Learning Bayesian networks given a data set consisting of samples that are not independent and identically distributed

Jessica Kasza Gary Glonek Patty Solomon

School of Mathematical Sciences, University of Adelaide

Keywords: Bayesian networks; high-dimensional data; complex mean structure; regulatory networks

ABSTRACT

Methods for the estimation of Bayesian networks, flexible frameworks allowing the representation of conditional independence relationships of sets of variables, typically require a data set that consists of independent and identically distributed samples. Often the data set available will be more complex, containing information on exogenous variables thought to affect the variables of interest. We consider score-based methods for learning Bayesian networks, and demonstrate that the use of a score metric that fails to take account of such complexities may result in the estimation of a network with many spurious edges. We provide two new score metrics that do take account of these complexities, extending the utility of score-based methods for learning Bayesian networks to data sets that do not necessarily consist of independent and identically distributed samples. The first is a fully Bayesian metric, where a prior distribution is placed upon the effects of exogenous variables, and the second is based upon a restricted maximum likelihood approach to account for the effects of exogenous variables, which we call the residual approach. The utility of these metrics is demonstrated through their application to simulated data, and to a gene expression data set that contains data on covariates thought to affect gene expression levels. Finally, these two metrics are compared, and theoretical justification for the use of the residual approach instead of the Bayesian approach is provided.

INTRODUCTION

Learning the structure of Bayesian networks given high-dimensional data sets is an important area of research, to which much effort has recently been dedicated. The network structure, which encodes conditional independence relationships of sets of variables, is useful in the exploration of systems of interacting genes, known as genetic regulatory networks. Bayesian networks provide tools for the representation of regulatory networks that are able to cope with noisy expression data obtained through experiments.

Consider a p-dimensional normally distrbuted random vector $X \sim N(0, \Sigma)$, where each X_i represents the expression level of a gene. A Bayesian network $B = (G, \Theta)$, $\Theta = \{\theta_1, \ldots, \theta_p\}$, for X consists of two components: a directed acyclic graph associated with X, $G = (V, E)$, $V = \{X_1, \ldots, X_p\}$, $E \subset V \times V$, and a set of conditional distributions $\{f(x_i|x_{P_i}, \theta_i)\}$. The set P_i is the set of those variables in G such that $(j, i) \in E$, and is called the set of parents of X_i. The joint distribution of X can be decomposed according to the graph in the following way:

$$f(x|\Theta) = \prod_{i=1}^{p} f(x_i|x_{P_i}, \theta_i).$$

We consider the case where both the parameters Θ and the structure of the directed acyclic graph encoding the conditional independence relationships of X are unknown. Given a data set d consisting of n samples of each of the p random variables, we focus on learning the graphical structure of the Bayesian network. Methods for learning the underlying graph can be grouped into two main classes: constraint-based methods, and score-based methods.

Constraint-based methods work by considering the local structure of each random variable, and testing for conditional independence relationships of each random variable. The results of these hypothesis tests are then combined to form a valid directed acyclic graph. These methods can be sensitive to type I and II errors, particularly in the case of small sample size. We will not discuss constraint-based methods further, and refer interested readers to Chapter 18 of Koller and Friedman (2009).

Score-based methods work by attempting to maximise an appropriately-chosen score metric that describes how well different graphical structures encode the conditional independence relationships of X. An obvious choice of score metric is the likelihood of directed acyclic graphs on p nodes, however, the structure that maximises this likelihood is the complete directed acyclic graph, which encodes no conditional independence relationships. Hence, we will consider Bayesian score metrics, which avoid the problem of overfitting.

Following Geiger and Heckerman (2002) among others, the Bayesian score of a directed acyclic graph G for a random vector X is proportional to the posterior probability of the graph given the data set d:

$$\begin{aligned} S(G|d) &= p(G)p(d|G) \\ &= p(G)\int p(d|G,\Theta)p(\Theta|G)d\Theta. \end{aligned} \quad (1)$$

We focus on the second component of this score, the marginal model likelihood of the data given a graph G, where the density of d given G and Θ is assumed to be an $n \times p$-dimensional normal density, with mean vector 0 and covariance matrix $\Sigma \otimes I_n$.

