Multiple Quantitative Trait Analysis in Statistical Genetics with Bayesian Networks

Marco Scutari

m.scutari@ucl.ac.uk
Genetics Institute
University College London

April 9, 2014

Gaussian BNs, between Classic and Modern Statistics

Bayesian networks (BNs) represent a flexible tool for quantitative [9], qualitative and causal [13] reasoning, and are one of the building blocks used to specify complex models and Monte Carlo inference techniques in machine learning [11].

However, BNs can also be approached from a perspective that is much closer to that of classic multivariate statistics by considering Gaussian Bayesian networks (GBNs):

- they allow the derivation of many closed form results because of the favourable properties of the multivariate normal distribution;
- they are related to such classic techniques as linear regression and covariance matrix decomposition;
- and they can be used to extend these techniques beyond their original scopes and definitions.

They have widespread applications in life sciences [12] and, as mentioned by Jean-Baptiste, in the upcoming [5, 17]

Gaussian Bayesian Networks (GBNs)

GBNs use a DAG G to represent the dependence structure of the multivariate distribution of $\mathbf{X} = \{X_1, \ldots X_p\}$ under the following assumptions [9]:

1. $\mathbf{X}$ has a multivariate normal distribution; and
2. dependencies between the X_is are linear.

Under these assumptions COV$(\mathbf{X}) = \Sigma$ is a sufficient statistics for the GBN and:

1. if X_i and X_j are graphically separated in G (d-separation, [9]), then $\Omega_{ij} = (\Sigma^{-1})_{ij} = 0$; and
2. the local distribution associated with each X_i is a linear regression on the parents Π_{X_i} of X_i, i.e.:

$$X_i = \mu_{X_i} + X_j\beta_j + \ldots + X_k\beta_k + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_i^2).$$

Note that $\beta_j = -\Omega_{ij}/\Omega_{ii}$ in the above [3].

GBNs in Genetics and GBLUP

The baseline model for association and prediction in statistical genetics is the linear mixed model [4], rebranded as GBLUP (Genetic BLUP, [10]). It is typically fitted on a single phenotypic trait X_t at a time using a large number S of genetic markers $\mathbf{X_S} = \{X_{s_1}, \ldots, X_{s_S}\}$ (e.g. SNPs, in the form of 0/1/2 allele counts) from a genome-wide profile:

$$X_t = \boldsymbol{\mu} + Z_{\mathbf{S}}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \mathbf{u} \sim N(\mathbf{0}, \mathbf{K}\sigma^2_{\mathbf{u}})$$

where $\boldsymbol{\mu}$ is the population mean, $Z_{\mathbf{S}}$ is the design matrix for the markers, $\mathbf{u}$ are random effects, $\boldsymbol{\varepsilon}$ is the error term and $\mathbf{K}$ is the kinship matrix encoding the relatedness between the individuals. When $\mathbf{K}$ can be expressed in the form $\mathbf{X_S}\mathbf{X_S}^{\mathrm{T}}$, GBLUP can be shown to be equivalent to the Bayesian linear regression

$$X_t = \boldsymbol{\mu} + \sum_{i=1}^{S} X^*_{s_i}\beta_i + \boldsymbol{\varepsilon} \quad \text{with SNP effect prior} \quad \boldsymbol{\beta} \sim N\left(\mathbf{0}, \frac{\sigma^2_g}{S}\mathbf{I}\right),$$

for some transformation of the X_{s_i} [14, 15].

GBNs and Multivariate Extension of GBLUP

If we wish to model traits $X_{t_1}, \dots X_{t_T}$ using a design matrix $\mathbf{Z_S}$ from $X_{s_1}, \dots X_{s_S}$ genetic markers, GBLUP can be extended [8] as follows

$$\left[\begin{array}{c} X_{t_1} \\ X_{t_2} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{\mu}_{t_1} \\ \boldsymbol{\mu}_{t_2} \end{array}\right] + \left[\begin{array}{cc} \mathbf{Z_S} & \mathbf{O} \\ \mathbf{O} & \mathbf{Z_S} \end{array}\right] \left[\begin{array}{c} \mathbf{u}_{t_1} \\ \mathbf{u}_{t_2} \end{array}\right] + \left[\begin{array}{c} \boldsymbol{\varepsilon}_{t_1} \\ \boldsymbol{\varepsilon}_{t_2} \end{array}\right],$$

where $\mathbf{u}_{t_1}, \mathbf{u}_{t_2}$ are random effects and $\boldsymbol{\varepsilon}_{t_1}, \boldsymbol{\varepsilon}_{t_2}$ are error terms, both normally distributed with covariances

