Gaussian Join Tree classifiers with applications to mass spectra classification

Victor Bellón,
Universitat de Barcelona,Spain
IIIA-CSIC, Spain
bellon@ub.edu

Jesús Cerquides
IIIA-CSIC, Spain
cerquide@iiia.csic.es

Ivo Grosse
Martin-Luther-Universität, Germany
grosse@informatik.uni-halle.de

Abstract

Classifiers based on probabilistic graphical models are very effective. In continuous domains, parameters for those classifiers are usually adjusted by maximum likelihood. When data is scarce, this can easily lead to overfitting. Nowadays, models are sought in domains where the number of data items is small and the number of variables is large. This is particularly true in the realm of bioinformatics. In this work we introduce Gaussian Join Trees (GJT) classifiers to try to partially overcome this issue by performing exact bayesian model averaging over the parameters. We use two different mass spectra classification datasets for cancer prediction to compare GJT classifiers with those learnt by maximum likelihood.

1 Introduction

Supervised classification is a basic task in data analysis and pattern recognition. It requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of variables. There are numerous classifier paradigms, among which the ones based on probabilistic graphical models (PGMs) (Lauritzen, 1996), are very effective and well-known in domains with uncertainty.

A widely used assumption is that data follows a multidimensional Gaussian distribution(Geiger and Heckerman, 1994). This is adapted for classification problems by assuming that data follows a multidimensional Gaussian distribution that is different for each class, encoding the resulting distribution as a Conditional Gaussian Network (CGN)(Bøttcher, 2004). In (Larrañaga et al., 2006), Larrañaga, Pérez and Inza introduce and evaluate classifiers based on CGNs with a more detailed description in (Pérez, 2010). They analyze different methods to identify a Bayesian network structure and a set of parameters such that the resultant CGN performs well in the classification task. In (Pérez, 2010) the authors propose to estimate the parameters directly from the sample mean and sample covariance matrix in the data, that is using maximum likelihood. Following this strategy can lead to model overfitting when data is scarce. In bioinformatics, models are sought in domains where the number of data items is small and the number of variables is large, such as classification of mass spectrograms or microarrays. To try to avoid overfitting, we introduce classifiers based on Gaussian Join Trees (GJT) that instead of estimating by maximum likelihood perform exact Bayesian av-

eraging over the parameters.

We start by reviewing hyper Markov laws (Dawid and Lauritzen, 1993), and the hyper inverse Wishart distribution in section 2. Those results are needed to introduce in section 3.1 the hyper normal inverse Wishart ($\mathcal{HNIW}$) distribution, and prove that it is a strong hyper Markov law. This means that Bayesian learning and inference can be performed locally and efficiently. The application of the $\mathcal{HNIW}$ to build GJT classifiers is provided in section 3.2 and a preliminary empirical comparison is presented in section 4.

The contributions of the work are: (i) the proposal of GJT classifiers including a proof that the $\mathcal{HNIW}$ law is strong hyper Markov, (ii) a preliminary empirical comparison with classifiers adjusting parameters by maximum likelihood for the task of mass spectra classification and, (iii) an open source implementation of the algorithms, made available for easy reproducibility of the results. Furthermore, in order to prove that the $\mathcal{HNIW}$ is strong hyper Markov we need to determine the marginals of a normal inverse Wishart distribution, a result that we have not found elsewhere in the literature.

2 Overview of hyper Markov laws

In this section we succintly review the fundamental results in (Dawid and Lauritzen, 1993). The interested reader can find some of the missing definitions and additional details in (Dawid and Lauritzen, 1993).

2.1 Markov distributions over decomposable graphs

In the following, let $\mathcal{G} = (V, E)$ be an undirected decomposable graph over a set of random variables. A graph is said to be decomposable when all of its prime components are complete subgraphs of G; we refer to all maximal prime components as cliques of the graph.

Definition 1. A distribution P on V is called Markov over $\mathcal{G}$ if for any decomposition (A, B) of $\mathcal{G}$

$$A \perp\!\!\!\perp B | A \cap B.$$

A graphical model $M(\mathcal{G})$ is a family of probability distributions which are Markov over $\mathcal{G}$.