When the data set d consists of independent and identically normally distributed samples, $\theta_i = \{\gamma_i, \psi_i\}$, and

$$x_i \mid x_{P_i}, \gamma_i, \psi_i \sim N_n\left(x_{P_i}\gamma_i, \psi_i I_n\right). \quad (2)$$

As can be seen in (1), in addition to the likelihood, a prior distribution over the space of parameters, $p(\Theta|G)$, is also required in the specification of the Bayesian score. As described by Geiger and Heckerman (2002), an Inverse-Wishart prior on Θ, or equivalent normal-inverse gamma priors on each θ_i, is required. The following priors are typically used in the calculation of the Bayesian score metric:

$$\begin{aligned} \gamma_i|\psi_i &\sim N_{|P_i|}\left(0, \tfrac{\psi_i}{\tau} I\right), \\ \psi_i^{-1} &\sim Ga\left(\tfrac{\delta+|P_i|}{2}, \tfrac{\tau}{2}\right), \end{aligned} \quad (3)$$

where τ and δ are user-specified parameters. Given these priors, the score metric of Equation 1 can be written as the product of the prior density on the space of directed acyclic graphs, $p(G)$, and p multivariate t densities.

The score metric can then be used in conjunction with an algorithm for moving through the space of directed acyclic graphs, to find the graph that maximises the score metric. The most commonly-used algorithm is greedy hill-climbing, see Chapter 18 of Koller and Friedman (2009), although there do exist others, such as high-dimensional Bayesian covariance selection, Dobra et al. (2004).

SCORE METRICS FOR NON-IID SAMPLES

The assumption in Equation 2 is only appropriate when the data set d consists of iid samples. Often, however, we will need to learn the structure of a Bayesian network given a more complex data set. The motivating example here is a data set consisting of expression levels of heat shock genes of grapes, where the grapes were sampled from three different vineyards, and temperatures at the times leading up to the picking of the grapes was recorded. Given the known relationships of these genes with temperature, and the disparities between vineyards, this data set cannot be considered to consist of independent and identically distributed samples. Due to possible common relationships with temperature and vineyard, failure to take account of these effects in learning Bayesian network structure may result in the inclusion of many spurious edges. Hence, there is a need to account for the presence of the exogenous variables of temperature and vineyard when learning graphical structure.

We will now suppose that contained within the data set d is information on exogenous variables thought to affect the expression levels of the genes considered. If Q is the $n \times m$ matrix containing data on the m exogenous variables, we suppose

$$x_i \mid x_{P_i}, \gamma_i, \psi_i, b_i \sim N_n\left(x_{P_i}\gamma_i + Qb_i, \psi_i I_n\right), \quad (4)$$

where b_i is the m-vector of the effects of the exogenous variable on x_i. In this specification, normality and linear dependence upon parents is retained, and more complex sampling schemes and the effects of exogenous variables are accounted for.

As the definition of the Bayesian score metric in (1) shows, a joint prior distribution for γ_i, ψ_i and b_i is required. Extension of the results in Geiger and Heckerman (2002) to the model in (4) indicates that the joint prior distribution for γ_i, ψ_i and b_i must have a normal-inverse gamma form. We use the priors for γ_i and ψ_i in (3), and suppose that

$$b_i|\phi_i \sim N_m(0, \phi_i I). \quad (5)$$

While there are several possibilities for the variance of the random effects ϕ_i, including placing a hyper prior on it and treating it as a constant, an extension of the results in Geiger and Heckerman shows that the only choice that results in a score metric with a closed form is taking $\phi_i = \upsilon^{-1}\psi_i$.

The score metric obtained through the use of this

prior density on the effects of exogenous variables can be shown to be

$$
\begin{aligned}
&S_B(G|d) = p(G)\prod_{i=1}^{p} f_\upsilon(x_i|x_{P_i}),\\
&x_i|x_{P_i} \sim t_{\delta+|P_i|}\left(0, \frac{\tau}{\delta+|P_i|}\Omega\right),\\
&\Omega = \left\{H_\upsilon - H_\upsilon x_{P_i}\left(\tau I + x_{P_i}^T H_\upsilon x_{P_i}\right)^{-1} x_{P_i}^T H_\upsilon\right\}^{-1}\\
&H_\upsilon = I - Q\left(\upsilon I + Q^T Q\right)^{-1} Q^T.
\end{aligned}
$$

Removal of random effects through analysis of residuals

Often the effects of exogenous variables are not of particular interest, and are included in the model to allow for structure to be learnt more accurately. However, in many situations, the effects of exogenous variables can be thought of as nuisances. Of course, ignoring such effects is not recommended. Instead, a non-parametric approach is developed for use in such situations.

