Exam Statistical Genomics

Date: Thursday 30 January, 2014
Time: 9.00-12.00
Place: Bernoulliborg Building, Room 5161.0222
Progress code: WISG-09

Rules to follow:

- The number of points per question are indicated within a box. Ten points are free.
- Do not forget to fill in your name and student number.
- We wish you success with the completion of the exam!

START OF EXAM

1. **Kullback-Leibler divergence, likelihood and deviance.** [20]
The Kullback-Leibler divergence is defined as $I(f;g) = E_f \log \frac{f(X)}{g(X)}$. We consider a graphical log-linear model. Let n be a table of counts, where $n(x)$ is a particular cell-count. Let $N = \sum_x n(x)$ be the total number of observations.

    - [10] Show that the log-likelihood for a table n can be written as
    $$l(p;n) = l(\frac{n}{N};n) - N \times I(\frac{n}{N};p),$$
    where p and n/N are interpreted as cell probabilities.
    - [10] Consider a particular graphical log-linear model M. Show that the deviance can be written as
    $$\text{Dev}(M) = 2\sum_x n(x) \log \frac{n(x)/N}{\hat{p}^M(x)},$$
    where $\hat{p}^M$ is the maximum likelihood estimator for p under M.

2. **Binary log-linear model.** [35]

    We study the three year survival (X_3) of 474 breast cancer patients according to nuclear grade (X_2) and diagnostic centre (X_1).

    (a) [5] Derive the MLE of $p(X_1 = 1, X_2 = 1, X_3 = 1)$ under the saturated model.

    (b) [5] Derive the MLE of $p(X_1 = 1, X_2 = 1, X_3 = 1)$ under the model $1.2 + 1.3$.

	$X_2 = 0$		$X_2 = 1$		
	$X_3 = 0$	$X_3 = 1$	$X_3 = 0$	$X_3 = 1$	
$X_1 = 0$	35	59	47	112	253
$X_1 = 1$	42	77	26	76	221
total	77	136	73	188	474

(c) [10] We can decompose the log-density $\log p$ in a log-linear way using u-terms. Derive the MLE of the u_1 term under the model $1.2 + 1.3$.

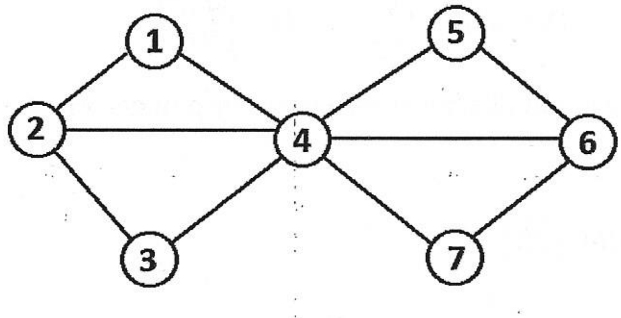
(d) [10] We want to test whether we can exclude the link $(2, 3)$ from the saturated model. Determine the edge exclusion deviance and test whether you can delete it at a 5% significance level. A chi-squared table can be found at the end of the exam.

(e) [5] Argue whether the model $1.2 + 1.3 + 2.3$ is graphical and/or hierarchical.

3. **Gaussian graphical model (1).** [15]
The sample variance matrix based on $N = 50$ observations from a Gaussian graphical model is

$$S = \begin{pmatrix} 3.9 & -0.1 & 1.7 & -1.1 & 0.3 & -0.2 & 0.1 \\ -0.1 & 1.2 & 0.9 & -0.6 & 0.2 & -0.1 & 0.0 \\ 1.7 & 0.9 & 3.0 & -1.9 & 0.5 & -0.4 & 0.1 \\ -1.1 & -0.6 & -1.9 & 3.5 & -1.0 & 0.7 & -0.2 \\ 0.3 & 0.2 & 0.5 & -1.0 & 1.7 & -1.2 & 0.3 \\ -0.2 & -0.1 & -0.4 & 0.7 & -1.2 & 4.8 & 0.7 \\ 0.1 & 0.0 & 0.1 & -0.2 & 0.3 & 0.7 & 2.2 \end{pmatrix}$$

Calculate the following 3 elements of the MLE of the variance covariance matrix Σ associated with the following conditional independence graph:



(a) [5] $\hat{\Sigma}_{12}$?

(b) [5] $\hat{\Sigma}_{13}$?

(c) [5] $\hat{\Sigma}_{17}$?

4. **Gaussian graphical model (2).** $\boxed{20}$
A sample covariance matrix for a sample from a Gaussian graphical model $N(0, V)$ of size $N = 10$ is given as

$$S = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

(a) $\boxed{10}$ Use the deviance $\text{Dev}(M_1)$ for $M_1 = X_3 \perp (X_2, X_1)$ to test whether M_1 fits the data.

Hint: You can use the fact that $|S| = 4$ and that the determinant of the top submatrix of S is $|S_{12,12}| = 3$. Moreover, the inverse of a 2×2 matrix, is given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

(b) $\boxed{10}$ Use (for example) the fact that

$$Inf(X_3 \perp X_2 | X_1) = -\frac{1}{2} \log \frac{|V|\ |V_{11}|}{|V_{12,12}|\ |V_{13,13}|}$$

to calculate the deviance $\text{Dev}(M_2)$ for $M_2 = X_3 \perp X_2 | X_1$ to test whether M_2 fits the data.

If you want to use directly the edge exclusion deviance, then that's fine too.

END OF EXAM

Chi-squared table.

$\nu \setminus \alpha$	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860

Table 1: Values of $\chi^2_{\alpha,\nu}$: entries correspond to values of x, such that $P(\chi^2_\nu > x) = \alpha$, where χ^2_ν correspond to a chi-squared distributed variable with ν degrees of freedom.