Stat521A Spring 2009: homework 4

1 Ising models are equivalent to Poisson log-linear models

(Source: [HTF09] ex. 17.12)
Consider an Ising model on d nodes

$$p(\mathbf{x}|\boldsymbol{\theta}) = \exp[\sum_{<j,k>} \theta_{jk} x_j x_k - \Phi(\boldsymbol{\theta})] \tag{1}$$

where the sum is over all edges $< j, k >$, $x_j \in \{0, 1\}$ and Φ is the log partition function

$$\Phi(\boldsymbol{\theta}) = \log \sum_{\mathbf{x}} \exp[\sum_{<j,k>} \theta_{jk} x_j x_k] \tag{2}$$

We assume there is a special node X_0 which is clamped to 1, and is connected to all the other nodes with weights θ_{0i}. The log-likelihood is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} [S(\mathbf{x}_i, \boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})] \tag{3}$$

$$S(\mathbf{x}_i, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{<j,k>} \theta_{jk} x_{ij} x_{ik} \tag{4}$$

The gradient is given by

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{jk}} = \sum_{i=1}^{n} x_{ij} x_{ik} - n \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \theta_{jk}} \tag{5}$$

$$\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \theta_{jk}} = \frac{1}{\sum_{\mathbf{x}'} e^{S(\mathbf{x}')}} [\sum_{\mathbf{x}} e^{S(\mathbf{x})} \frac{\partial}{\partial \theta_{jk}} S(\mathbf{x})] \tag{6}$$

$$= \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\theta}) x_j x_k = E_{\boldsymbol{\theta}}[X_j X_k] \tag{7}$$

Hence at the MLE, we have that the empirical moments match the model moments, as is standard for an exponential family model:

$$\hat{E}[X_j X_k] = E_{\boldsymbol{\theta}}[X_j X_k] \tag{8}$$

where we have defined

$$\hat{E}[X_j X_k] = \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik} \tag{9}$$

One can fit this model using gradient methods (where computing the gradient takes $O(nd^2 + 2^d)$ time). Alternatively, we can treat this model as a generalized linear model, and use IRLS, as we show below.

1. Explain why the constant node $X_0 = 0$ must be included. (Consider a model with just two variables.)

2. Consider a Poisson regression model with d binary covariates, $x_{ij} \in \{0, 1\}$, and response variable $y_i \in \{0, 1, \ldots, 2^d - 1\}$, with distribution

$$P(Y = y|\mathbf{X} = \mathbf{x}) = \frac{e^{-\mu(\mathbf{x})}\mu(\mathbf{x})^y}{y!} \tag{10}$$

We assume a log-linear model with first-order interactions for the mean:

$$\log \mu(\mathbf{x}) = \theta_{00} + \sum_{<j,k>} x_j x_k \theta_{jk} \tag{11}$$

where $x_{i0} = 1$ as before. Consider the log likelihood

$$\ell(\boldsymbol{\theta}) = \sum_i \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \tag{12}$$

Show that the gradient equation of ℓ wrt θ_{00} computes the log partition functon (Equation 2).

3. Show that the gradient equation of ℓ wrt the other θ terms yields the moment matching equation in Equation 8.

4. Explain how we can fit an Ising model by fitting a Poisson regression model (using, say, IRLS or Newton's method). Give an example. What is the computational complexity of this procedure? Hint: the book [Agr02] may be helpful.

5. The Poisson regression model is a conditional model $p(Y|\mathbf{x}, \boldsymbol{\theta})$, whereas the Ising model is an unconditional multinomial model, $p(\mathbf{x}|\boldsymbol{\theta}) = \text{Mu}(\mathbf{x}|1, \boldsymbol{\theta})$. Explain how to convert the former to the latter. (Hint: a Poisson conditioned on the count $N = \sum_i y_i$ is a Multinomial.)

