Entropy

Anders Hansson and Martin Andersen

Linköping University and Technical Unversity of Denmark

April 18, 2024

Chebyshev Bounds

Consider a probability space $(\Omega, \mathcal{F}, P)$ and a random variable $X : \Omega \to S \subseteq \mathbf{R}^n$.

Let the pdf of X be $p : S \to \mathbf{R}_+$.

Assume that

$$Ef_i(X) = \int_S f_i(x)p(x)dx = a_i, \qquad 0 \leq i \leq n$$

where $f_i : \mathbf{R}^n \to \mathbf{R}$ with $f_0(x) = 1$ and $a_0 = 1$, and where $a_i \in \mathbf{R}$ are given.

Chebyshev Bounds ctd.

We want to maximize $P(X \in C)$ over p for $C \subseteq S$ or equivalently with variable p solve

$$\begin{aligned} \text{maximize} \quad & \int_C p(x)dx \\ \text{subject to} \quad & \int_S p(x)f_i(x)dx = a_i, \quad 0 \leq i \leq n \\ & p(x) \geq 0, \quad \forall x \in S \end{aligned}$$

Let $\mathcal{D} = \{p \in \mathbf{R}^{\mathbf{R}^n} \mid \int_S p(x)dx = 1,\ p(x) \geq 0)$, and let $L : \mathcal{D} \times \mathbf{R}^{n+1} \to \mathbf{R}$ be defined by

$$L[p, \lambda] = \int_C p(x)dx + \sum_{i=0}^{n} \lambda_i \left(a_i - \int_S p(x)f_i(x)dx \right)$$

We then have that

$$\sup_p L[p, \lambda] = \sum_{i=0}^{n} \lambda_i a_i + \sup_p \left(\int_C (1 - f(x))p(x)dx - \int_{S \setminus C} f(x)p(x)dx \right)$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is given by $f(x) = \sum_{i=0}^{n} \lambda_i f_i(x)$.

Chebyshev Bounds ctd.

Then

$$\sup_p L[p, \lambda] \leq \sum_{i=0}^{n} \lambda_i a_i$$

if

$$1 - \inf_C f(x) \leq 0; \qquad - \inf_{S \setminus C} f(x) \leq 0$$

Since $Ef_i(X) = a_i$ we also have that $\sup_p P(X \in C) \leq \sum_{i=0}^{n} \lambda_i a_i$ under the same conditions. Can compute smallest possible such upper bound by solving with variable λ

$$\begin{aligned} \text{minimize} \quad & \sum_{i=0}^{n} a_i \lambda_i \\ \text{subject to} \quad & 1 - \inf_C f(x) \leq 0 \\ & - \inf_{S \setminus C} f(x) \leq 0 \end{aligned}$$

which is a convex problem, since $\inf_C f(z) = \inf_C \sum_{i=0}^{n} \lambda_i f_i(x)$ is convex in λ and similarly for second constraint.

Maximum Entropy Principle

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Entropy measures the amount of uncertainty in a probability distribution.

Assume that we observe values of a random variable $X : \Omega \to \mathbf{R}$ for outcomes of experiments and estimate the mean of the random variable.

What is then the most likely distribution of the random variable?

The *maximum entropy principle* says that it is the one that maximizes the entropy among all possible probability distributions that have the same estimated mean.

Entropy

Let $\Omega = \mathbf{N}_n$ and

$$\mathcal{D}_n = \left\{ p = (p_1, \ldots, p_n) \in [0,1]^n \mid p_k \geq 0,\ k \in \mathbf{N}_n;\ \sum_{k=1}^{n} p_k = 1 \right\}$$

be the set of finite probability functions for $X : \mathbf{N}_n \to \mathbf{R}$.

