10/36-702 Homework 5
Due Friday 5/3/2013

1 Graphical Models

1. Consider a p-dimensional Gaussian graphical model $P_X \sim \mathcal{N}(0, \Sigma)$ defined on $X = (X_1, \dots, X_p)$. Let $\Omega = \Sigma^{-1}$ denote the precision matrix. In this problem, you will show that $\Omega_{ij} = 0$ if and only if X_i is conditionally independent of X_j given the remaining variables.

   (a) Partition $X = (Y, Z)$ where Y is a subset of the p variables and Z denotes the remaining variables. What is the conditional distribution $P(Y|Z)$?

   (b) Denoting the precision matrix in block for $\Omega = [\Omega_{YY}\Omega_{YZ}; \Omega_{ZY}\Omega_{ZZ}]$, show that $\Omega_{YY}^{-1} = \text{Var}(Y|Z)$. (Hint: Use the form of the inverse of a block matrix in terms of Schur completement)

   (c) Using the above two results, argue that $\Omega_{ij} = 0$ if and only if X_i is conditionally independent of X_j given the remaining variables.

   The above results motivate the Graphical Lasso (Glasso) algorithm. Suppose we have data $(X^1, \dots, X^n)$ where each X^i is a 0-mean p-dimensional multivariate Gaussian. Then we perform the optimization:

   $$\arg\min_{\Omega \in \mathcal{S}_p^+} \underbrace{-\log\det(\Omega) + \text{tr}(\Omega S_n)}_{\text{negative log likelihood}} + \underbrace{\lambda||\Omega||_1}_{\text{regularization}}$$

   where $\mathcal{S}_p^+$ is the set of all $p \times p$ positive semidefinite matrices, $S_n = \frac{1}{n}\sum_{i=1}^n X^i X^{iT}$ is the sample covariance, and $||\Omega||_1 = \sum_{i,j} |\Omega_{i,j}|$ is a ℓ_1 penalty on every element of Ω.

   The optimization produces an inverse covariance matrix with many zero-entries, which corresponds to a sparse graph.

2. Meinshausen and Buhlmann in 2006 derived an alternative method for estimating a sparse Gaussian graphical model. Recall from Stat 705:

   - Let X, Y be two random vectors and suppose we want to regress X onto Y. Then the expected-least-square regression function $m(X)$ is $\mathbb{E}[Y|X]$, i.e:

   $$\arg\min_{m: X \to Y} \mathbb{E}[||Y - m(X)||_2^2] = \mathbb{E}[Y|X]$$

   For this problem let $X = (X_1, \dots, X_p)$ be a p-dimensional random Gaussian vector.

   (a) Suppose we regress all X_j for $j \neq i$ onto the variable X_i. Prove that the expected-least-square regression function $m(\{X_j\}_{j \neq i})$ is $\sum_{j \neq i} \beta^i(j) X_j$ where $\beta^i(j) = -\frac{\Omega_{ij}}{\Omega_{ii}}$ (Hint: Use Schur Complement).

This motivates the multiple lasso procedure. Which solves:

$$\arg\min_{\beta^i \in \mathbb{R}^{p-1}} \frac{1}{2n} ||X_i - \sum_{j \neq i} \beta^i(j) X_j||_2^2 + \lambda ||\beta^i||_1$$

for all $i = 1, \ldots, p$. We then put an edge between X_i and X_j if either $\beta^i(j) \neq 0$ or $\beta^j(i) \neq 0$.

2 PCA

1. Let $X = [x_1, \ldots, x_n]$ be n centered data points in $\mathbb{R}^p$. Show that:

   $$\arg\min_{L} ||X - L||_2 \text{ s.t. } \operatorname{rank}(L) \leq k$$

   is equivalent to PCA. Here $||\cdot||_2$ denotes the spectral norm (i.e. $||A||_2 = \max_{x \neq 0} ||Ax||/||x||$) which is also the largest singular value.

2. PCA can be sensitive to gross (large in magnitude) corruptions of the data. Give an example of a dataset where changing a single entry in X can completely change the first principal component.

3. The formulation in pat 1 above can be modified to define a version of PCA that is stable under a few gross corruptions of the data. The robust PCA method is:

   $$\arg\min_{L,S} ||X - L - S||_2 \text{ s.t. } \operatorname{rank}(L) \leq k_1, \operatorname{card}(S) \leq k_2$$

   Derive a convex relaxation of this method.

4. Intuitively, justify why the solution to this problem might not be sensitive to corruptions in the data. Also, argue why the above formulation is different than sparse PCA discussed in class.

3 Density Clustering

Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a sample from a distribution P with density p. Let

$$L_t = \Big\{ x : \ p(x) > t \Big\}.$$

Let

$$\widehat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{||x - X_i||}{h} \right)$$

be the kernel density estimator. Define

$$\widehat{L}_t = \Big\{ x : \ \widehat{p}(x) > t \Big\}.$$

Define the Hausdorff distance

$$H(L_t, \widehat{L}_t) = \inf \Big\{ \epsilon : \ L_t \subset \widehat{L}_t \oplus \epsilon \quad \text{and} \quad \widehat{L}_t \subset L_t \oplus \epsilon \Big\}$$

where

$$A \oplus \epsilon = \bigcup_{x \in A} B(x, \epsilon)$$

and $B(x, \epsilon)$ denotes a ball of radius ϵ centered at x. Show that, with probability at least $1 - \delta$,

$$H(L_t, \widehat{L}_t) \le C_1 \sqrt{\frac{\log n}{nh^d}} + C_2 h.$$

Hint 1: You may assume any regularity conditions on p that you need. In addition to the usual smoothness conditions, you will need to assume that there are no "flat splots" in the density. In other words, assume that the gradient of p does not vanish near the boundary of the level set L_t.

Hint 2: You may use this fact: with probability at least $1 - \delta$,

$$||\widehat{p} - \overline{p}||_\infty \le C_1 \sqrt{\frac{\log n}{nh^d}}$$

where $\overline{p}(x) = \mathbb{E}(\widehat{p}(x))$.