Directed and Undirected Graphical Models **2013/2014**

Lecture 6 — November 6th

Lecturer: Guillaume Obozinski *Scribe: Lucas Plaetevoet, Ismael Belghiti*

6.1 Moment vector

Definition 6.1 (Moment vector) *We define the* moment vector *(or* moment parameter *as:*

$$\mu(\eta) = \nabla A(\eta) = E_\eta[\phi(X)].$$

6.1.1 Examples of moment vectors

Bernoulli

For a Bernoulli distribution, we can write:

$$p(x) = \pi^x(1-\pi)^{1-x} = e^{x\log\pi - x\log(1-\pi)+\log(1-\pi)} = e^{x\eta - A(\eta)}$$

with $\eta = \log\frac{\pi}{1-\pi}$ and $A(\eta) = -\log(1-\pi)$.

From this we get that $\pi = (1-\pi)e^\eta$ and thus $\pi = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}} = \sigma(\eta)$. Remark that in logistic regression we have $\eta = w^\top x$.

Moreover, we can write $A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$ and the moment vector is:

$$\mu(\eta) = E_\eta[\phi(X)] = E_\eta[X] = \pi.$$

Multinomial

In the multinomial case we consider $Z \to \{0,1\}^k$. We have $\phi(Z) = \begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix}$ and the moment vector is:

$$\mu(\eta) = E_\eta[\phi(Z)] = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \end{pmatrix}.$$

Gaussian

In the gaussian model, we have $\phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$ and we obtain:

$$\mu(\eta) = E_\eta \begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{pmatrix} \mu \\ \sigma^2 + \mu^2 \end{pmatrix}$$

6.2 Hessian of A

Proposition 6.2 *The hessian of A is the covariance matrix of the sufficient statistic:*

$$\nabla^2 A(\eta) = E[(\phi(X) - \mu(\eta))(\phi(X) - \mu(\eta))^\top] = Cov(\phi(X))$$

Proof We can write:

$$\begin{aligned}
\nabla^2 A(\eta) = \nabla\nabla A(\eta) = \nabla\left(\frac{\nabla Z(\eta)}{Z(\eta)}\right) &= \frac{\nabla^2 Z(\eta)}{Z(\eta)} + \nabla Z(\eta)\left(\frac{-\nabla Z(\eta)}{Z(\eta)^2}\right)^\top \\
&= \frac{\nabla^2 Z(\eta)}{Z(\eta)} - \left(\frac{\nabla Z(\eta)}{Z(\eta)}\right)\left(\frac{\nabla Z(\eta)}{Z(\eta)}\right)^\top
\end{aligned}$$

Moreover we have $[\nabla^2 Z(\eta)]_{k,k'} = E\left[\phi_k(X)\phi_{k'}(X)\right] Z(\eta)$ ie:

$$\nabla^2 Z(\eta) = E[\phi(X)\phi(X)^\top] Z(\eta).$$

Consequently:

$$\begin{aligned}
\nabla^2 A(\eta) &= E[\phi(X)\phi(X)^\top] - \mu(\eta)\mu(\eta)^\top \\
&= E[(\phi(X) - \mu(\eta))(\phi(X) - \mu(\eta)^\top] \\
&= Cov(\phi(X))
\end{aligned}$$

■

Remark: Z can be seen as a moment generating function $t \to Z(\eta + t)$ and A as the cumulative generating function $t \to A(\eta + t)$.

Corollary 6.3 *We have the three following properties:*

1. $\nabla^2 A(\eta) \succeq 0$ *(semi-positive definite).*
2. A *is convex.*
3. A *is strictly convex on* $\mathring{\Omega}$ *if, and only if,* $\phi(X)$ *is a minimal representation of the exponential family.*

Proof

1. $\forall c, c^\top \nabla^2 A(\eta) c = E[c^\top(\phi-\mu)(\phi-\mu)^\top c] = E[((\phi-\mu)^\top c)^2] \geq 0$
2. Since $\nabla^2 A \succeq 0$, A is convex.
3. If A is not strictly convex, then there exists η and c such that $c^\top \nabla^2 A(\eta) c = 0$ therefore, for all x, $\mathrm{Var}(c^\top \phi(x)) = 0$ thus $c^\top \phi(x) = -c_o$. We can thus write: $\forall x, c_0 + c_1\phi_1(x) + \ldots + c_k\phi_k(x) = 0$. Since we can go backward, we have the equivalence.