2 EM for family-based genetic association II

This exercise is an extension of the question in HW3. Consider a set of N_f families, each of which has a mother, father and N_c children. We measure the alleles at N_g gene/SNP locations; these have values $\{AA, Aa, aa\}$. Let $G^g_{fc} \in \{0, 1, 2\}$ represent the number of copies of the minor allele (a) for child c in family f for gene g. Similarly, G^g_{fm} is the maternal allele and G^g_{fp} is the paternal allele. Let $i \in \{m, f, 1, \ldots, N_c\}$ index a generic family member. We do not observe the G^g_{fi} directly, but instead observe a noisy copy, $O^g_{fi} \in \{0, 1, 2\}$. In addition, we observe a covariate (assumed scalar for simplicity), $X_{fi} \in \mathbb{R}$, representing environmental factors, and a response/ phenotype (assumed continuous for simplicity), $Y_{fi} \in \mathbb{R}$. The goal is to infer if the genes or environment "cause" the response, and if it is the genes, which ones. We assume a linear regression model of the form

$$Y_{fi} \sim \mathcal{N}(Y_{fi}|\sum_{g=1}^{N_g} \beta_g G^g_{fi} + \beta_{N_g+1} X_{fi} + \beta_{N_g+2}, 1/\lambda) \tag{13}$$

Thus if gene g is causal, its coefficient should be large (positive or negative), and it will have an additive effect on the response. The more copies of the minor allele at locus g, the greater the effect on the response. If multiple genes are involved, the effects should be larger (unless their coefficients are of opposite signs! We ignore that issue here).
Some synthetic data is shown in Figure 1. We assume $N_f = 50$ familes, each of which has $N_c = 2$ children. There are $N_g = 2$ genes. The environmental factor X is random. The coefficient vector is $\boldsymbol{\beta} = [-49.4, 0.6, 0.5, 1.7]$, so gene 1 has a negative effect on the response, gene 2 is irrelevant, the environment is irrelevant, and the offset term is negligible. Consequently we see that the response Y reflects the pattern of alleles in gene 1: when $G^1_{fi} = 2$, the response is large and negative, and when $G^1_{fi} = 0$, the response is near zero.
In the test data, you observe X, Y, and O, and you have to infer $\boldsymbol{\beta}$. From this, you can estimate which genes are involved (if any), and how strong their effect is. The simplest approach is to compute a MAP estimate of $\boldsymbol{\beta}$ using EM, treating G as the missing data. (Ideally we would want some measure of uncertainty in our conclusions, too, e.g., a