Definition 2. We say that distributions Q over A and R over B are consistent if both yield the same distribution over $A \cap B$.

The next theorem tells us that given a set of marginal probability distributions over the cliques of a graph that are consistent between them, we can assess the unique joint probability distribution having those marginals. Let $\mathcal{C}$ be the set of cliques of $\mathcal{G}$, and $\mathcal{S}$ the corresponding collection of separators (including possible repetitions) (see (Dawid and Lauritzen, 1993; Cowell et al., 1999) for details).

Theorem 1. *Given a pairwise consistent collection of distributions $\{Q_C : C \in \mathcal{C}\}$, Q_C being a distribution over C, the unique Markov distribution over $\mathcal{G}$ having $\{Q_C\}$ as its marginals is*

$$p(x) = \frac{\prod_{C \in \mathcal{C}} p_C(x_C)}{\prod_{S \in \mathcal{S}} p_S(x_S)}, \qquad (1)$$

where p_C, the marginal of p over the clique C is Q_C and p_S is the marginal of any of the cliques in which the separator S is included.

2.2 Hypermarkov laws

Let θ be a quantity parameterising a graphical model $M(\mathcal{G})$. A hyper Markov law is then defined by a property which mimics the global Markov property, at the parameter level

Definition 3. A law on $M(\mathcal{G})$ is called (weak) hyper Markov over $\mathcal{G}$ if for any decomposition (A, B) of $\mathcal{G}$

$$\theta_A \perp\!\!\!\perp \theta_B | \theta_{A \cap B}$$

Definition 4. A law on $M(\mathcal{G})$ is called strong hyper Markov over $\mathcal{G}$ if for any decomposition (A, B) of $\mathcal{G}$

$$\theta_{B|A} \perp\!\!\!\perp \theta_A$$

The family of hyper Markov laws and the family of strong hyper Markov laws each form a conjugate family for the sampling family $M(\mathcal{G})$. Strong hyper Markov laws produce an especially simple decomposition of the Bayesian analysis into a collection of subanalyses for smaller problems. Thus, the posterior after observing some data can be assessed locally.

The next two propositions from (Dawid and Lauritzen, 1993) will be needed later to prove that the hyper normal inverse Wishart distribution is strong hyper Markov.

Proposition 1. *Given a set of hyperconsistent laws $\{\mathcal{M}_C\}$ over clique marginals, there is a unique hyper Markov law over $\mathcal{G}$ satisfying those marginals, which is called the hyper Markov combination of $\{\mathcal{M}_C\}$.*

Proposition 2. *Let $\mathcal{P} \subseteq M(\mathcal{G})$ be a subfamily of the Markov models over $\mathcal{G}$. Assume that $\mathcal{P}$ is weak meta Markov, and for any complete set S in $\mathcal{G}$ the model $\mathcal{P}_S$ form a full exponential family. Let $\mathcal{L}$ be a hyper Markov law such that, for any clique C, the law of θ_C is a conjugate prior distribution for the model $\mathcal{P}_C$. Then $\mathcal{L}$ is strong hyper Markov.*

2.3 Hyperinverse Wishart distribution

Let $\mathcal{G}$ be a decomposable graph over a set of continuous variables. We are interested in the subfamily of models which are in $M(\mathcal{G})$ and which are assumed to be jointly multivariate normal with mean equal to zero and unknown positive definite covariance matrix Σ, that is $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. For this particular family, it is possible to define a strong hyper Markov distribution. It was introduced in (Dawid and Lauritzen, 1993), under the name of hyper inverse Wishart distribution

Definition 5. Let $\Phi^{\mathcal{C}} = \{\Phi^C, C \in \mathcal{C}\}$ be a collection of positive definite dispersion matrices, where Φ^C is the dispersion matrix over the variables in clique C. Assume that the matrices are consistent, that is, for each $B \subseteq C_1 \cap C_2$, the sub-matrices[1] $\Phi^{C_1}[B, B]$ and $\Phi^{C_2}[B, B]$ are identical. Let δ be a positive real number. We can define a collection of hyper consistent Markov laws by defining over each matrix Σ^C the following law:

$$\mathcal{L}(\Sigma^C) = \mathcal{IW}(\delta; \Phi^C).$$

The law that results from the hyper Markov combination of this collection of laws is called hyper inverse Wishart and noted $\mathcal{HIW}(\delta, \Phi^{\mathcal{C}})$.