■

6.3 Log-Likelihood of an exponential function

Denoting $\bar{\phi} = \frac{1}{n}\sum_i \phi(x_i)$, we have:

$$-l(\eta) = -\eta^\top \bar{\phi} n + nA(\eta)$$

and

$$-\nabla l(\eta) = -\bar{\phi} n + n\mu(\eta).$$

Consequently, we have the following equivalence:

$$\nabla l(\eta) = 0 \Leftrightarrow \mu(\eta) = \bar{\phi}$$

Theorem 6.4 *The maximum likelihood estimator η is such that $\bar{\phi} = \mu(\eta)$. This result is called "Moment Matching".*

$$\boxed{\bar{\phi} = E_\eta[\phi(x)] = \mu(\eta)}$$

$$\eta \underset{\text{learning}}{\overset{\text{inference}}{\rightleftarrows}} \mu(\eta) = \bar{\phi}$$

6.4 Link between Maximum Likelihood and Maximum Entropy

The Maximum Entropy principle can be applied: we want to find the disribution p such that $E[\phi(X)] = \bar{\phi}$ and has maximal entropy.

We can write this as a convex optimization problem:

$$\begin{array}{|ll|}\hline \underset{p}{\text{Minimize}} & -H(p) \\ \text{subject to} & \begin{cases} E_p[\phi(X)] = \bar{\phi} \\ p(x) \geq 0 \\ \sum_x p(x) = 1 \end{cases} \\ \hline \end{array}$$

Let us introduce the corresponding Lagrangian:

$$\mathcal{L}(p, \lambda, c) = \sum_x p(x) \log p(x) - \lambda^\top \left(\sum_x p(x)\phi(x) - \bar{\phi} \right) + c \left(\sum_x p(x) - 1 \right)$$

Since the problem is convex, we have strong duality:

$$\min_p \max_{\lambda,c} \mathcal{L}(p, \lambda, c) = \max_{\lambda,c} \min_p \mathcal{L}(p, \lambda, c)$$

Slater's condition corresponds to the existence of p in the relative interior of the domain of the function that is in $\mathbb{R}_{+*}^{|X|}$ and such that $\sum_{x\in\mathcal{X}} p(x) = 1$. If we do not find such a p then we can reduce our set taken $\mathcal{X}' = \mathcal{X} \setminus \{x|p(x) = 0\}$.

Without loss of generality, we can hence assume that $p > 0$ and that the moment condition holds. The gradient of the Lagrangian with respect to p is given by:

$$\nabla_p \mathcal{L}(p, \lambda, c) = \log p(x) + 1 - \lambda^\top \phi(x) + c$$

and we have:

$$\begin{aligned} \nabla_p \mathcal{L} = 0 &\Leftrightarrow \log p(x) = \lambda^\top \phi(x) - (c + 1) \\ &\Leftrightarrow p(x) = Ce^{\lambda^\top \phi(x)} \text{ with } C = e^{-(c+1)} \end{aligned}$$

We recognize here an exponential family. Reinjecting this value of p and maximizing with respect to λ and c, we obtain the maximum likelihood estimator.

Theorem 6.5 *If $X_1, \ldots, X_n$ is an iid sample and $\phi(X)$ a statistic, then the maximum entropy estimator satisfying the equality $E_p[\phi(X)] = \bar{\phi}$ is the maximum likelihood distribution in the exponential family with sufficient statistic ϕ.*

6.5 Gaussian graphical models

6.5.1 Canonical parameterization

We consider a Gaussian random variable $X \in \mathbb{R}^p : X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$, $\Sigma \succ 0$. We recall the expression of its density:

$$p(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$

Denoting $\eta = \Sigma^{-1}\mu$ et $\Lambda = \Sigma^{-1}$ we get:

$$
\begin{aligned}
(x-\mu)^T \Sigma^{-1} (x-\mu) &= x^T\Sigma^{-1}x - x\mu^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu \\
&= x^T\Lambda x - 2\eta^T x + \eta^T\Lambda^{-1}\eta \\
p(x,\mu,\Lambda) &= \exp\left[\eta^T x - \frac{1}{2}x^T\Lambda x - A(\eta,\Lambda)\right] \\
A(\eta,\Lambda) &= \frac{1}{2}\eta^T\Lambda^{-1}\eta + \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Lambda|
\end{aligned}
$$

$\theta = \{\Lambda, \eta\}$ are the canonical parameters. Λ is called the *precision matrix*, and η is the *loading vector*. We have the following sufficient statistic, which is not a minimal representation:

$$
\Phi(x) = \begin{pmatrix} x \\ -\frac{1}{2}\operatorname{Vec}\left(xx^T\right) \end{pmatrix}
$$

