Modeling skin and ageing phenotypes using latent variable models in Infer.NET

Anonymous Author(s)
Affiliation
Address
email

Abstract

We demonstrate and compare three unsupervised Bayesian latent variable models implemented in Infer.NET [2] for biomedical data modeling of 42 skin and aging phenotypes measured on the 12,000 female twins in the Twins UK study [7]. We address various data modeling problems include high missingness, heterogeneous data, and repeat observations. We compare the proposed models in terms of their performance at predicting disease labels and symptoms from available explanatory variables, concluding that factor analysis type models have the strongest statistical performance in this setting. We show that such models can be combined with regression components for improved interpretability.

This work is being performed in collaboration with the Department of Twin Research and Genetic Epidemiology (DTR) at King's College London. The DTR manages the largest UK adult twin registry of around 12,000 female monozygotic and dizygotic twins, established in 1992 [7]. The data has characteristics common to many biomedical applications, each of which we our able to address using our modeling framework.

1. *High missingness.* Many variables have up to 80% missing, and the level of overlap between phenotypes varies considerably. This level of missingness motivates Bayesian methods which are able to naturally deal with missingness, rather than attempting crude imputation procedures.
2. *Heterogeneous data.* The data contains continuous, categorical (including binary), ordinal and count data. We show in simulation experiments that using appropriate likelihood functions for each of these data types improves statistical power.
3. *Multiple observations.* Often the same underlying phenotype is recorded as multiple measurements, and the measurements may not be consistent. Allowing the model to combine these measurements into a single phenotype aids interpretability, improves statistical power and helps deal with the missingness problem.
4. *High dimensional.* The Twins UK database contains over 6000 phenotype and exposure variables, measured at multiple time points. Modern healthcare records are of the same nature. For a subset of 800 individuals we have 10,000 gene expression measurements in three different tissues, and the genotype of 600k Single Nucleotide Polymorphisms (SNPs).

Our modeling framework allows these issues to be straightforwardly and rigorously addressed, and provides an efficient inference platform using Variational Message Passing under the Infer.NET framework. Although the models we use all provide some form of dimensionality reduction, which is essential for the high dimensional nature of the data, we currently only analysis around 40 phenotypes of particular relevance to skin and aging. Scaling these models to handle the full dataset, including gene expression and genotype data, is ongoing research.

An attribute of the data that we have not fully explored how to model at this stage is that it is time series data. Most individuals in the study group have made multiple visits to be medically