The $\mathcal{HIW}$ is proved strong hyper Markov in (Dawid and Lauritzen, 1993).

3 Gaussian join tree classifiers

In this section we introduce Gaussian join tree classifiers as an alternative to conditional Gaussian network classifiers in (Larrañaga et al., 2006; Pérez, 2010). We start by introducing a new hyper Markov distribution, the hyper normal inverse Wishart distribution, that generalizes the Hyper inverse Wishart distribution in the case when the mean is also unknown. After that, we provide the algorithm to perform learning and inference using these distributions.

3.1 Hyper normal inverse Wishart distribution

$\mathcal{HIW}$ laws can only be used when the multivariate mean is known to be $\mathbf{0}$. We are interested in the more general subfamily of models in $M(\mathcal{G})$ with any possible mean, that is $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. We introduce the hyper normal inverse Wishart distribution and prove that it is strong hyper Markov.

Definition 6. Let $\Phi^{\mathcal{C}} = \{\Phi^C, C \in \mathcal{C}\}$ be a collection of matrices as in definition 5. Let $\boldsymbol{\mu}^{\mathcal{C}} = \{\boldsymbol{\mu}^C, C \in \mathcal{C}\}$ be a collection of vectors such that for each $B \subseteq C_1 \cap C_2$, $\boldsymbol{\mu}^{C_1}[B] = \boldsymbol{\mu}^{C_2}[B]$. Let κ and δ be non-negative real numbers. Since the marginal of a normal inverse Wishart depends only on the corresponding sub-vector and submatrix (see section B.3), the collection of laws

$$\mathcal{L}(\Sigma^C) = \mathcal{NIW}(\boldsymbol{\mu}^C, \kappa, \delta, \Phi^C)$$

is hyper consistent. The hyper Markov combination of its elements is called hyper normal inverse Wishart and noted $\mathcal{HNIW}(\boldsymbol{\mu}^{\mathcal{C}}, \kappa, \delta, \Phi^{\mathcal{C}})$. By proposition 1 the $\mathcal{HNIW}$ law is hyper Markov.

Theorem 2. *The $\mathcal{HNIW}$ distribution is strong hyper Markov.*

Proof. Is a direct application of proposition 2. First, note that in our case, the set of models is the set of multidimensional Gaussian models

[1] Given a matrix M and sets of indexes I,J, we note $M[I, J]$ the sub-matrix that keeps the rows with indexes in I and the columns with indexes in J. Equivalent notation is used for vectors.

that factorize over $\mathcal{G}$. Thus, it is weak meta Markov and over a complete set, it forms a full exponential family. Since the $\mathcal{HNIW}$ law is hyper Markov, and for any clique C, it is a conjugate distribution for the multidimensional Gaussian over C, by proposition 2 it is strong hyper Markov. □

3.1.1 Conjugacy

Since the $\mathcal{HNIW}$ is strong hyper Markov, we can update each of the marginal distributions locally. Furthermore, since the local laws are $\mathcal{NIW}$ the update can be done using the result in appendix B.1. Hence, in order to learn from data we only have to update our hyperparameters using the equations (4)-(7).

3.1.2 Predictive distribution

The predictive distributions for each clique are multivariate t distributions, as given by equation (8). To get the distributions over a separator, we marginalize the distribution of one of the cliques that it separates using the result provided in section A.1 about the marginal of a multivariate t.

3.2 The Gaussian join tree classifier

As in CGN classifiers, Gaussian join tree classifiers start by determining a dependency structure and then adjust the model parameters for that structure (see Algorithm 1).

Algorithm 1 General GJT classifier

```
function GJTLearner(D)
    T := DetermineCliqueTree(D)
    Θ := LearnGJTParameters(T, D)
    return ⟨T, Θ⟩
end function
```

In Gaussian join tree classifiers the structure is represented as a join tree. We formally define join trees following (Cowell et al., 1999).