Mean and covariance

The mean and covariance of X are given by :

$$
\begin{aligned}
\nabla_\theta A(\eta,\Lambda) &= \mathbb{E}_\theta[\Phi(X)] \\
&= \begin{pmatrix} \mathbb{E}_\theta[X] \\ -\frac{1}{2}\mathbb{E}_\theta\left[XX^T\right] \end{pmatrix} \\
\mathbb{E}_\theta[X] &= \nabla_\eta A(\eta,\Lambda) \\
&= \Lambda^{-1}\eta \\
&= \mu \\
-\frac{1}{2}\mathbb{E}_\theta\left[XX^T\right] &= \nabla_\Lambda A(\eta,\Lambda) \\
&= -\frac{1}{2}\Lambda^{-1}\eta\eta^T\Lambda^{-1} - \frac{1}{2}\Lambda^{-1} \\
&= -\frac{1}{2}\left[\mu\mu^T + \Lambda^{-1}\right]
\end{aligned}
$$

Hence

$$
\begin{aligned}
\operatorname{Cov}[X] &= \mathbb{E}_\theta\left[XX^T\right] - \mathbb{E}_\theta[X]\,\mathbb{E}_\theta[X]^T \\
&= \Lambda^{-1} \\
&= \Sigma
\end{aligned}
$$

Please note that we could have also computed the covariance with:

$$
\nabla^2_\theta A(\eta,\Lambda) = \begin{pmatrix} \operatorname{Cov}(X) & \dots \\ \dots & \operatorname{Cov}(\operatorname{Vec}(XX^T)) \end{pmatrix}
$$

and $\nabla^2_\eta A(\eta,\Lambda) = \Lambda^{-1}$

6.5.2 Conditioning and marginalization in Gaussian GM

We partition the random variable $X \in \mathbb{R}^p$ into two components $X_1 \in \mathbb{R}^{p_1}$ and $X_2 \in \mathbb{R}^{p_2}$ such that $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $p = p_1 + p_2$. We now seek to determine the law of X_1 and $X_2|X_1$.

$$X_1 \sim ?, \quad X_2|X_1 \sim ?$$

Before doing so, we need to partition the moment parameters μ, Σ and the canonical parameters Λ, η in the same way:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

from which we get a partitioned form for the joint distribution:

$$p(x_1, x_2) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \Lambda \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right] \tag{6.1}$$

In what follows, we will introduce a tool to block diagonalize partitioned matrices. We will then be able to develop general formulas for marginalization and conditioning in the multivariate Gaussian setting.

6.5.3 Digression on Schur complement

Let us consider the block matrix $M = \begin{pmatrix} A & L \\ R & U \end{pmatrix}$. Our goal is to explicit the blocks of its inverse in terms of the initial blocks A, L (L stands for left), U (U stands for upper) and R (R stands for right).

We can zero out the L and R by *premultiplying* M by D and *postmultiplying* by D. We denote Δ this block diagonal matrix.

$$\begin{aligned}
\begin{pmatrix} I & 0 \\ -RA^{-1} & I \end{pmatrix} \times \begin{pmatrix} A & L \\ R & U \end{pmatrix} \times \begin{pmatrix} I & -A^{-1}L \\ 0 & I \end{pmatrix} &= D \times M \times G \\
&= \begin{pmatrix} I & 0 \\ -RA^{-1} & I \end{pmatrix} \times \begin{pmatrix} A & 0 \\ R & U - RA^{-1}L \end{pmatrix} \\
\Delta &= \begin{pmatrix} A & 0 \\ 0 & U - RA^{-1}L \end{pmatrix}
\end{aligned}$$

Definition 6.6 *The* Schur complement *of the matrix* $M = \begin{pmatrix} A & L \\ R & U \end{pmatrix}$ *with respect to* A *is* $\left[M_{/A}\right] = U - RA^{-1}L$.

By symmetry we obtain the Schur complement of M with respect to U: $\left[M_{/U}\right] = A - LU^{-1}R$

Lemme 6.7 (Determinant lemma)

$$|M| = |A| \times \left|\left[M_{/A}\right]\right| = |U| \times \left|\left[M_{/U}\right]\right|$$

Proof

$$|\Delta| = \underbrace{|D|}_{=1} |M| \underbrace{|G|}_{=1} = |M|$$

and we have also

$$|\Delta| = |A| \left|\left[M_{/A}\right]\right|$$

and

$$|\Delta| = |U| \left|\left[M_{/U}\right]\right|$$

■

Lemme 6.8 (Positivity lemma) *If M is symmetric then $M \succcurlyeq 0$ if and only if $A \succcurlyeq 0$ and $\left[M_{/A}\right] \succcurlyeq 0$.*

Please note that we have the same lemma for strict inequalities.