$$\mathbf{G} = \mathsf{COV}\left(\left[\begin{array}{c} \mathbf{u}_{t_1} \\ \mathbf{u}_{t_2} \end{array}\right]\right) = \left[\begin{array}{cc} \mathbf{G}_{t_1 t_1} & \mathbf{G}_{t_1 t_2} \\ \mathbf{G}^{\mathsf{T}}_{t_1 t_2} & \mathbf{G}_{t_2 t_2} \end{array}\right],$$

$$\mathbf{R} = \mathsf{COV}\left(\left[\begin{array}{c} \boldsymbol{\varepsilon}_{t_1} \\ \boldsymbol{\varepsilon}_{t_2} \end{array}\right]\right) = \left[\begin{array}{cc} \sigma^2_{t_1}\mathbf{I} & \sigma^2_{t_1 t_2}\mathbf{I} \\ \sigma^2_{t_1 t_2}\mathbf{I} & \sigma^2_{t_2}\mathbf{I} \end{array}\right].$$

GBNs can be shown to be equivalent to GBLUP by considering the joint distribution of traits and genetic markers (through the random effects), which leads to

$$\Sigma = \mathsf{COV}\left(\left[\begin{array}{c} X_{t_1} \\ X_{t_2} \\ \hline \mathbf{u}_{t_1} \\ \mathbf{u}_{t_2} \end{array}\right]\right) = \left[\begin{array}{c|c} \mathbf{Z_S G Z_S}^T + \mathbf{R} & \mathbf{Z_S G} \\ \hline (\mathbf{Z_S G})^T & \mathbf{G} \end{array}\right].$$

Assumptions for Genetic Data

In the spirit of commonly used additive genetic models [7, 10], we make some further assumptions on the GBN to obtain a sensible causal model:

1. traits can depend on SNPs (i.e. $X_{s_i} \to X_{t_j}$) but not vice versa (i.e. not $X_{t_j} \to X_{s_i}$), and they can depend on other traits (i.e. $X_{t_i} \to X_{t_j}, i \neq j$);
2. SNPs can depend on other SNPs (i.e. $X_{s_i} \to X_{s_j}, i \neq j$); and
3. dependencies between traits follow the temporal order in which they are measured.

Under these assumptions, the local distribution of each trait is

$$\begin{aligned} X_{t_i} &= \boldsymbol{\mu}_{t_i} + \Pi_{X_{t_i}} \boldsymbol{\beta}_{t_i} + \boldsymbol{\varepsilon}_{t_i} \\ &= \boldsymbol{\mu}_{t_i} + \underbrace{X_{t_j}\beta_{t_j} + \ldots + X_{t_k}\beta_{t_k}}_{\text{traits}} + \underbrace{X_{s_l}\beta_{s_l} + \ldots + X_{s_m}\beta_{s_m}}_{\text{SNPs}} + \boldsymbol{\varepsilon}_{t_i}, \quad \boldsymbol{\varepsilon}_{t_i} \sim N(0, \sigma^2_{t_i}\mathbf{I}) \end{aligned}$$

and the local distribution of each SNP is

$$X_{s_i} = \boldsymbol{\mu}_{s_i} + \underbrace{X_{s_l}\beta_{s_l} + \ldots + X_{s_m}\beta_{s_m}}_{\text{SNPs}} + \boldsymbol{\varepsilon}_{s_i}, \qquad \boldsymbol{\varepsilon}_{s_i} \sim N(0, \sigma^2_{s_i}\mathbf{I}).$$

Learning GBNs from Genetic Data

We used the R packages bnlearn [16] and penalized [6] to implement the following hybrid approach to GBN learning [18].

1. Structure Learning.
   - 1.1 For each trait X_{t_i}, use the SI-HITON-PC algorithm [1] and the t-test for correlation to learn its parents and children; this is sufficient to identify the Markov blanket $\mathcal{B}(X_{t_i})$ because of the assumptions on the GBN. The choice of SI-HITON-PC is motivated by its similarity to single-SNP analysis.
   - 1.2 Drop all the markers which are not in any $\mathcal{B}(X_{t_i})$.
   - 1.3 Learn the structure of the GBN from the nodes selected in the previous step, setting the directions of the arcs as discussed above. We identify the optimal structure as that which maximises BIC.
2. Parameter Learning. Learn the parameters of the local distributions using ridge regression.