The entropy $H_n : \mathcal{D}_n \to \mathbf{R}$ should satisfy

1. $H_n(p) = H_n(\pi(p))$, where π is any permutation function.
2. $H_n(p) \leq H_n(\bar{p})$, where $\bar{p} = (1/n, \ldots, 1/n)$.
3. $H_n(p) = H_{n+1}(q)$, where $q = (p, 0)$
4. If $r \in \mathcal{D}_{mn}$ is a joint probability distribution with marginal distributions $p \in \mathcal{D}_n$ and $q \in \mathcal{D}_m$, then $H_{mn}(r) = H_n(p) + \sum_{k:p_k \neq 0} p_k H_m(r_k/p_k)$, where $r_k = (r_{k,1}, \ldots, r_{k,m})$.

$$\Longrightarrow$$

$$H_n(p) = -k \sum_{k=1}^{n} p_k \log p_k$$

Maximum Entropy Optimization Problem

$$\text{maximize} \quad -\sum_{k=1}^{n} p_k \ln p_k \tag{1}$$

$$\text{subject to } Ap = b \tag{2}$$

$$\mathbf{1}^T p = 1 \tag{3}$$

$$p \geq 0 \tag{4}$$

with variable p, where $A \in \mathbf{R}^{m \times n}$. *Convex Optimization Problem*

Expectation Constraints

Every row in the first constraint is an expectation constraint.

Define random variables a_i that takes values $A_{i,j}$ with probability p_j. Then the expected value of a_i is $\sum_{j=1}^{n} A_{i,j} p_j$, which is the left hand side of the ith row.

In an application the right hand side b_i could be obtained from an empirical estimate of the expected value.

Simple example of a_i is when $A_{i,j} = 1$ if $j = i$ and zero otherwise. Hence the expected value is p_j and if the right hand side is an empirical estimate of this expected value, we have effectively constrained p_j to be equal to its empirical estimate.

If a_i takes values f_j with probability p_j then by defining $A_{i,j} = f_j^2$ we define a constraint on the second moment of the random variable.

Example of Maximum Entropy Problem

Consider

$$\text{maximize } H_n(p) \tag{5}$$
$$\text{subject to } f^T p = b \tag{6}$$
$$\mathbf{1}^T p = 1 \tag{7}$$
$$p \geq 0 \tag{8}$$

with variable p, where $f \in \mathbf{R}^n$ and $b \in \mathbf{R}$ given.

Lagrangian $L : \mathbf{R}^n \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ where

$$L(p, \lambda, \mu) = H_n(p) + \lambda(b - f^T p) + \mu(1 - \mathbf{1}^T p)$$

is a concave function.

Lagrange Dual Function

Maximum of Lagrange function subject to $p_k \geq 0$ obtained when

$$\frac{\partial L}{\partial p_k} = -\ln p_k - 1 - \lambda f_k - \mu = 0$$

$$\Longrightarrow$$

$$p_k = \frac{e^{-\lambda f_k}}{e^{1+\mu}} \geq 0$$

Lagrange dual function $g : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ where

$$g(\lambda, \mu) = \frac{\sum_{k=1}^{n} e^{-\lambda f_k}}{e^{1+\mu}} + \lambda b + \mu$$

which is a convex function.

Minimization of Lagrange Dual Function

Minimum of g with respect to μ when

$$\frac{\partial g}{\partial \mu} = -\frac{\sum_{k=1}^{n} e^{-\lambda f_k}}{e^{1+\mu}} + 1 = 0$$

$$\Longrightarrow$$

$$\Phi(\lambda) := e^{1+\mu} = \sum_{k=1}^{n} e^{-\lambda f_k}$$

where $\phi : \mathbf{R} \to \mathbf{R}_{++}$.