Definition 7. A clique tree $\mathcal{T} = \{T_1, \ldots, T_n\}$ is a tree where each node $T_i \subseteq X$ is a set of variables. If T_i is a parent of T_j, the separator between T_i and T_j is the set of variables $S_{ij} = T_i \cap T_j$. A join tree $\mathcal{T}$ is a clique tree such that for every pair of nodes T_i, T_j, $T_i \cap T_j$ is a subset of every separator on the unique path from T_i to T_j.

The parameters of a GJT classifier are determined combining the results described in section 3.1.1 and 3.1.2. For each of the classes, each of the cliques and separators will be assigned a multivariate t distribution. The algorithmic details are provided in Algorithm 2.

3.3 Predicting

Given a new unclassified data point, Bayes formula is used to determine the posterior probability for each class. We use Laplace formula to assess the prior probability for class c (which is equivalent to assuming a Dirichlet prior) and get the posterior by multipling it by joint probability obtained using equation (1).

4 Empirical comparison

In this section we compare GJT classifiers with classifiers making similar independency assumptions, but whose parameters are adjusted by maximum likelihood, namely CGN classifiers, as proposed in (Pérez, 2010). We use two different datasets from the bioinformatics domain, one for ovarian cancer and one for pancreatic cancer. Both datasets have been obtained from the NIH and contain high resolution spectrograms coming from surface-enhanced laser desorption/ionization time of flight mass spectrometry (SELDI-TOF MS). In both datasets the objective is to distinguish spectrograms coming from cancer patients from those coming from control individuals.

The ovarian cancer dataset contains a total of 216 spectrograms, 121 from cancer patients and 95 controls. The m/z values do not coincide along the different spectrograms. Thus, to create the variables, the m/z axis data has been discretized into different bins, creating a variable for each bin, for a total of 11300 variables. Thus, the number of variables largely exceeds the number of data points. For each spectrogram, the average of the values of each bin has been assigned to that bin's variable.

The pancreatic cancer dataset contains a total of 181 spectrograms, 101 from cancer patients and 80 controls. Each spectrogram is defined by 6771 variables which in this case are aligned, so no discretization is needed.

Algorithm 2 Bayesian learning of GJT parameters

function LEARNGJTPARAMETERS($\mathcal{T}, \mathcal{D}$)
  for each class c **do**
    Set the parameters of the prior distribution κ, δ,$\boldsymbol{\eta}$, and Ψ.
    With the data from class c assess the number of data points $n(c)$, the sample mean $\overline{\mathbf{y}}(c)$
    and the sample covariance matrix $\Sigma(c)$.
    $\kappa'(c) := \kappa(c) + n(c)$
    $\delta'(c) := \delta(c) + n(c)$
  end for
  $\mathcal{C} := Cliques(\mathcal{T})$
  for $C \in \mathcal{C}$ **do**
    for each class c **do**
      $\boldsymbol{\eta}' := \frac{\kappa(c)\boldsymbol{\eta}[C]+n(c)\overline{\mathbf{x}}(c)[C]}{\kappa(c)+n(c)}$
      $\Psi' := \Psi[C,C] + (n(c)-1)\Sigma(c)[C,C] + \frac{\kappa n(c)}{\kappa+n(c)}(\overline{\mathbf{x}}(c)[C] - \boldsymbol{\eta}[C])(\overline{\mathbf{x}}(c)[C] - \boldsymbol{\eta}[C])^T$.
      $d(C,c) := t_{\delta'}(\boldsymbol{\eta}', \frac{\kappa'+1}{\kappa'\delta'}\Psi')$
      for each clique C' child of C in $\mathcal{T}$ **do**
        $d(C' \cap C, c) = t_{\delta'}(\boldsymbol{\eta}'[C' \cap C], \frac{\kappa'+1}{\kappa'\delta'}\Psi'[C' \cap C, C' \cap C])$
      end for
    end for
  end for
end function

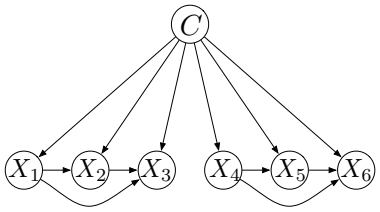
We have used two different families of structures. Both are based on the hypothesis that those variables that represent close m/z relations are more likely to have large correlations than those whose m/z values are further away.