The Importance of Preprocessing and Feature Selection

Even though SI-HITON-PC scales extremely well, structure learning is still $O(p^2)$. This makes data pre-processing crucial:

- we can remove SNPs that are nearly constant (i.e. one allele, the minor allele, is almost absent from the data);
- we can remove highly correlated SNPs, which would form dense clusters in G and increase model and computational complexity for little gain in explaining the traits; and
- we can remove the influence of population structure from the traits to reduce the number of spurious relationships in the GBN.

Using the Markov blankets for feature selection makes learning even simpler, because we learn the full GBNs from a small subset of the original variables.

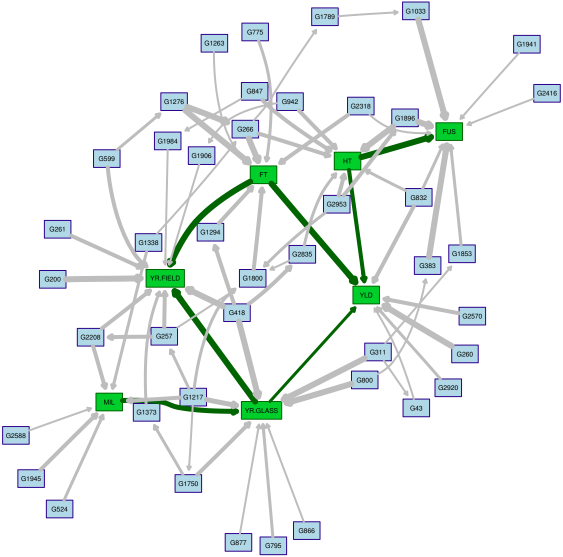
The Data: a MAGIC Wheat Population

The MAGIC data (Multiparent Advanced Generation Inter-Cross) include 721 wheat varieties, 16K markers and the following phenotypes:

- flowering time (FT);
- height (HT);
- yield (YLD);
- yellow rust, as measured in the glasshouse (YR.GLASS);
- yellow rust, as measured in the field (YR.FIELD);
- mildew (MIL); and
- fusarium (FUS).

Varieties with missing phenotypes or family information and markers with $> 20\%$ missing data, minor allele frequencies < 0.01 and COR > 0.95 were dropped. The phenotypes were adjusted for family structure via BLUP, leaving 600 varieties and 3.2K SNPs.

GBN from Model Averaging, $\alpha = 0.10$



50 nodes
(7 traits, 43 SNPs)

78 arcs, interpreted as putative causal effects

Thickness represents arc strength, computed as the frequency of each arc in the GBNs used in model averaging.

Type I error threshold for the test is $\alpha = 0.10$.

Predictive Performance

		YLD	FT	HT	YR FIELD	YR GLASS	MIL	FUS	Avg.
ENET	ρ_G	0.15	0.30	0.48	0.39	0.59	0.21	0.27	0.34
GBLUP	ρ_G	0.10	0.15	0.19	0.22	0.32	0.21	0.12	0.19
BN	ρ_G	0.20	0.29	0.46	0.37	0.60	0.12	0.22	0.32
$(\alpha = 0.01)$	ρ_C	0.38	0.29	0.45	0.44	0.62	0.13	0.33	0.37
BN	ρ_G	0.18	0.27	0.46	0.39	0.61	0.12	0.25	0.33
$(\alpha = 0.05)$	ρ_C	0.34	0.27	0.45	0.44	0.63	0.14	0.32	0.37
BN	ρ_G	0.18	0.28	0.45	0.40	0.62	0.13	0.25	0.33
$(\alpha = 0.10)$	ρ_C	0.34	0.28	0.45	0.45	0.63	0.14	0.31	0.37

ρ_G = predictive correlation given all SNPs in the model.
ρ_C = predictive correlation given putative causal effects identified by the BN.

Computed averaging 10×10-fold cross-validations, $\sigma = 0.01$ for traits and $\sigma = 0.005$ for the average. ENET is a single-trait elastic net penalised regression [19]; GBLUP is also in its classic single-trait form.

Inference and Interpretation

Conditional probability queries provide an ideal means for many different inferential tasks.

- Contrasting high and low values of traits makes it possible to identity SNPs tagging known genes; if

  $$|\,\mathrm{E}(X_{S_i}|X_{t_j} > c_{\mathrm{HIGH}}) - \mathrm{E}(X_{S_i}|X_{t_j} < c_{\mathrm{LOW}})|$$

  is large, it suggests that one allele of X_{S_i} is linked with low values of X_{t_i} and the other with high values. Several known genes were correctly identified this way (*Rht-D1b* for HT and FUS, *Ppd-D1* for FT, several genes for resistance to NIL and YR.GLASS).
- Confounding can be detected and accounting for; otherwise, we find that YLD increases with FUS (it doesn't when conditioning against HT, which is adjacent to both).
- Known causal relationship between traits can be quantified and validated by experts in the field (e.g. HT and FT affecting YLD).