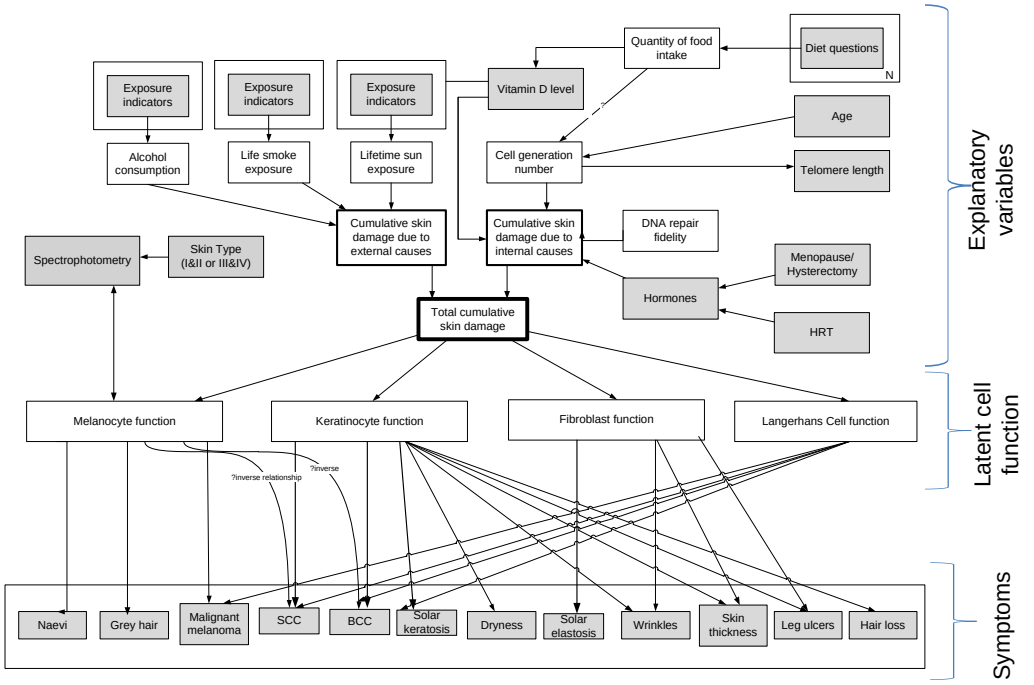


Figure 1: Schematic of key processes and variables involved in skin and aging conditions. This is not a probabilistic graphical model, but a representation of our prior knowledge of how the variables in the dataset are related. Gray boxes denote variables we have direct measurements of in at least some individuals, white boxes represent latent variables.

assessed, typically on a time frame of 3 to 5 years. Additionally many have answered surveys and self-assessment forms between these visits. Healthcare data is typically of this asynchronous time series nature. Currently we only use data from within three years of the most recent visit.

Another aspect of modeling phenotypic data is that there is an enormous amount of prior knowledge of the relationships between variables from decades of medical research and practice. Figure shows a schematic of the key processes involved in skin and aging, devised in collaboration with an experienced dermatologist. Although we are only using this prior knowledge in a very crude way at the moment (separating explanatory variables and symptoms) we intend to incorporate more structure into our models using this.

1 Models

We compare three Bayesian latent variable models. The first is a mixture model which attempts to cluster individuals. The second is a factor analysis model extended to allow different observed data types using various likelihood functions. The third is a combined regression and factor analysis model aimed at providing the expressive power of the factor analysis model and the interpretability of a regression model.

1.1 Mixture model

We assume that each individual sample was generated from one of K clusters. The variable $z_{nk} \in \{0, 1\}$ indicates whether individual n was generated from cluster k.

$$\pi \sim \text{Dir}(\alpha) \tag{1}$$
$$z_n \sim \text{Discrete}(\pi) \qquad \forall n \in \{1, \ldots, N\} \tag{2}$$
$$\tag{3}$$

The factor graph of this model is shown in Figure 2(a).

Continuous variables. For continuous variables y^c_{nd} each cluster has a mean $\mathbf{m}_k$ and variance $\mathbf{v}_k$, which are given normal and inverse-Gamma distributions respectively:

$$m_{dk} \sim N(m_{dk}; 0, 1) \qquad \forall d \in \{1, \ldots, D^c\}, k \in \{1, \ldots, K\} \tag{4}$$
$$v_{dk} \sim IG(v_{dk}; 1, 1) \qquad \forall d \in \{1, \ldots, D^c\}, k \in \{1, \ldots, K\} \tag{5}$$
$$y^c_{nd} \sim N(y^c_{nd}; m_{dz_n}, v_{dz_n}) \qquad \forall n \in \{1, \ldots, N\}, d \in \{1, \ldots, D^c\} \tag{6}$$

Binary variables. For binary variables y^b_d each cluster has a probability $\mathbf{p}_k$, which is given a uniform Beta prior.

$$p_{dk} \sim \text{Beta}(p_{dk}; 1, 1) \qquad \forall d \in \{1, \ldots, D^c\}, k \in \{1, \ldots, K\} \tag{7}$$
$$y^b_{nd} \sim \text{Bernoulli}(p_{dz_n}) \qquad \forall n \in \{1, \ldots, N\}, d \in \{1, \ldots, D^b\} \tag{8}$$

Categorical variables. For categorical variables y^c_d each cluster has a probability vector $\mathbf{p}_{dk}$, which is given a uniform Dirichlet prior.

$$\mathbf{p}_{dk} \sim \text{Dirichlet}(p_{dk}; \mathbf{1}) \qquad \forall d \in \{1, \ldots, D^c\}, k \in \{1, \ldots, K\} \tag{9}$$
$$y^b_{nd} \sim\sim Discrete(\mathbf{p}_{dz_n}) \qquad \forall n \in \{1, \ldots, N\}, d \in \{1, \ldots, D^b\} \tag{10}$$

1.2 Factor Analysis model

We assume each observation is generated as a linear combination of K underlying, latent factors.

