582653 Computational Methods of Systems Biology

Lecture 9: Unsupervised inference/Correlation networks

24.2.2011
In previous lecture, we looked how a biological network can be completed using supervised machine learning.

This lecture we examine the question of inferring the network when no prior information on the interactions is assumed.

We concentrate on undirected networks.
Covariance

- Assuming we have a set of experimental data given as a \( m \times d \) matrix \( X \) with one row per sample, one column per variable
  - e.g. gene expression profiles over time or under different conditions
- Co-variance as the measure of strength of belief of an edge being present: \( \sigma(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) \)
- Empirical covariance given sample \((x_i, y_i), i = 1, \ldots, m\) of data \( s(x, y) = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_x)(y_i - \mu_y) \)
- Note: to get an unbiased estimate one should actually divide by \( m - 1 \) (we ignore that aspect here)
Covariance

- Given a data set covariance tells us about how the data lies with respect the variables
- Example on the right:
  - Top: no co-variance between $x$ and $y$, $x$ has higher variance than $y$, diagonal co-variance matrix with inequal entries
  - Middle: no co-variance between $x$ and $y$, equal variance for $x$ and $y$, diagonal co-variance matrix with equal entries
  - Bottom: $x$ and $y$ co-vary, co-variance matrix will have non-zero off-diagonal entries
Covariance matrix

- Given a set of samples (rows of data matrix $X$

$$x_i = (x_{i1}, \ldots, x_{id}), \ i = 1, \ldots, m$$

with mean vector $\mu = \frac{1}{m} \sum_{i=1}^{m} x_i$, the empirical co-variance matrix is as a $d \times d$ matrix given by

$$S = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T$$

containing all pairwise empirical co-variances $s_{ij} = s(x_i, x_j)$

- By $\Sigma = (\sigma_{x_i,x_j})_{i,j=1}^{m}$ we denote the ”true” co-variance matrix generating the data.
Correlation matrix and correlation network

- Correlation is computed from covariance by normalizing by the standard deviations $\sigma_x, \sigma_y$ of the variables

$$Corr(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{Cov(x, y)}{\sqrt{Cov(x, x)Cov(y, y)}}$$

- Empirical correlation matrix $R$ collects all pairwise empirical correlations among the variables

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

- Correlation network for the data is obtained from $R$ by defining a threshold $0 \leq \tau \leq 1$ and drawing an edge between vertex $v_j$ and $v_k$ if $r_{ij} \geq \tau$

- Different thresholds give different network
Weakness of correlation thresholding

- Obtaining the edges of the graph by thresholding the correlation(or covariance) matrix is simple
- However, the method but is not sensitive in detecting spurious correlations that are due to other (controlling) variables
- For example: Protein interactions $x - y$, and $x - z$ may reflect as correlation of $y - z$. However, the may not be no physical interaction between them
- Correlation is an inherently pairwise concept: Adding variables to the data does not have effect on correlation between existing vertices
Inverse covariance matrix

- An alternative approach is to rely on the inverse covariance matrix $\Theta = \Sigma^{-1}$
- Connection to conditional independence: if
  - Variables $X_1, \ldots, X_n$ have multivariate normal distribution, and
  - Variables $X_i$ and $X_j$ are conditionally independent given other variables $X_k, k \neq i, j$,

Then the inverse covariance $\theta_{ij} = 0$

- Inverse covariance has interpretation as *partial correlation* (remaining correlation after the effect of other variables has been removed): $\rho_{X_i, X_j \mid \setminus\{X_i, X_j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$
Graphical lasso

- Here we review a method introduced in Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics (2008), 9, 3 pp 432-441
- The name ‘lasso’ comes from ‘least absolute shrinkage and selection operator’ coined by Tibshirani in 1996
- Nowadays there is a large family of lasso methods, the common factor being the approach of penalizing the $\ell_1$-norm, $||w||_1 = \sum_j |w_j|$ of parameters to make feature selection and avoid overfitting
Aim in graphical lasso