Back substitution into the Lagrange dual function results in $h : \mathbf{R} \to \mathbf{R}$ defined by

$$h(\lambda) = \lambda b + \ln \Phi(\lambda)$$

Minimization of h

Minimum of h with respect to λ when

$$\frac{\partial h}{\partial \lambda} = b + \frac{\partial \Phi}{\partial \lambda} / \Phi(\lambda) = b - \frac{\sum_{k=1}^{n} f_k e^{-\lambda f_k}}{\Phi(\lambda)} = 0$$

$$\Longleftrightarrow$$

$$\sum_{k=1}^{n} (f_k - b)\, e^{-\lambda f_k} = 0$$

This equation has a solution λ if $f_k \neq f_l$ for some $k \neq l$, since then $G(\lambda) = \sum_{k=1}^{n} (f_k - b)\, e^{-\lambda f_k}$ is strictly decreasing as function of λ with

$$\lim_{\lambda \to -\infty} G(\lambda) = \infty, \qquad \lim_{\lambda \to +\infty} G(\lambda) = -\infty$$

Natural Parameters

If we do not want to parameterize the distribution in terms of b, then we can do it in terms of the *natural parameter* λ, i.e.

$$p_k = \frac{e^{-\lambda f_k}}{\sum_{l=1}^{n} e^{-\lambda f_l}}$$

Distribution is called the *Categorial, Gibbs* or *Boltzmann* distribution.

Belongs to the family of *Exponential distributions*

Normaliziation used in logistic regression:

$$p_k = \begin{cases} \frac{e^{z_k}}{1+\sum_{l=1}^{n-1} e^{z_l}}, & k \in \mathbf{N}_{n-1} \\ \frac{1}{1+\sum_{l=1}^{n-1} e^{z_l}}, & k = n \end{cases}$$

where $z_k = \lambda(f_n - f_k)$, $k \in \mathbf{N}_{n-1}$.

Quality of Lumber

Prices for lumber are $f_1 = 1$ for lowest grade, $f_2 = 1.1$ for middle grade, and $f_3 = 1.2$ for highest grade.

On average the price is $b = 1.05$.

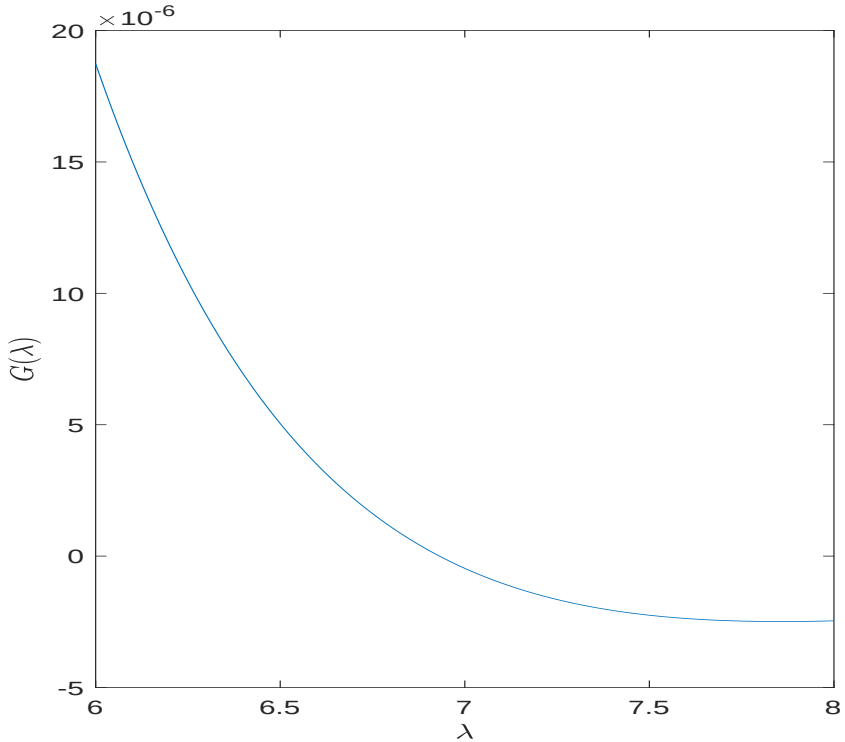
Use the maximum entropy principle to estimate what the probabilities are that low (p_1), middle (p_2) or high grade (p_3) lumber is delivered.

