Modelling skin/ageing phenotypes with latent variable models in Infer.NET

David Knowles[1] University of Cambridge
Leopold Parts Wellcome Trust Sanger Institute
Daniel Glass King's College London
John M. Winn Microsoft Research Cambridge
[1]dak33@cam.ac.uk

Abstract

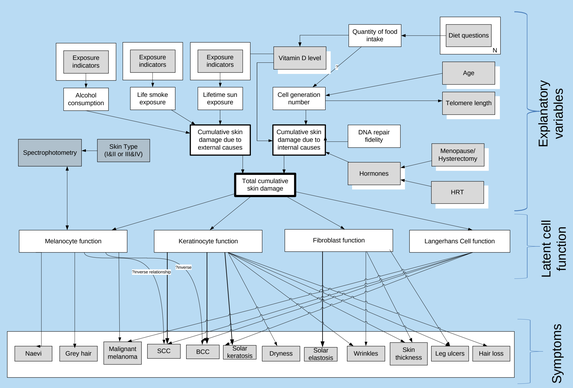
We demonstrate and compare three unsupervised Bayesian latent variable models implemented in Infer.NET [1] for biomedical data modeling of 42 skin and ageing phenotypes measured on the 12,000 female twins in the Twins UK study [2].

Data characteristics

Like many biomedical applications:

1. *High missingness.* Many variables have up to 80% missing: Bayesian methods are able to naturally deal with missingness
2. *Heterogeneous data.* Continuous, categorical (including binary), ordinal and count data: using appropriate likelihood functions for each of these data types improves statistical power.
3. *Multiple observations.* Combine into a single phenotype: aids interpretability, improves statistical power and helps with missingness.
4. *High dimensional.* 6000 phenotype and exposure variables, measured at multiple time points: use dimensionality reduction

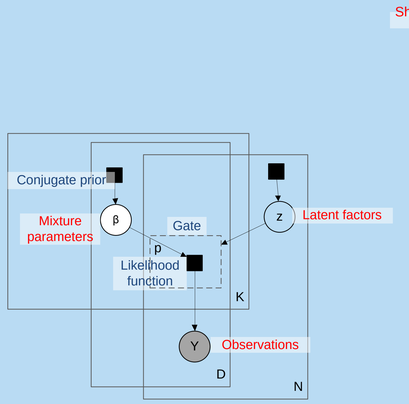
Medical expertise: prior knowledge



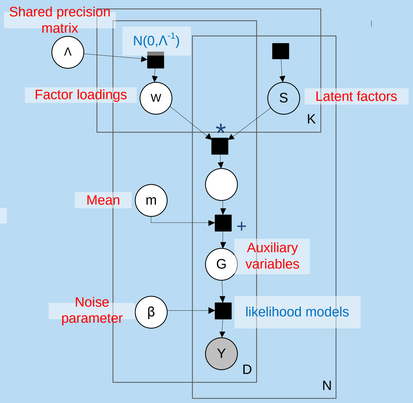
Key processes involved in skin and ageing, devised in collaboration with an experienced dermatologist. We use this prior knowledge in a very crude way at the moment (separating explanatory variables and symptoms) but we intend to use such knowledge to incorporate more structure into our models.

Models

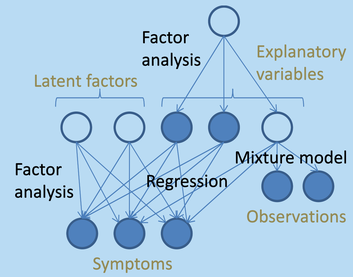
Factor graphs for the three proposed models.



1. Generalised mixture model. Clusters individuals. Suitable conjugate prior for each data type.

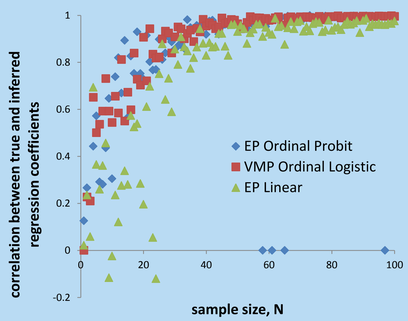


2. Generalised factor analysis model. Allows different observed data types using various likelihood functions

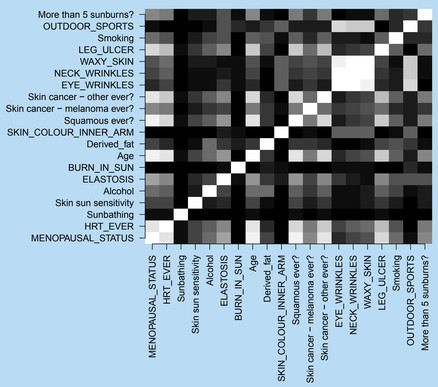


3. Combined regression and factor analysis model. Provides the expressive power of FA and interpretability of regression.

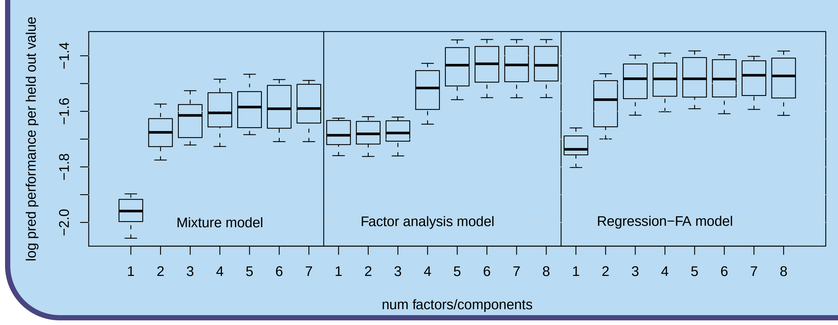
Results



Synthetic data test. Ordinal regression with 5 output values, $P = 20$ observed explanatory variables and varying sample size.



Correlation under the model. The fitted FA model implies a particular covariance structure for the variables of interest.



Imputation performance (real data). For a random 10% of individuals treat symptoms (e.g. skin cancer, wrinkles) as missing, but leave the explanatory variables (e.g. age, smoking, sun exposure), and infer the predictive posterior over the held out values.

Methods

We use Variational Message Passing under the Infer.NET framework. To support these models various factors were added to the framework: e.g. logistic regression, ordinal regression, "sum where".

Conclusions

1. Using appropriate likelihood models allows optimal integration of different data types
2. FA models have superior predictive performance to mixture models in this setting
3. Combining regression and FA components eases interpretability but at some cost to predictive performance (this may be due to scheduling problems or local minima)
4. Infer.NET allows us to use complex models

Future work

1. *Time series.* Multiple asynchronous visits, different phenotypes recorded each time.
2. *Scalability.* Although our message passing algorithms are efficient, scaling modern healthcare size datasets remains a challenge. Parallelization is a potential solution.
3. *Online learning.* This would allow new data could be incorporated as it is recorded.
4. *Nonlinearities.* We are currently experimenting with Gaussian Process and Mixture of Experts models to accommodate nonlinearity.

References

[1] T. Minka, J.M. Winn, J.P. Guiver, and D.A. Knowles. Infer.NET 2.4, 2010. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

[2] Tim D. Spector and Alex J. MacGregor. The St. Thomas' UK Adult Twin Registry. *Twin Research*, 5:440–443(4), 1 October 2002.

Funding

DK was supported by Microsoft Research through the Roger Needham Scholarship at Wolfson College, Cambridge.