A k-BOX structure can be defined over an ordered set of variables $V = \langle X_1, \ldots, X_n \rangle$. It is a restriction of *semi Naïve Bayes* (Larrañaga et al., 2006), also known as *JAN* (Pérez, 2010). In a JAN structure the variables are joined in groups to form multivariate distributions. In the k-BOX structure we divide the variables in disjoints sets of k contiguous variables. The network structure can be seen in Figure 1(a) and the corresponding covariance matrix in Figure 2(a). In our case the ordering is provided by the m/z value.

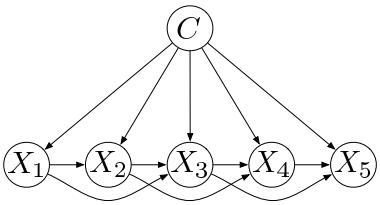
The second proposed structure is the k-BAND structure. In k-BAND, we assume that each variable is independent of all the remaining variables given the $k-1$ variables that precede it and the class variable. Therefore, the k-BAND structure is formed by cliques of size k with separators of $k-1$ variables.

The covariance matrix for a k-BAND structure is a band of size k around the diagonal, as is shown in Figure 2(b). An example of the structure is shown in Figure 1(b).

We have run a sequence of experiments on

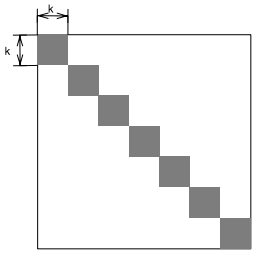


(a) k-BOX model

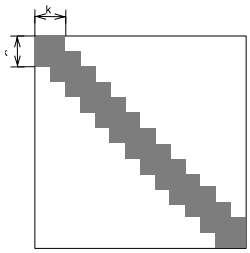


(b) k-BAND model

Figure 1: Structure of the graph for different models



(a) k-BOX model



(b) k-BAND model

Figure 2: Structure of the covariance matrix

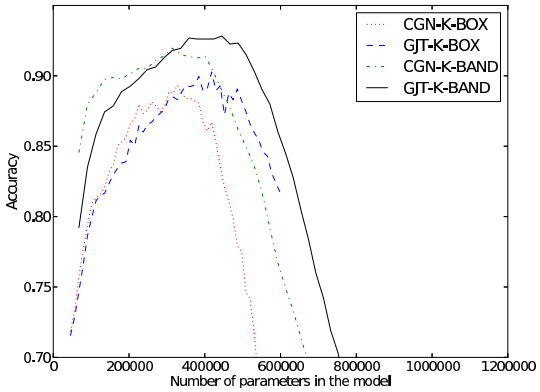


Figure 3: Prediction of ovarian cancer. Accuracy versus number of parameters in model.

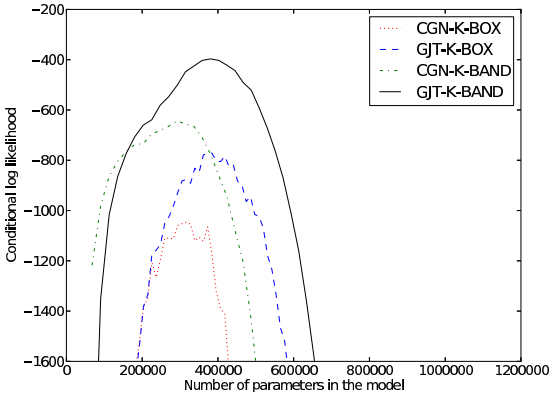


Figure 4: Prediction of ovarian cancer. CLL versus number of parameters in model.

each dataset to compare the different structures (k-BOX and k-BAND) and parameter learning methods (CGN and GJT) varying the k parameter.

We have run 5 repetitions of 10-fold cross validation and assessed the accuracy: the ratio of the number of data classified correctly to the total number of data classified; and conditional log-likelihood (CLL): the sum of the logarithm of the probability of the real class of the data given by the classifier. While accuracy gives us information about how many patients are correctly classified, CLL measures how accurately the probabilities for each class are estimated, which is very relevant for adequate decision making.