$$g_{nd} = \mathbf{w}_{d:}\mathbf{s}_{n:} + m_d \tag{11}$$
$$\mathbf{w}_{d:} \sim N_K(\mathbf{0}, \Lambda^{-1}) \tag{12}$$
$$\Lambda \sim \text{Wishart}(\mathbf{10}, 0.1\mathbf{I}) \tag{13}$$
$$\mathbf{s}_{d:} \sim N_K(\mathbf{0}, \mathbf{I}) \tag{14}$$
$$m_d \sim N(m_d; 0, 1) \tag{15}$$

The factor graph for this model is shown in Figure 2(b). The hierarchical prior on $\mathbf{w}_{d:}$ is a form of Automatic Relevance Determination which helps suppress extra unnecessary features. We found this choice of prior superior in terms of predictive performance compared to no hierarchy or having an precision matrix for each observed dimension, which would encourage greater sparsity in an analogous way to using a student-T prior.

Continuous variables. Continuous variables are modeled simply by adding diagonal Gaussian noise to g_{nd}:

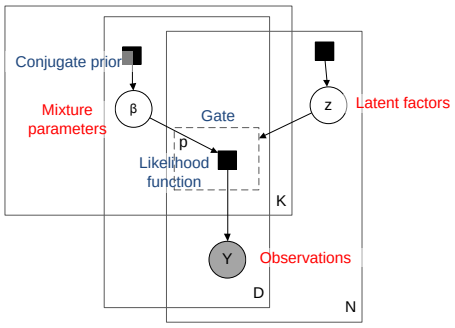
$$y^c_{nd} \sim N(y^c_{nd}; 0, \sigma^2_d) \tag{16}$$
$$\sigma^2_d \sim IG(\sigma^2_d; 1, 1) \tag{17}$$

where IG is the inverse-Gamma distribution.

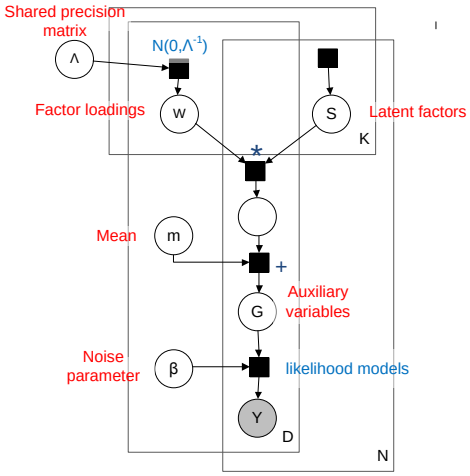
Binary variables. For binary variables we use a logistic link function $\sigma(x) = (1 + e^{-x})^{-1}$ in an analogous manner to logistic regression. We experimented with a probit link function but found little difference in empirical performance. The logistic link may be preferred in general due to its longer tails.

$$y^b_{nd} \sim \text{Bernoulli}(\sigma(g_{nd})) \tag{18}$$

In simulation studies we found that adding an additional noise term was unnecessary since the scale of g_{nd} effectively models varying noise levels. This component of our framework is closely related to [1] and [6] although we perform full Bayesian inference rather than maximum likelihood fitting.



(a) Mixture model. The conjugate priors and likelihood functions used for each data type are described in the text.



(b) Factor analysis model. For all data types a continuous auxiliary variable is the output from the factor analysis component. A different likelihood model/link function is used depending on the data type, as described in the text.

Figure 2: Factor graph representation of the mixture model and factor analysis model. Circles represent random variables. A white background represents a latent variable, whereas a gray background denotes an observed variable (or at least partially observed in this case). Solid rectangles represent plates (repetitive structures) and dashed rectangles represent gates [4], denoting an *if* or *switch* statement as used to build a mixture distribution.

Ordinal variables. Ordered categorical (ordinal) variables are common in biomedical data, for example, severity of a condition. Assume we have a Gaussian predictor variable g and an observed ordinal variable $y \in [1, .., J]$. Let the likelihood function be

$$P(Y = j|g) = \sigma(\tau_j - g) - \sigma(\tau_{j-1} - g) \qquad = \sigma(g - \tau_{j-1}) - \sigma(g - \tau_j) \tag{19}$$

where the logistic function $\sigma(x) = 1/(1 + e^{-x})$ and $\{\tau_j : j = 0..J\}$ are interval boundaries with $\tau_0 = -\infty, \tau_{j-1} < \tau_j, \tau_J = +\infty$. This aspect our of framework relates to the work in [5], although we use deterministic rather than MCMC based methods.