Proof $G = D^T$. $A \succcurlyeq 0$ and $\left[M_{/A}\right] \succcurlyeq 0 \Leftrightarrow \forall x,\ x^T \Delta x \geqslant 0 \Leftrightarrow \forall x, \left(D^T x\right)^T M \left(D^T x\right) \geqslant 0$, hence $\forall y, y^T M y \geqslant 0$ because $G = D^T$ is invertible.

■

Woodbury-Sherman-Morrison inversion formula for partitioned matrices

We have that M is invertible if and only if $A \succcurlyeq 0$ and $\left[M_{/A}\right] \succcurlyeq 0$. Then $\Delta^{-1} = G^{-1} M^{-1} D^{-1}$, and $M^{-1} = G \Delta^{-1} D$. The explicit computation of this matrix product gives the so-called Woodbury-Sherman-Morrison formula:

$$\begin{aligned} M^{-1} &= \begin{pmatrix} I & -A^{-1}L \\ 0 & I \end{pmatrix} \times \begin{pmatrix} A^{-1} & 0 \\ 0 & \left[M_{/A}\right]^{-1} \end{pmatrix} \times \begin{pmatrix} I & 0 \\ -RA^{-1} & I \end{pmatrix} \\ &= \begin{pmatrix} A^{-1} + A^{-1}L\left[M_{/A}\right]^{-1} RA^{-1} & -A^{-1}L\left[M_{/A}\right]^{-1} \\ -\left[M_{/A}\right]^{-1} RA^{-1} & \left[M_{/A}\right]^{-1} \end{pmatrix} \end{aligned} \tag{6.2}$$

Similarly we obtain:

$$M^{-1} = \begin{pmatrix} \left[M_{/U}\right]^{-1} & -U^{-1}R\left[M_{/U}\right]^{-1} \\ -\left[M_{/U}\right]^{-1} LU^{-1} & U^{-1} + U^{-1}R\left[M_{/U}\right]^{-1} LU^{-1} \end{pmatrix}$$

6.5.4 Back to the problem

We now use the Woodbury formula (6.2) to compute an interesting expression for the quadratic form of the multivariate Gaussian distribution.

$$\begin{aligned}(x-\mu)^T\Sigma^{-1}(x-\mu) &= \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} I & -\Sigma_{11}^{-1}\Sigma_{12} \\ 0 & I \end{pmatrix} \dots \\ &\times \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \left[\Sigma_{/\Sigma_{11}}\right]^{-1} \end{pmatrix} \times \begin{pmatrix} I & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\ &= (x_1-\mu_1)^T(x_1-\mu_1) + (x_2-\mu_2-b)^T\left[\Sigma_{/\Sigma_{11}}\right]^{-1}(x_2-\mu_2-b)\end{aligned} \tag{6.3}$$

where we denoted $b = \Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1)$.
Now recall that we have $|\Sigma| = |\Sigma_{11}|\left|\left[\Sigma_{/\Sigma_{11}}\right]\right|$. The joint distribution can be expressed as:

$$\begin{aligned}p(x_1,x_2) = &\underbrace{\frac{1}{\sqrt{(2\pi)^{p_1}|\Sigma_{11}|}}\exp\left[-\frac{1}{2}\left((x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right)\right]}_{p(x_1)} \times \dots \\ &\underbrace{\frac{1}{\sqrt{(2\pi)^{p_1}\left|\left[\Sigma_{/\Sigma_{11}}\right]\right|}}\exp\left[-\frac{1}{2}\left((x_2-\mu_2-b)\left[\Sigma_{/\Sigma_{11}}\right]^{-1}(x_2-\mu_2-b)\right)\right]}_{p(x_2|x_1)}\end{aligned} \tag{6.4}$$

From (6.4) we deduce that $X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$, et $X_2|X_1 \sim \mathcal{N}\left(\mu_2+b, \left[\Sigma_{/\Sigma_{11}}\right]\right)$.