The ovarian data has been modelled using k-BOX and k-BAND structures with k ranging from 1 to 50. In Figure 3 we show the mean accuracy versus the number of parameters in the model and in Figure 4 we show the mean CLL versus the number of parameters in each structure. We see that in that dataset, k-BAND models are more accurate that k-BOX models. For low values of k, CGN performs better than GJT, but GJT has a largest accuracy and a highest CLL at its peak, and shows a more graceful decay as the number of parameters grows beyond that peak. The pancreatic cancer data has been classified using k-BOX and k-BAND structures with k ranging from 1 to 30. Figure 5 and Figure 6 show respectively

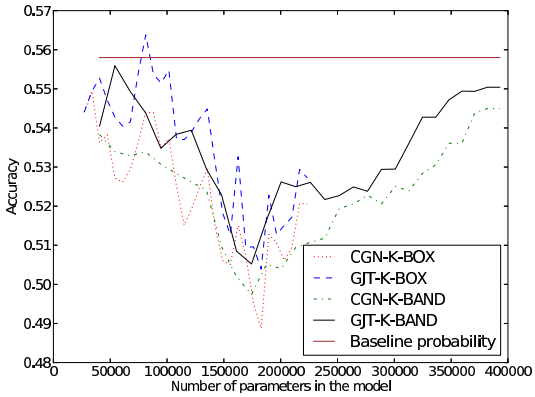


Figure 5: Prediction of pancreatic cancer. Accuracy versus number of parameters in model.

the mean accuracy and mean CLL against the number of parameters. Note that the accuracy for this dataset is much lower and close to the frequency of the largest class, that is 55.8% in this datasets. Previous studies have shown that the accuracy results for this dataset are much lower than for the previous one. The k-BOX model using GJT appears to reach the highest accuracy and the k -BAND model reaches the highest CLL also for pancreatic cancer.

The source code is available at *http://www.iiia.csic.es/ cerquide/pypermarkov*

5 Conclusions and future work

We have introduced a new family of classifiers for continuous domains, namely Gaussian join Tree classifiers that perform exact Bayesian av-

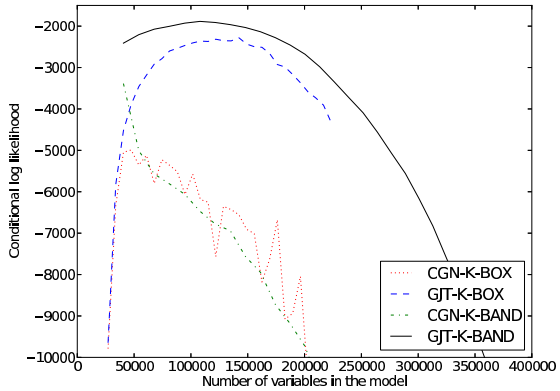


Figure 6: Prediction of pancreatic cancer. CLL versus number of parameters in model.

eraging over the parameters by virtue of the hyper normal inverse Wishart law, that we have introduced and proved to be strong hyper Markov. To assess the benefits we have compared GJT classifiers with GCN classifiers which assuming the same set of independencies adjust parameters using maximum likelihood. We performed our comparison with two high resolution mass spectrometry datasets for ovarian and pancreatic cancer prediction. We have seen that, for two simple dependency structures, our classifiers reach a better peak accuracy and a consistently better conditional log likelihood.

We have introduce k-BOX and k-BAND structure, which are part of the same family of join trees structures. For example different sizes of the separators define different models in the family. The study of this family remains also as future work.

The GJT parameter averaging can be performed over any set of dependencies that can be encoded into a decomposable graph. We can adapt any algorithm for learning the structure of a CGN classifier so that it outputs a join tree by moralizing and triangulating the DAG that encodes the structure of the CGN, and then running maximum cardinality search. A future line of work is comparing along the datasets in (Pérez, 2010) using the same structure learning algorithms that they suggest, thus testing if the benefits extend to domains with a smaller number of attributes.

Letac and Massam have generalized the hyper inverse Wishart distribution in (Letac and Massam, 2007). Studying the use of this generalized hyper inverse Wisharts for classification remains as future work.