This approach, instead of directly using the gene expression data, is based upon the use of linear combinations of residuals left over when the data is regressed upon the effects of the exogenous variables. We call this the "residual approach", and it is motivated by the restricted maximum likelihood procedure used in inference for mixed linear models; see for example Section 12.2 of Davison (2003), and Patterson and Thompson (1971).

This approach makes no assumptions about the form of the distribution of the random effects. Since no assumptions are made, the approach is correct no matter what the true distribution of the random effects may be. Hence, in situations when the assumption that $b_i \mid \psi_i \sim N_m\left(0, \upsilon^{-1}\psi_i I\right)$ is not satisfied, the residual approach provides a useful alternative to the BGeCM score metric. Further justification for this metric is provided below.

We consider an $n \times (n-m)$ matrix P such that

$$
\begin{aligned}
P^T Q &= 0,\\
P^T P &= I_{n-m},\\
PP^T &= I_n - Q(Q^T Q)^{-1} Q^T.
\end{aligned}
$$

Instead of considering x_i, $P^T x_i$ are used. It can be shown that the score metric obtained using the residual approach has the form:

$$
\begin{aligned}
&S_R(G|ld) = p(G)\prod_{i=1}^{p} f_R(P^T x_i|P^T x_{P_i}),\\
&P^T x_i|P^T x_{P_i} \sim t_{\delta+|P_i|}\left(0, \frac{\tau}{\delta+|P_i|}\Omega_R\right),\\
&\Omega_R = \left\{I - P^T x_{P_i}\left(\tau I + x_{P_i}^T PP^T x_{P_i}\right)^{-1} x_{P_i}^T P\right\}^{-1}.
\end{aligned}
$$

Limiting Behaviour of $S_B(G|d)$

Consideration of the limiting behaviour of S_B and S_R, the score metrics obtained under the Bayesian and residual approaches is illuminating. We consider the cases as υ approaches 0 and ∞. Note that large values of υ correspond to situations where the effects of exogenous variables are not *a priori* thought to contribute greatly to the variability of the X_i, while small values of υ correspond to situations where the variability of each X_i is thought to be largely driven by the variability of the exogenous variables.

Upon examination of the score metrics S_B it can be seen that as $\upsilon \to \infty$, S_B approaches the score metric obtained when no exogenous variables are present. This implies that when when the variances of the effects of exogenous variables are small, the Bayesian networks estimated using the full Bayesian approach will not be markedly different from those estimated when the exogenous variables are ignored.

When υ is small, on the other hand, H_υ is close to PP^T, and the Bayesian score metric is an improper score metric. In this situation, most of the variability in each X_i is due to variation in the effects of exogenous variables, and when S_B is used, all Bayesian networks will have a score of zero. Hence, for small values of υ, it is recommended that the residual approach be used instead of the Bayesian approach.

COMPARISON OF BAYESIAN AND RESIDUAL APPROACHES

Results above indicate that the residual approach may be thought of as an approximation, in a sense, to the Bayesian approach, and a quantification of the closeness of this approximation is important. The residual and Bayesian approaches are compared using the Kullback Leibler divergence to measure the distance between the posterior distributions obtained under each of these approaches. This comparison provides further theoretical justification for the residual approach by showing that the distance between the posterior densities obtained under

the Bayesian and residual approaches is generally small, and decreasing as sample size increases.

The divergence

The Kullback Leibler divergence, Kullback and Leibler (1951), between two posterior distributions $f(\theta|x)$ and $g(\theta|x)$ is given by

$$D(f,g) = \int \log\left\{\frac{f(\theta|x)}{g(\theta|x)}\right\} f(\theta|x)d\theta, \qquad (6)$$

which is always non-negative, and minimised when $f = g$. The Kullback Leibler divergence is a standard measure of the distance between two distributions, and can be thought of as a measure of the loss of information about θ when using $g(\theta|x)$ instead of using $f(\theta|x)$ to describe the posterior distribution of θ.