We denote by (μ_1, Σ_1), respectively $(\mu_{2|1}, \Sigma_{2|1})$, the moment parameters of the marginal distribution of x_1, respectively the moment parameters of the conditional distribution of x_2 given x_1. We have a similar notation for the canonical parameters η and Λ. We summarize our results in the following:

Moment parameterization summary

$$\begin{cases} \mu_1 = \mu_1 \\ \Sigma_1 = \Sigma_{11} \\ \mu_{2|1} = \mu_2 + b = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1) \\ \Sigma_{2|1} = \left[\Sigma_{/\Sigma_{11}}\right] = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{cases}$$

Canonical parameterization summary

$$\begin{cases} \eta_1 = \left[\Lambda_{/\Lambda_{22}}\right]\mu_1 = \eta_2 - \Lambda_{12}\Lambda_{22}^{-1}\eta_2 \\ \Lambda_1 = \Sigma_{11}^{-1} = \left[\Lambda_{/\Lambda_{22}}\right] = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} \\ \eta_{2|1} = \Lambda_{22|1} \times \mu_{2|1} = \Lambda_{22}\mu_2 - \Lambda_{21}(x_1-\mu_1) = \eta_2 - \Lambda_{21}x_1 \\ \Lambda_{22|1} = \Lambda_{22} \end{cases}$$

We can notice that in the moment parameterization, the marginalization operation is simple and the conditioning is complicated and the opposite holds in the canonical parameterization.

6.5.5 Zeros of the precision matrix and Markov properties

Let $p(x_1, ..., x_p)$ a joint Gaussian distribution. We denote $I = \{i, j\}$ and we consider $p\left(x_i, x_j | x_B\right)$, with $B = \{1, \ldots, p\} \setminus \{i, j\}$. Using the canonical parameterization:

$$\eta_I | B = \begin{pmatrix} \eta_i - \Lambda_{iB} x_B \\ \eta_j - \Lambda_{jB} x_B \end{pmatrix} \qquad \text{and} \qquad \Lambda_{II|B} = \Lambda_{II} = \begin{pmatrix} \lambda_{ii} & \lambda_{ij} \\ \lambda_{ji} & \lambda_{jj} \end{pmatrix}$$

we have the following expression for the covariance matrix of $X_I | X_B$:

$$\operatorname{Cov}\left(X_I | X_B\right) = \Sigma_{II|B} = \Lambda_{II|B}^{-1} = \frac{1}{|\Lambda_{II}|} \begin{pmatrix} \lambda_{jj} & -\lambda_{ji} \\ -\lambda_{ij} & \lambda_{ii} \end{pmatrix}$$

Hence $\operatorname{Cov}\left(x_i, x_j | X_B\right) = \frac{-\lambda_{ij}}{\sqrt{\lambda_{ii} \times \lambda_{jj}}}$ and $\lambda_{ij} = 0 \Rightarrow X_i \perp X_j \mid X_B$.

Proposition 6.9 *The non zero coefficients in Λ correspond to edges in the underlying graphical model.*

Indeed, the distribution is proportional to $\exp(\eta^T - \frac{1}{2} x \Lambda x^T) = \prod_i \exp(\eta_i x_i) \prod_{ij} \exp(-\frac{1}{2} x_i \lambda_{ij} x_j)$

6.5.6 Matrix inversion lemma

A useful consequence of the Schur component is to prove rigorously the following inversion lemma:

Lemme 6.10 *(Matrix inversion) Let $X \in \mathbb{R}^{p \times n}$*

$$\left(\operatorname{Id} + \lambda\, X^T X\right)^{-1} = \operatorname{Id} - \lambda X \left(\operatorname{Id} + \lambda\, X X^T\right)^{-1} X^T$$

In practice, we often want to invert matrix such as $\left(\operatorname{Id} + \lambda X^T X\right)$ where $X \in \mathbb{R}^{p \times n}$ is a *design* matrix. n represents an i.i.d sample while p represents the features, and we usually have $n \gg p$. In that case, the inversion lemma 6.10 replaces the problem of inverting a $n \times n$ matrix (complexity in $O(n^3)$) by a less costly one: inverting a $p \times p$ matrix.

Proof We consider $M = \begin{pmatrix} \operatorname{Id} & X \\ X^T & -\frac{1}{\lambda} \operatorname{Id} \end{pmatrix} = \begin{pmatrix} A & L \\ R & U \end{pmatrix}$, then $\left[M_{/U}\right]^{-1} = \left(\operatorname{Id} + \lambda X^T X\right)$. Recall the Woodbury formula (6.2), we have:

$$\left[M_{/U}\right]^{-1} = A^{-1} + A^{-1} L \left[M_{/A}\right]^{-1} R A^{-1}$$

which gives us the inversion lemma since here $\left[M_{/U}\right]^{-1} = \operatorname{Id} + X \left(-\frac{1}{\lambda} \operatorname{Id} - X X^T\right)^{-1} X^T$. ■