1.3 Regression-FA model

This model attempts to combine the statistical performance of the factor analysis model with greater interpretability. It is generally possible to split measurements into explanatory variables (for example: age, smoking, alcohol, sun exposure) and outcomes (e.g. heart disease, melanoma, wrinkles). It is of direct interest to known if there are (causal) interactions between these groups of variables. To achieve this, some of the factors from the factor analysis model are set to known explanatory variables. These are encoded as for standard regression: binary variables as $\{0, 1\}$ and a categorical variable y with C categorical is expanded into $C - 1$ variables, where $y_c = \mathbb{I}[y = c + 1]$.

1.4 Two layer model.

We often have multiple variables representing a single underlying phenotype. For example, whether an individual is undergoing Hormone Replacement Therapy (HRT) is known to effect their skin, so this is an important explanatory variable to include in the model. However, there are four different variables in the dataset since this question was asked on different questionnaires. We approach this problem by instantiating a latent variable representing the "true" value of this phenotype. The repeat observations are then given some probability conditional on the value of the latent variable. For categorical variables these will simply be conditional probability tables, each row of which is given

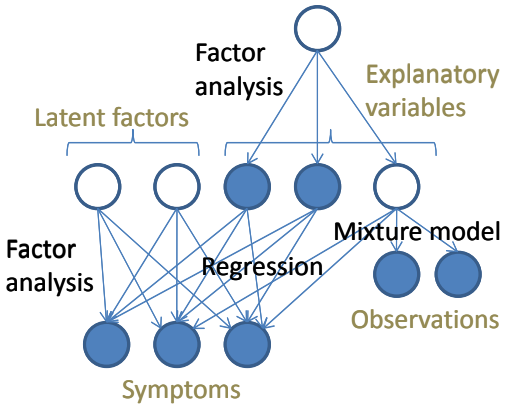


Figure 3: The factor analysis-regression model with the two layer summarization of latent exposures. We show the Directed Acyclic Graph (DAG) of the model here rather than the full factor graph for clarity.

a Dirichlet prior:

$$P(u_r|y) = \text{Discrete}(u_r|\pi_y^r) \tag{20}$$

$$\pi_y^r \sim \text{Dirichlet}(1, ..., 1) \tag{21}$$

where u_r is the r-th measurement relating to a particular phenotype, y is the true underlying binary value, and π_y^r is a probability vector. The "true" phenotype will have a Beta variational posterior, and can be used as an output straightforwardly in the mixture model, using the logistic link function as for observed binary variables in the factor analysis model, or even as an explanatory variable in the regression model. All these options are supported by Infer.NET [2] using Variational Message Passing [8].

2 Results

We present some initial results on synthetic and real data.

2.1 Synthetic data

We have validated the models and inference code on various synthetic data tasks. Due to space limitations we cannot document all of these tests here, but give one example. Consider an ordinal regression problem, with 5 ordinal output values, $P = 20$ observed explanatory variables and sample size N. The explanatory variables and regression coefficients are drawn from independent standard normals. The intervals τ are set as follows: $\tau_j = j - J/2$. The likelihood function described in Section 1.2 for ordinal data is used for both data generation and inference. Note that this is a simple instance of the regression-FA model of Section 1.3. Given synthetic data we measure the algorithm's ability to infer the vector of regression coefficients, in terms of correlation with the true value. Figure 2.1 shows the results for different sample sizes and three different models: 1. EP Ordinal Probit Regression (uses the Expectation Propagation (EP) algorithm [3], and the probit link function rather than logistic) 2. VMP ordinal logistic (our proposed model for this data type) 3. EP linear (again uses the EP algorithm but with a Gaussian likelihood function). The results highlight the value of using the appropriate likelihood function rather than just modeling all data as Gaussian. The performance of EP and VMP on this problem seems very similar, so we use VMP as it is able to handle the factor analysis and mixture components that we require, unlike EP.

