I.B. Skaaret, M. Silva, T.B. Johannesen, M. Nygård, Low proportion of unreported cervical treatments in the cancer registry of Norway between 1998 and 2013, Acta Oncol. 57 (12) (2018) 1663–1670, https://doi.org/10.1080/0284186X.2018.1497296.
[44] G. Ursin, S. Sen, J.-M. Mottu, M. Nygård, Protecting privacy in large datasets-first we assess the risk; then we fuzzy the data, Cancer Epidemiol Biomarkers Prevdoi, 2017. https://doi.org/10.1158/1055-9965.EPI-17-0172.