We consider the posterior densities of γ_i and ψ_i obtained under each of the Bayesian and residual approaches, and obtain the Kullback Leibler distance between these posteriors. The posteriors obtained under the Bayesian and residual approaches are shown in Appendix A. We will use $f_B(\gamma_i, \psi_i|x_i, x_{P_i})$ to denote the joint posterior of γ_i and ψ_i obtained under the Bayesian approach, and $f_R(\gamma_i, \psi_i|x_i, x_{P_i})$ to denote the joint posterior obtained under the residual approach.

The Kullback Leibler divergence between f_B and f_R, which we will denote by $D(f_B, f_R)$, is shown in Appendix B. If the graphical structure of the Bayesian network for X is known, the divergence between $f_B(\Sigma|X)$, the posterior density of Σ (the marginal covariance matrix of X) obtained under the Bayesian approach, and $f_R(\Sigma|X)$, the posterior obtained under the residual approach, is then available:

$$D_\Sigma\left\{f_B(\Sigma|X), f_R(\Sigma|X)\right\}$$
$$= \sum_{i=1}^{p} D\left\{f_B(\gamma_i, \psi_i|x_i, x_{P_i}), f_R(\gamma_i, \psi_i|x_i, x_{P_i})\right\}.$$

Typically the underlying graphical structure will not be known, but the divergence for a graph will be bounded by the divergence for the covariance matrix corresponding to the empty graph and the divergence corresponding to the covariance matrix of an arbitrary full graph.

The following theorem justifies the use of the residual approach as an approximation to the Bayesian approach:

Thm. 1 *As* $n \to \infty, D_\Sigma\left\{f_B(\Sigma|lX), f_R(\Sigma|X)\right\} \to 0.$

The proof of this theorem is omitted. The theorem tells us that as sample size increases, the posteriors obtained using the residual approach more closely approximate those obtained using the Bayesian approach.

EXAMPLE 1

Here we present an example that demonstrates the importance of accounting for exogenous variables in learning Bayesian network structure. Ten data sets were generated according to the following system of linear recursive equations

$$X_{ijk} = b_{ij} + \epsilon_{ijk}, \quad \epsilon_i \sim N(0, \psi_i),$$

$i = 1, \ldots, 100$, $j = 1, 2$, $k = 1, \ldots, 50$. The values of ψ_i were obtained by sampling from an Inverse Gamma$(1, 1/2)$ distribution, and are constant for each of the samples generated. Similarly, $b_i = (b_{i1}, b_{i2})^T$, $i = 1, \ldots, 100$, are fixed across data sets, obtained by sampling from

$$b_{ij} \sim N(0, \psi_i), \quad i = 1, \ldots, 100;\ j = 1, 2,$$

corresponding to $\upsilon = 1$. The non-zero mean structure of this example corresponds to two groups, and the true underlying directed acyclic graph is the empty graph.

Directed acyclic graphs were learnt for each of these 10 data sets, first using the original score metric, then using S_B and S_R, using the high-dimensional Bayesian covariance selection, Dobra et al. (2004), to move through the graph space. The mean and standard deviation of the number of edges obtained using each of these three score metrics is summarised in Table 1.

Table 1: Results of Example 1. The mean number of edges obtained when each score metric is used. Standard deviation in brackets.

Score metric	Number of edges
Original	117.2 (5.25)
S_B	1.0 (0.94)
S_R	1.7 (1.16)

These results indicate that when exogenous variables are ignored, the highest-scoring graph found tends to have many more edges than the true graph. When either S_B or S_R is used, a graph that is much closer to the true graph is obtained. This demonstrates the necessity of these two score metrics.

GRAPE GENE EXAMPLE

The data set considered here consists of fifty samples of gene expression levels for 26 grape genes.

The gene expression levels were derived from grape berry tissue samples grown in three different vineyards in South Australia. The 26 grape genes considered here are known to code for heat shock proteins (HSPs), Wang et al. (2004), which are responsible for protecting grapes against heat-induced stress. Accordingly, air temperature at each vineyard was recorded every hour from 5.5 hours to 0.5 hours before grapes were sampled.