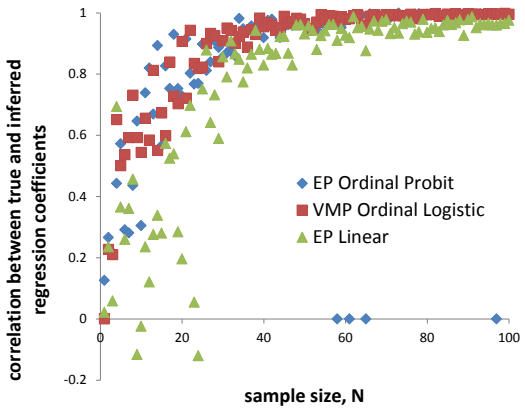


Figure 4: Synthetic data test.

2.2 Real data

We currently focus on a subset of around 40 variables across 3000 of the individuals with the least missing data. We use imputation performance to assess the fit of the proposed models to this data. For a randomly chosen 10% of individuals we treat symptoms (e.g. skin cancer, wrinkles) as missing, but leave the explanatory variables (e.g. age, smoking, sun exposure), and use the model to infer a predictive posterior over the held out values. The likelihood of the true values under the predictive posterior gives a measure of how well the model is fitting the data which is robust to overfitting. Figure 2.2 shows the imputation performance (higher is better) for the three models with different numbers of factors or mixture components. The variation shown by the box plots comes from taking a different 10% held out set 10 times.

The mixture model shows improved performance up to around five mixture components. More components do not seem to help, but it is encouraging to see that using our Bayesian approach overfitting still does not occur. The factor analysis model has generally superior performance to the mixture model, suggesting that this is a more appropriate model for this type of data. The factor analysis again seems to perform best with five factors. We are currently investigating the rapid jump in performance from 3 to 4 factors, since it is surprising that the second and third factors do not seem to contribute much. This may be an initialization or message passing schedule problem. The regression-FA model has predictive performance close to but not quite as high as the factor analysis model. Only three factors are required by this model, fewer than for the FA model, which is to be expected since the explanatory variables can be used directly in the regression, rather than via factors. For example in the factor analysis model we find one factor which is effectively the age of the individual, whereas in the regression model age is used directly. The regression-FA should have similar expressive power to the factor analysis model, so the slight decrease in performance relative to the factor analysis model may be attributable to being stuck in a local minimum, not using enough factors to fill in missing explanatory variables (we used two, and plan to run experiments to find the optimal number), or an initialization issue. Since the regression-FA model is simpler to interpret the choice between the FA model and regression-FA is effectively one of statistical performance versus interpretability.

Although the factor analysis model may not be as obviously interpretable as the regression model the fitted FA model does imply a particular covariance structure for the variables. This is shown in Figure 2.2. Although these are preliminary results it is interesting to note certain strong correlations, such as between smoking and two out of the three skin cancer types.

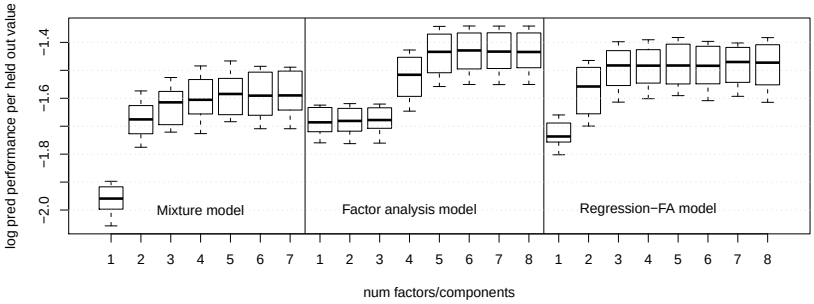


Figure 5: Predictive performance (higher is better) of the three models with different numbers of factors/mixture components.