Acknowledgments

We would like to thank Aritz Pérez, Steffen Lauritzen, Phil Dawid and Karina Gibert. This work has been funded by projects EVE (TIN2009-14702-C02-01 and TIN2009-14702-C02-02), CSIC 201050I008, and the Generalitat of Catalunya (2009-SGR-1434).

References

Susanne Gammelgaard Bøttcher. 2004. *Learning Bayesian Networks with Mixed Variables*. Ph.D. thesis, Aalborg University.

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. 1999. *Probabilistic Networks and Expert Systems*. Springer-Verlag.

A. P. Dawid and S. L. Lauritzen. 1993. Hyper Markov Laws in the Statistical Analysis of Decomposable. *The Annals of Statistics*, 21(3):1272–1317.

Dan Geiger and David Heckerman. 1994. Learning gaussian networks. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-94)*, pages 235–243.

Andrew Gelman, JB Carlin, HS Stern, and Rubin. 2004. *Bayesian data analysis*.

Samuel Kotz and Saralees Nadarajah. 2004. *Multivariate T-Distributions and Their Applications*. Cambridge University Press, Cambridge.

Pedro Larrañaga, P, Aritz Pérez, Iñaki Inza, and Pedro Larra. 2006. Supervised classification with conditional Gaussian networks : Increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning*, 43(January):1–25.

Steffen L. Lauritzen. 1996. *Graphical models*. Oxford University Press.

Gérard Letac and Hélène Massam. 2007. Wishart distributions for decomposable graphs. *The Annals of Statistics*, 35(3):1278–1323, July.

Aritz Pérez. 2010. *Supervised classification in continuous domains with Bayesian networks*. Ph.D. thesis, Universidad del Pais Vasco.

A Multivariate t-distributions

A p-dimensional random vector $\mathbf{x}$ is said to have the p-variate t distribution with degrees of freedom ν, mean vector μ , and correlation matrix R (that is $x \sim t_\nu(\mu, R)$) if its joint pdf is given by

$$p(\mathbf{x}|\nu,\mu,R) = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|R|^{1/2}}\cdot \cdot [1+\frac{1}{\nu}(x-\mu)R^{-1}(x-\mu)]^{-(\nu+p)/2} \quad (2)$$

A.1 Marginals

Let $\mathbf{x}$ be a p-dimensional random vector $\mathbf{x} \sim t_\nu(\boldsymbol{\mu}, R)$. Furthermore let $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2})$ where $\mathbf{x_1}$ is p_1 dimensional, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and $R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$. We have that

$$\mathbf{x_1} \sim t_\nu(\boldsymbol{\mu}_1, R_{11}) \quad (3)$$

For more information regarding multivariate t-distributions see (Kotz and Nadarajah, 2004).

B Normal Inverse Wishart distributions

The inverse Wishart distribution is defined on real-valued positive-definite $p \times p$ matrices. It has two parameters: a real number $\delta > 0$ and a $p \times p$ positive-definite matrix Ψ. The probability density function is

$$\mathcal{IW}(X|\delta,\Psi) = \frac{|\Psi|^{\frac{\delta+p-1}{2}}|X|^{-\frac{\delta+2p}{2}}}{2^{\frac{(\delta+p-1)p}{2}}\Gamma_p(\frac{\delta+p-1}{2})}e^{-\frac{1}{2}tr(\Psi X^{-1})}$$

The normal inverse Wishart distribution is defined on pairs composed of (i) vectors of dimension p and (ii) real-valued positive-definite $p \times p$ matrices. It has four parameters: a p dimensional vector $\boldsymbol{\eta}$ that encodes the location, a positive real number κ that acts as scaling factor, and δ and Ψ as in the inverse Wishart. The probability density function is

$$\mathcal{NIW}(\boldsymbol{\mu},\Sigma|\boldsymbol{\eta},\kappa,\delta,\Psi) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\eta},\frac{1}{\kappa}\Sigma)\cdot\mathcal{IW}(\Sigma|\delta,\Psi)$$