Pros & Cons of GBNs

Pros:

- SNPs that are associated with more than one trait (pleiotropic effects) are included in the GBN even when association with just a single trait is detected; at that point they can be linked to all the relevant traits.
- GBNs model correlation between traits effectively, unlike single-trait models such as GBLUP and the elastic net.
- Confounding in genetic effects is reduced.
- The combination of a compact model and a graphical representation makes GBNs ideal for qualitative reasoning.
- Lots of literature of causal reasoning [2, 9, 13].

Cons:

- SNPs that are jointly associated but individually independent from a trait (epistatic effects) are not correctly modelled by the GBN because they violate the faithfulness assumption in SI-HITON-PC.
- Performing feature selection impacts the ability of predicting traits influenced by many small genetic effects (multigenic traits).

Conclusions

- GBNs provide a general modelling framework in statistical genetics, extending and subsuming existing models.
- Inference in GBNs in more flexible than in most of these models.
- The graphical component of a GBN is a valuable tool in disseminating results to non-statisticians.

This work is currently accepted for publication in *Genetics* as:

Scutari M, Howell P, Balding DJ, Mackay I (2014).
Multiple Quantitative Trait Analysis Using Bayesian Networks.
Genetics, to appear.

Acknowledgements

NIAB

Ian Mackay data preparation and general support

Phil Howell has run the MAGIC programme and collected disease scores and yield data

Nick Gosman involved in the running of the MAGIC programmes

Rhian Howells collected the flowering time data

Richard Hornsell performed crossing to create the MAGIC population and preparation of DNA

Pauline Bancept collected the glasshouse yellow rust data

UCL

David Balding my Supervisor

References

References I

[1] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Xenofon.
Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation.
J. Mach. Learn. Res., 11:171–234, 2010.

[2] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter.
Probabilistic Networks and Expert Systems.
Springer-Verlag, New York, 2007.

[3] D. R. Cox and N. Wermuth.
Multivariate Dependencies: Models, Analysis and Interpretation.
Chapman & Hall, Boca Raton, 1996.

[4] E. Demidenko.
Mixed Models: Theory and Applications with R.
Wiley, 2nd edition, 2009.

[5] J.-B. Denis and M. Scutari.
Réseaux Bayésiens avec R : Élaboration, Manipulation et Utilisation en Modélisation Appliquée.
Pratique R. EDP, 2014.
In preparation. This is a French translation of "Bayesian Networks with Examples in R".

[6] J. J. Goeman.
penalized R package, 2012.
R package version 0.9-41.

[7] Y. Guan and M. Stephens.
Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems.
Annals of Applied Statistics, 5(3):1780–1815, 2011.

References II

[8] C. R. Henderson and R. L. Quaas.
Multiple trait evaluation using relatives' records.
J. Anim. Sci., 43:1188–1197, 1976.

[9] D. Koller and N. Friedman.
Probabilistic Graphical Models: Principles and Techniques.
MIT Press, Cambridge, 2009.

[10] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard.
Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.
Genetics, 157:1819–1829, 2001.

[11] K. P. Murphy.
Machine Learning: A Probabilistic Perspective.
MIT Press, 2012.

[12] R. Nagarajan, M. Scutari, and S. Lèbre.
Bayesian Networks in R with Applications in Systems Biology.
Use R! series. Springer, 2013.

[13] J. Pearl.
Causality: Models, Reasoning and Inference.
Cambridge University Press, 2nd edition, 2009.

[14] H.-P. Piepho.
Ridge Regression and Extensions for Genomewide Selection in Maize.
Crop Sci., 49(4):1165–1176, 2009.

References III

[15] H.-P. Piepho, J. O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo, and F. Technow.
Efficient Computation of Ridge-Regression Best Linear Unbiased Prediction in Genomic Selection in Plant Breeding.
Crop Sci., 52(3):1093–1104, 2012.

[16] M. Scutari.
Learning Bayesian networks with the bnlearn R package.
J. Stat. Soft., 35(3):1–22, 2010.

[17] M. Scutari and J.-B. Denis.
Bayesian Networks with Examples in R.
Chapman & Hall, 2014.
In print.

[18] M. Scutari, P. Howell, D. J. Balding, and I. Mackay.
Multiple Quantitative Trait Analysis Using Bayesian Networks.
Genetics, 2014.
Submitted.

[19] H. Zou and T. Hastie.
Regularization and variable selection via the elastic net.
J. Roy. Stat. Soc. B, 67(2):301–320, 2005.