- Intuitively: find inverse covariance matrix $\Theta = \Sigma^{-1}$ that produces a *good fit* to the data well but has a *lot of zeroes*.
- Concretely: maximize penalized log-likelihood of the data

$$\arg\max_{\Theta} \log L(x_1, \ldots, x_m; \Theta) - \rho \sum_{k=1}^{d} \sum_{l=1}^{d} |\theta_{kl}|$$

- The double sum is the $\ell_1$ norm of the matrix $\Theta$: sum of absolute values of the elements
- The larger the $\ell_1$ norm, the higher the penalty
- $\rho \geq 0$ is a co-efficient controlling the balance between the goodness-of-fit (log-likelihood) and the penalty.
  - High values of $\rho$ give very sparse matrices (low number of edges in the resulting graph), but less good fit to data
  - Low values of $\rho$ give denser matrices but may overfit the data
Log-likelihood of multivariate gaussian data

For a set multivariate normal distributed data points the log-likelihood is given by

$$\log L(x_1, \ldots, x_m; \Theta) = \log \left[ \prod_{i=1}^{m} (2\pi)^{-1/2} \det \Theta^{-1/2} \exp \left( (x_i - \mu)^T \Theta (x_i - \mu) \right) \right]$$

- $\Theta = \Sigma^{-1}$ controls the shape of the multivariate distribution
- The more co-variance, the more ellipsoid-shaped the distribution is
Log-likelihood of multivariate gaussian data

For a set multivariate normal distributed data points the log-likelihood is given by

\[
\log L(x_1, \ldots, x_m; \Theta) = \log \left[ \prod_{i=1}^{m} (2\pi)^{-\frac{1}{2}} (\det \Theta)^{\frac{1}{2}} \exp \left( (x_i - \mu)^T \Theta (x_i - \mu) \right) \right]
\]

- The expression within the \( \exp() \) gives the unnormalized density of the distribution at \( x_i \);
- \( \det \Theta \) is the determinant of matrix \( \Theta \);
- normalizing the equation to correspond to a log-probability
Penalized log-likelihood

With some manipulation (which we skip) and dropping constant terms the maximization problem can be equivalently be written as

\[
\arg \max_{\Theta} \left( \log \det \Theta - \langle S, \Theta \rangle_F - \rho \sum_{k=1}^{d} \sum_{l=1}^{d} |\theta_{kl}| \right)
\]

- \( S = \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T \) is the empirical covariance matrix
- \( \langle S, \Theta \rangle_F = \sum_{k=1}^{d} \sum_{l=1}^{d} s_{kl}\theta_{kl} \) the Frobenius inner product of the matrices (can be seen as similarity of two matrices)
- Intuitively: try to push \( S \) and \( \Theta \) away from each other while pushing as many coefficients towards zero as possible, and keeping the scores as valid log-probabilities
The type problem to be maximized has several optimization algorithms:

- Interior-point methods for "log det" type of problems (Yuan and Lin, 2007)
- Subgradient based approaches (Banerjee et al. 2007; Friedman et al. 2007), used in Graphical Lasso:
  - Iteratively optimizing single variable or small set of variables while keeping the others constant
  - Looking for steepest ascent direction of the objective, given by the (sub-)gradient
- Details of the algorithms are out of scope of this course (belongs to the field of non-linear optimization)
Cells were perturbed in 9 different ways, total of 7466 observations

The abundance of 11 Signaling proteins were observed for each cell

Sachs et al. performed a Bayesian network analysis of this data
Analysis with Graphical Lasso

- The data was used to build the empirical covariance matrix of the data
- Graphical Lasso was used to learn the inverse covariance matrix with different values for the penalization parameter
- Networks extracted with different parameter settings are shown
Review session of the extra exercise set: after the break from 11.15-12.00

Course exam: Monday 28.2 at 16.00, auditorium B123

Examined contents: lectures and exercises

Junker & Schreiber book is not explicitly examined, for background reading if you like.

Type of exam:
- 5 questions, 10 points each
- 1 question on short definitions of concepts, 2 essay questions, 2 technical questions (exercise type)

Thank you for all participating!