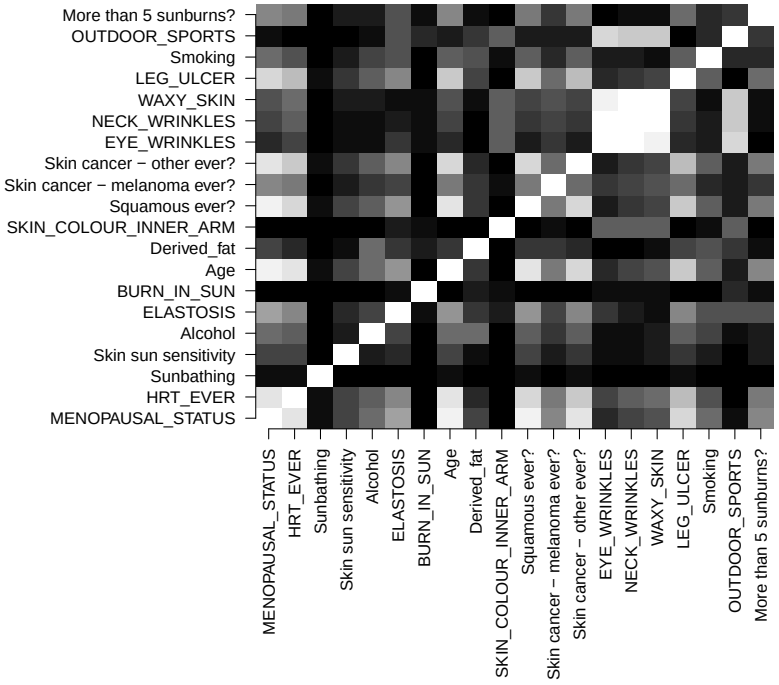


Figure 6: Correlation implied by the fitted factor analysis model. Lighter gray implies higher correlation. Notes on variable names: *Squamous* - squamous cell carcinoma is a type of skin cancer, *Derived fat* - a measure of fat metabolism in blood, *BURN_IN_SUN* - an ordinal 1-4 variable denoting how easily one burns in the sun, a standard measure of skin type. *HRT_EVER* - whether the individual has ever or is currently undergoing Hormone Replacement Therapy. *Sunbathing*, *HRT_EVER* and *MENOPAUSAL_STATUS* are derived from multiple observation as described in Section 1.4.

3 Discussion

We have described a biomedical data modeling framework we are currently constructing, with three different latent variable models. Our Bayesian model fitting allows missingness and noise to be naturally handled. Extending the flexibility of the Infer.NET package has allowed us to model and integrate a wide range of data types. The deterministic algorithms used allow us to scale these models to datasets far larger than would be feasible with MCMC methods. Infer.NET also allows us to write down more complex models that would otherwise be complex to keep track of, for example including the two layer model of Section 1.4 to reduce multiple observations to one underlying "true" phenotype, with associated uncertainty. Compared to a simple GLM type model, we can handle missingness in the explanatory variables, and confounding effects in both the explanatory variables and the symptoms by using factor analysis components.

Various issues remain to be resolved. The time series nature of the data is currently being ignored, which is clearly undesirable. Scaling these models to modern healthcare size datasets remains a challenge. Fortunately message passing algorithms lend themselves naturally to parallelization, an avenue we intend to explore in the future. If such a system were to be employed in a real world situation, online learning would also be beneficial, so that new data could be incorporated as it is recorded. Although this work is preliminary, the results are encouraging and we believe our framework and its extensions should be valuable modeling tools for biomedical researchers and potentially one day be useful at the front line of health care provision.

References

[1] Jan de Leeuw. Principal component analysis of binary data by iterated singular value decomposition. *Comput. Stat. Data Anal.*, 50(1):21–39, 2006.

[2] T. Minka, J.M. Winn, J.P. Guiver, and A. Kannan. Infer.NET 2.3, 2009. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

[3] Tom Minka. Expectation propagation for approximate bayesian inference. In *Uncertainty in Artificial Intelligence*, 2001.

[4] Tom Minka and John Winn. Gates: A graphical notation for mixture models. *NIPS*, 2008.

[5] Kevin M. Quinn. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, 12:338–353(16), 2004.

[6] Andrew Schein, Lawrence Saul, and Lyle Ungar. A generalized linear model for principal component analysis of binary data. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[7] Tim D. Spector and Alex J. MacGregor. The st. thomas' uk adult twin registry. *Twin Research*, 5:440–443(4), 1 October 2002.

[8] John Winn, Christopher M. Bishop, and Tommi Jaakkola. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.