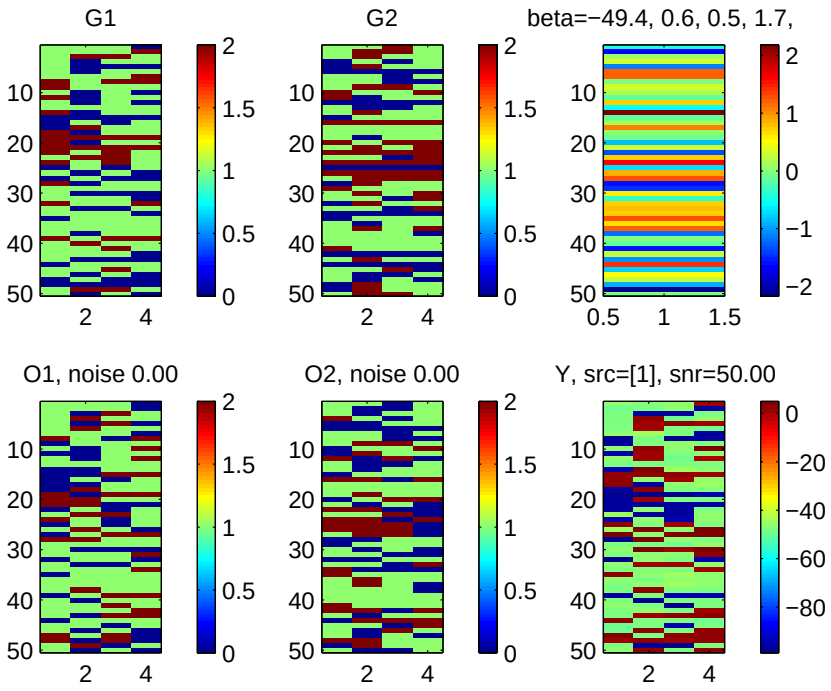
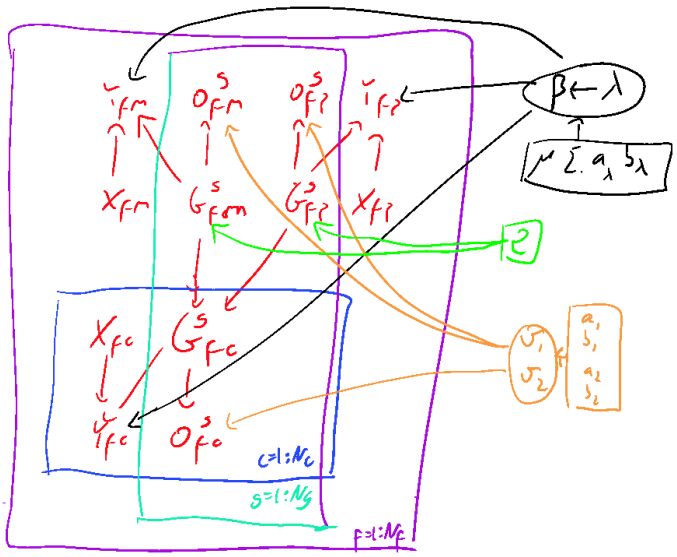
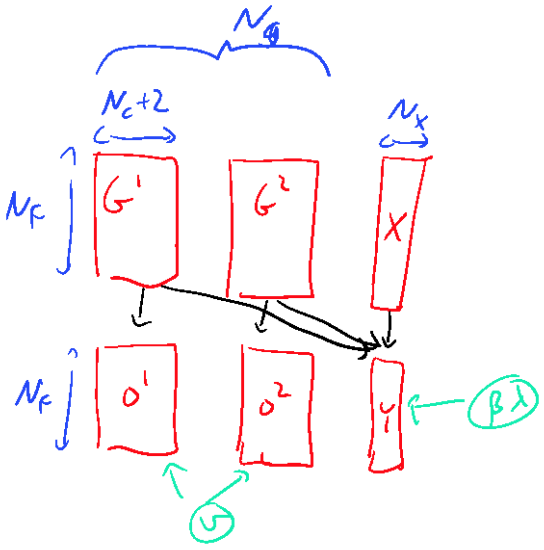


Figure 1: Family-based genetic association, dataset 1. We see that the response Y mirrors gene 1, but ignores gene 2 and the environment X (top right). Note that, even though $\nu = 0$, O is still not exactly equal to G, since AA (2) can get flipped to aa (0) and vice versa. However, in this noise free case, Aa is observed perfectly (so the patches of green in G and O are the same).



(a)



(b)

Figure 2: (a) Graphical model for family-based genetic association. (b) Visualization of how the data was generated. Each row of the G matrix is a single family, and has a correlation structure as shown in (a).

posterior credible interval. We will leave that to future exercises.) In general, we may be missing some of the entries for X, Y and O (corresponding to unmeasured people). This is a trivial extension which we will not worry about.
The data was generated from the model shown in Figure 2 (although the parameters ν, β and λ were set by hand rather than sampled). You can make use of this fact when performing inference. First, the root nodes, G^g_{fp} and G^g_{fm}, have the following CPD:

AA	Aa	aa
$(1-\rho)^2$	$2\rho(1-\rho)$	ρ^2

We used $\rho = 0.5$; you may assume this constant is known. (It can be estimated from the population frequencies of each allele type for each gene.) The child nodes, G^g_{fc}, have the following CPD, which encodes Mendel's laws:

F	M	$p(C = AA)$	$p(C = Aa)$	$p(C = aa)$
AA	AA	1	0	0
AA	Aa	0.5	0.5	0
Aa	AA	0.5	0.5	0
AA	aa	0	1	0
aa	AA	0	1	0
Aa	Aa	0.25	0.5	0.25
Aa	aa	0	0.5	0.5
aa	Aa	0	0.5	0.5
aa	aa	0	0	1