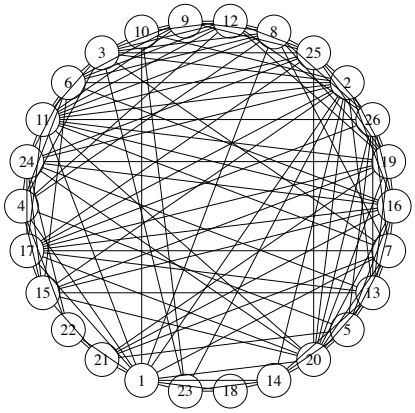
Learning the conditional dependence structure of the genes

Understanding the stress tolerance mechanisms of plants is important, and the heat shock protein network, as discussed by Kotak et al. (2007) and Wang et al. (2004), is very complex. Precisely how Hsps interact with one another and how they protect against heat stress is not yet completely understood, and here we seek to gain some insight into the heat shock protein network by examining the conditional dependence structure of these genes.

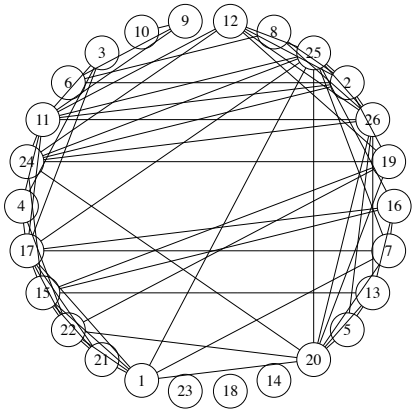
Given the known functions of the genes considered here and the climatic and geographic disparities between vineyards, it is important to account for the exogenous variables of temperature and vineyard in the estimation of a Bayesian network for the grape genes. If the expression levels of genes are strongly influenced by temperature, and temperature is ignored, then an observed conditional dependence relationship between two genes may simply be due to a common response to temperature, and not due to any regulatory mechanism between the genes.

Initially, no attempt is made to account for the effects of temperature and vineyard in the estimation of a Bayesian network for the genes. The Bayesian score metric is used in conjunction with the high-dimensional Bayesian covariance selection algorithm, Dobra et al. (2004), and the highest-scoring network found has 55 edges, the moralised version of which has 130 edges.The moralised version is shown in Figure 1(a).

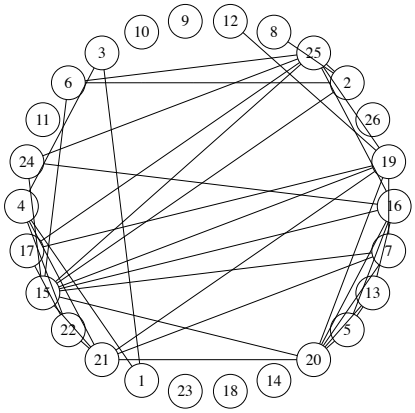
How best to include temperature and vineyard effects in the model for gene expression was investigated using linear regression models with forward and backward selection. The largest model considered for each gene contains separate intercepts for each vineyard, and terms for each of the 6 temperatures recorded before the picking of the grapes. Each of the vineyard or temperature coefficients is significant in at least one of the 26 regressions. We first consider the model including only vineyard effects, and then consider the model including vineyard and temperature effects.



(a)



(b)



(c)

Figure 1: The moral versions of the highest-scoring graphs obtained for the grape genes when (a) the effects of temperature and vineyard are ignored, and when the residual approach is used to include (b) vineyard effects, and (c) vineyard effects and main temperature effects.

We use the residual approach, with score metric S_R, in conjunction with the high-dimensional Bayesian covariance selection algorithm, to find high-scoring graphs when we include the effects of the exogenous variables of temperature and vineyard in the model. We use S_R since the score metric S_B uses the prior assumption that the effects of the exogenous variables are independent and identically distributed. In the case of the model that includes only temperature, such an assumption may be valid. However, it is probably not realistic to assume that the effects of temperature and vineyard are iid.