B.1 Conjugacy

The normal inverse Wishart distribution is conjugate to the multivariate normal (Gelman et al., 2004). Thus, if we assume as prior a $\mathcal{NIW}(\boldsymbol{\mu},\Sigma|\eta,\kappa,\delta,\Psi)$ and we are given a sample X from a multivariate normal, the posterior will be a $\mathcal{NIW}(\boldsymbol{\mu},\Sigma|\eta',\kappa',\delta',\Psi')$ where

$$\boldsymbol{\eta}' = \frac{\kappa\boldsymbol{\eta}+n\overline{\mathbf{x}}}{\kappa+n}, \quad (4)$$

$$\kappa' = \kappa+n, \quad (5)$$

$$\delta' = \delta+n, \quad (6)$$

$$\Psi' = \Psi+(n-1)S+\frac{\kappa n}{\kappa+n}(\overline{\mathbf{x}}-\boldsymbol{\eta})(\overline{\mathbf{x}}-\boldsymbol{\eta})^T. \quad (7)$$

and S is the sample covariance.

B.2 Predictive distribution

The predictive distribution is

$$p(\mathbf{x}|\boldsymbol{\eta},\kappa,\delta,\Psi) = = \int_{\boldsymbol{\mu},\Sigma}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)\mathcal{NIW}(\boldsymbol{\mu},\Sigma|\boldsymbol{\eta},\kappa,\delta,\Psi) = = t_\delta(\boldsymbol{\eta},\frac{\kappa+1}{\kappa\delta}\Psi). \quad (8)$$

B.3 Marginals

Proposition 3. *Let* $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, $\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix}$ *and* $\Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}$. *If* $(\boldsymbol{\mu},\Sigma) \sim \mathcal{NIW}(\boldsymbol{\eta},\kappa,\nu,\Psi)$, *then* $(\boldsymbol{\mu}_1,\Sigma_{11}) \sim \mathcal{NIW}(\boldsymbol{\eta}_1,\kappa,\nu,\Psi_{11})$

Proof. The marginal can be assessed as follows

$$\begin{aligned} &P(\boldsymbol{\mu}_1,\Sigma_{11}|\boldsymbol{\eta},\kappa,\delta,\Psi) = \\ &= \int_{\boldsymbol{\mu}_2,\Sigma_{12},\Sigma_{22}}\mathcal{NIW}(\boldsymbol{\mu},\Sigma|\boldsymbol{\eta},\kappa,\delta,\Psi) = \\ &= \int_{\boldsymbol{\mu}_2,\Sigma_{12},\Sigma_{22}}\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\eta},\frac{1}{\kappa}\Sigma)\cdot\mathcal{IW}(\Sigma|\delta,\Psi) = \\ &= \int_{\boldsymbol{\mu}_2,\Sigma_{12},\Sigma_{22}}\mathcal{N}(\boldsymbol{\mu}_1|\boldsymbol{\eta}_1,\frac{1}{\kappa}\Sigma_{11})P(\boldsymbol{\mu}_2|\boldsymbol{\mu}_1,\boldsymbol{\eta},\kappa,\delta,\Psi)\cdot \\ &\cdot\mathcal{IW}(\Sigma_{11}|\delta,\Psi_{11})P(\Sigma_{22},\Sigma_{21}|\Sigma_{11},\delta,\Psi) = \\ &= \mathcal{N}(\boldsymbol{\mu}_1|\boldsymbol{\eta}_1,\frac{1}{\kappa}\Sigma_{11})\mathcal{IW}(\Sigma_{11}|\delta,\Psi_{11})\cdot \\ &\cdot\int_{\boldsymbol{\mu}_2,\Sigma_{12},\Sigma_{22}}P(\boldsymbol{\mu}_2|\boldsymbol{\mu}_1,\boldsymbol{\eta},\kappa,\delta,\Psi)P(\Sigma_{22},\Sigma_{21}|\Sigma_{11}\delta,\Psi) = \\ &= \mathcal{N}(\boldsymbol{\mu}_1|\boldsymbol{\eta}_1,\frac{1}{\kappa}\Sigma_{11})\mathcal{IW}(\Sigma_{11}|\delta,\Psi_{11}) \quad (9) \end{aligned}$$

□