The observed genotypes have the following noise model:

G	$p(O = AA)$	$p(O = Aa)$	$p(O = aa)$
AA	$\frac{1-\nu_2}{2}$	ν_2	$\frac{1-\nu_2}{2}$
Aa	ν_1	$1 - 2\nu_1$	ν_1
aa	$\frac{1-\nu_2}{2}$	ν_2	$\frac{1-\nu_2}{2}$

We varied ν depending on the dataset. You should estimate these values. You can use an informative beta prior, which encodes the belief that ν is unlikely to exceed 0.1.
Finally, we sampled Y according to Equation 13. We set $\lambda = 1$ but varied β according to the data set. You should estimate these values. Use a conjugate prior of the form

$$p(\boldsymbol{\beta}, \lambda) = \text{Ga}(\lambda | a_\lambda, b_\lambda) \mathcal{N}(\boldsymbol{\beta} | \boldsymbol{\mu}, \boldsymbol{\Sigma}/\lambda) \tag{14}$$

Use the vague hyper-parameters such as $a_\lambda = b_\lambda = 0.01$, $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Sigma} = 100\mathbf{I}$.

1. Derive an EM algorithm for MAP estimation in this model. This is essentially the same as HW3, except now we have priors, the likelihood model for O is slightly different, and the regression model for Y is slightly different.

2. Implement your algorithm. You may find the file `familyTreeGeneDataMystery`, in `pmtk/examples`, helpful. (Download the latest version of PMTK (1.4.0) first.) It specifies then generative model, without revealing the parameter values. When performing inference in the DGM (for the E step), since there are only 4 hidden nodes per family, you can use any inference method you want, including brute force enumeration (which explicitly builds the joint containing 2^4 entries) or variable elimination. (See `inheritedDiseaseVarElim` in `pmkt/examples/dgmDistExamples` for an example of how to do inference in PMTK in a tree with discrete hidden nodes and continuous observed child nodes.) For larger families, one should use belief propagation (see future exercise).

3. Load the files `familyTreeDataX.mat`, for $X = 1, \ldots, 5$, from `pmtk/data`. Each one should contain variables of the following form:

```
X: [50x1 double]
Y: [50x4 double]
O: [50x4x2 double]
```

We have $N_f = 50$ families, each with $N_c = 2$ children, and $N_g = 2$ genes. The values of O are in $\{1, 2, 3\}$ (not $\{0, 1, 2\}$ as above), and represent AA,Aa, and aa. The matrix is indexed as follows: $O(f, i, g)$, for family $f = 1 : N_f$, person $i = 1 : N_c + 2$, and gene $g = 1 : N_g$. These datasets were generated with different parameter values. Fit a separate model (using EM) to each dataset and state your MAP estimate of $\boldsymbol{\beta}$. Make a plot similar to Figure 1, where for the unobserved G variables you plot the mode of the marginal $p(G_{fi}^g|\mathcal{D}, \hat{\boldsymbol{\theta}})$.

4. Try fitting different models to each dataset, in which different subsets of $\boldsymbol{\beta}$ are forced to zero. Use a model selection criterion such as BIC to pick the best. Which genes (if any) cause the response in each dataset?

5. Try ignoring the family structure between the Gs and see if your conclusions change.

6. Pretend that O are noise-free versions of G. (You will have to "soften" the CPD for $p(G_{fc}^g|G_{fp}^g, G_{fm}^g)$ to prevent a probability of zero being assigned to non-mendelian data.) Rerun your experiments and see if your conclusions change. In this case, there is no missing data, so you do not need EM. In fact, you can compute the exact posterior over the parameters, and use the marginal likelihood for each model, instead of using BIC (this requires using the $p(\mathcal{D})$ formula for Bayesian linear regression). See if this makes any difference. Try making the dataset smaller (say, 10 families); one expects the Bayesian score to be better than BIC for small sample sizes.

References

[Agr02] A. Agresti. *Categorical Data Analysis*. 2002. 2nd ed.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. 2nd edition.