Figure 1(b) and (c) shows the highest-scoring graphs obtained using this approach. Immediately apparent is that as more of the variation of the expression levels is accounted for, the sparser the moral versions of the highest-scoring graphs obtained become. Due to the relationships of the considered genes with temperature, the most useful graph is that in Figure 1(c). Seven nodes in this graph are completely disconnected from all other nodes, indicating that the expression levels of these genes are independent of the expression levels of all other genes, once the effects of temperature and vineyard have been accounted for.

Three of these seven disconnected nodes correspond to Hsp 81 - an early response to dehydration. In discussions of the response of plants to heat stress, summarised in Kotak et al. (2007) and Wang et al. (2004), Hsp70 proteins (genes 1,2 and 3) and small HSPs (genes 5 and 13-26) are most focussed on. It is thought that the heat shock pathway consists of interactions between small Hsps, Hsp60, Hsp70, Hsp90, and Hsp100. Hsp 81 proteins (genes 9, 10 and 11), are not mentioned in connection with the heat shock protein network. The special role of Hsp81 proteins in *Arabidopsis thaliana* has been discussed in Yabe et al. (1994), in which it was noted that an increase in the expression level of Hsp81-1 is possibly caused by a regulatory pathway other than the heat shock pathway. Our analysis indicates that Hsp81 is not implicated in the protection of grapes against heat stress, and is evidence that the role of these genes requires further investigation.

Comparison of the Residual and Bayesian approaches

In the above section, S_R, the residual approach score metric, was used to estimate conditional dependence structure. We noted that S_R was appropriate when we included the effects of both vineyard and temperature, since these effects are probably not iid. However, when only vineyard is included in the model, use of the Bayesian score metric S_B may be more appropriate, since these effects can reasonably be thought of as iid.

Figure 2: Upper and lower bounds of the divergence for the marginal covariance matrix of the 26 grape genes, when vineyards are included as exogenous variables in the analysis. The solid line corresponds to the divergence for the empty graph, the dashed line to the divergence of an arbitrary full graph.

Using the Kullback Leibler divergence, we compare the residual and Bayesian approaches. This comparison can be interpreted as the amount of information lost about the marginal covariance matrix if the residual approach, instead of the Bayesian approach, is used. If $b_i = (b_{i1}, b_{i2}, b_{i3})^T$ represents the effects of the three vineyards on the expression of gene i, the prior density for b_i assumed under the Bayesian approach is $b_i \sim N_3(0, \upsilon^{-1}\psi_i I)$.

Since the true underlying graph is unknown, and υ is an unknown parameter, we calculate bounds for $D_\Sigma(f_B, f_R)$ as described above. The results are displayed in Figure 2. This figure indicates that for all considered values of υ, if the true underlying graph is sparse, as many biological networks are thought to be, the loss of information about the marginal covariance matrix when the residual approach is used will be minimal.

CONCLUSIONS

The score metrics presented here, S_R and S_B, extend the utility of score-based methods for learning Bayesian network structure to situations where the available data set does not consist of iid samples. Through consideration of the Kullback Leibler divergence between the posteriors obtained under each of these approaches, further justification for the residual approach was provided. The necessity

of score metrics that take account of the effects of exogenous variables was demonstrated. The use of these score metrics in the analysis of the grape gene example resulted in biologically plausible conclusions being drawn, and a conditional dependence structure that can be used in the generation of hypotheses about the interaction of grape heat shock genes. We recommend that the presence of exogenous variables in a data set should always be accounted for in the estimation of graphical structure, and when a score-based method for learning structure is taken, these score metrics provide methods for doing so.

Acknowledgments

The first and third authors acknowledge the support of Australian Research Council Discovery Project Grant DP110102028.

References

Davison, A. C. (2003), *Statistical Models*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Dobra, A., C. Hans, B. Jones, J. Nevins, and M. West (2004), Sparse graphical models for exploring gene expression data, *Journal of Multivariate Analysis 90*(1), 196–212.

Geiger, D. and D. Heckerman (2002, October), Parameter priors for directed acyclic graphical models and the characterization of several probability distributions, *The Annals of Statistics 30*(5), 1412–1440.

Koller, D. and N. Friedman (2009), *Probabilistic graphical models*. MIT Press.

Kotak, S., J. Larkindale, U. Lee, P. von Koskull-Döring, E. Vierling, and K.-D. Scharf (2007), Complexity of the heat stress response in plants, *Current opinion in plant biology 10*, 310–316.

Kullback, S. and R. A. Leibler (1951), On information and sufficiency, *The Annals of Mathematical Statistics 22*, 79–86.

Patterson, H. D. and R. Thompson (1971, December), Recovery of inter-block information when block sizes are unequal, *Biometrika 58*(3), 545–554.

Wang, W., B. Vinocur, O. Shoseyov, and A. Altman (2004, May), Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response, *Trends in Plant Science 9*(5), 244–252.

Yabe, N., T. Takahashi, and Y. Komeda (1994), Analysis of tissue-specific expression of *Arabidopsis thaliana* HSP90-family gene HSP81, *Plant cell physiology 35*(8), 1207–1219.

APPENDIX A

The posteriors obtained under the Bayesian approach are given by

$$\gamma_i|\psi_i, x_i, x_{P_i} \sim N_{|P_i|}\left(\mu_B, \psi_i\left(\tau I + x_{P_i}^T H_v x_{P_i}\right)^{-1}\right),$$
$$\mu_B = \left(\tau I + x_{P_i}^T H_v x_{P_i}\right)^{-1} x_{P_i}^T H_v x_i,$$
$$\psi_i|x_i, x_{P_i} \sim Inv.Gam.\left(\frac{\delta + n + |P_i|}{2}, \beta_B\right),$$
$$\beta_B = \frac{\tau}{2} + \frac{1}{2}x_i^T H_v x_i$$
$$- \frac{1}{2}x_i^T H_V x_{P_i}\left(\tau I + x_{P_i}^T H_v x_{P_i}\right)^{-1} x_{P_i}^T H_v x_i.$$

The posteriors obtained under the residual approach are given by

$$\gamma_i|\psi_i, x_i, x_{P_i} \sim N_{|P_i|}\left(\mu_R, \psi_i\left(\tau I + x_{P_i}^T PP^T x_{P_i}\right)^{-1}\right),$$
$$\mu_R = \left(\tau I + x_{P_i}^T PP^T x_{P_i}\right)^{-1} x_{P_i}^T PP^T x_i,$$
$$\psi_i|x_i, x_{P_i} \sim Inv.Gam.\left(\frac{\delta + n - m + |P_i|}{2}, \beta_R\right),$$
$$\beta_R = \frac{\tau}{2} + \frac{1}{2}x_i^T PP^T x_i$$
$$- \frac{1}{2}x_i^T PP^T x_{P_i}\left(\tau I + x_{P_i}^T PP^T x_{P_i}\right)^{-1} x_{P_i}^T PP^T x_i.$$

APPENDIX B

The Kullback Leibler divergence between f_B and f_R is given by

$$D(f_B, f_R) = \frac{1}{2}\log\left(\frac{\left|\tau I + x_{P_i}^T H_V x_{P_i}\right|}{\left|\tau I + x_{P_i}^T PP^T x_{P_i}\right|}\right)$$
$$+ \frac{1}{2}tr\left\{\left(\tau I + x_{P_i}^T PP^T x_{P_i}\right)\left(\tau I + x_{P_i}^T H_V x_{P_i}\right)^{-1}\right\} - \frac{|P_i|}{2}$$
$$+ \frac{\delta + n + |P_i|}{4\beta_B}\left(\mu_R - \mu_B\right)^T\left(\tau I + x_{P_i}^T PP^T x_{P_i}\right)\left(\mu_R - \mu_B\right)$$
$$+ \frac{\delta + n - m + |P_i|}{2}\log\left(\frac{\beta_B}{\beta_R}\right) + \log\left\{\frac{\Gamma\left(\frac{\delta+n-m+|P_i|}{2}\right)}{\Gamma\left(\frac{\delta+n+|P_i|}{2}\right)}\right\}$$
$$+ \frac{\delta + n + |P_i|}{2}\left(\frac{\beta_R}{\beta_B} - 1\right) + \frac{m}{2}\text{Digamma}\left(\frac{\delta + n + |P_i|}{2}\right).$$