Moment constraint

$$1 \times p_1 + 1.1 \times p_2 + 1.2 \times p_3 = 1.05$$

$G(\lambda)$

$$G(\lambda) = (1-1.05)e^{-\lambda\times 1}+(1.1-1.05)e^{-\lambda\times 1.1}+(1.2-1.05)e^{(-\lambda\times 1.2)}$$



is zero for $\lambda = 6.9$, and hence

$$p_1 = 0.5386; \quad p_2 = 0.2701; \quad p_3 = 0.1913$$

Binary-Valued Random Vectors

Consider m-dimensional vectors X of random variables, and introduce the bijection $\mathcal{D}_x = \{0,1\}^m \leftrightarrow \mathbf{N}_n$, where $n = 2^m$ such that for $x \in \mathcal{D}_x$ and $k \in \mathbf{N}_n$ it holds that $k = \sum_{l=1}^{n} x_l 2^{l-1}$, where $x = (x_1, \ldots, x_m)$ with $x_l \in \{0,1\}$.

With abuse of notation we let $p(x) = p_k$, with $p = (p_1, \ldots, p_n) \in \mathcal{D}_n$.

Then $H_n(p) = -\sum_{x \in \mathcal{D}_x} p(x) \ln(p(x))$.

Maximum Entropy for Binary-Valued Random Vectors

Constraints on the first moment of X and on the cross-moments between the components of X:

$$\text{maximize } H_n(p) \tag{9}$$

$$\text{subject to } \sum_{x \in \mathcal{D}_x} x p(x) = m \tag{10}$$

$$\sum_{x \in \mathcal{D}_x} x x^T p(x) = M \tag{11}$$

$$\sum_{x \in \mathcal{D}_x} p(x) = 1 \tag{12}$$

$$p(x) \geq 0, \quad x \in \mathcal{D}_x \tag{13}$$

with variable p, where with abuse of notation m is the vector of first moments, and M is the matrix of second moments, where we do not specify the diagonal.

Lagrangian

Lagrangian $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{S}^m \times \mathbf{R} \to \mathbf{R}$ via

$$L(p, \lambda, \Lambda, \mu) = H_n(p) + \lambda^T \left(\sum_{x \in \mathcal{D}_x} x p(x) - m \right) + \mathbf{tr}\, \Lambda \left(\sum_{x \in \mathcal{D}_x} x x^T p(x) - M \right) + \mu \left(\sum_{x \in \mathcal{D}_x} p(x) - 1 \right)$$

where Λ has a zero diagonal.

Maximized when

$$\frac{\partial L}{\partial p(x)} = -\ln p(x) - 1 + \lambda^T x + \mathbf{tr}\, \Lambda x x^T + \mu = 0$$

Ising Distribution

$$p(x) = \exp\left(\lambda^T x + \mathbf{tr}\,\Lambda x x^T - 1 + \mu\right)$$

Lagrange dual function $g : \mathbf{R}^m \times \mathbf{S}^m \times \mathbf{R} \to \mathbf{R}$ by back substitution of p into the Lagrangian. Minimizing this function with respect to μ results in choosing μ such that the probabilities sum up to one. With $A = 1 - \mu$ this holds if

$$A = \ln \sum_{x \in \mathcal{D}_x} \exp\left(\lambda^T x + \mathbf{tr}\,\Lambda x x^T\right)$$

We see that A is a function of λ and Λ, and we write $A : \mathbf{R}^m \times \mathbf{S}^m \to \mathbf{R}$.

Equivalent Formulation

With the *energy function* $E : \{0,1\}^m \to \mathbf{R}$ given by

$$E(x) = -\lambda^T x - \mathbf{tr}\,\Lambda x x^T$$

where Λ has a zero diagonal, we have

$$p(x) = \exp\left(-E(x) - A(\lambda, \Lambda)\right)$$

where

$$A(\lambda, \Lambda) = \ln \sum_x \exp(-E(x))$$

Relation to Moments

Back-substitute $\mu = 1 - A(\lambda, \Lambda)$ into the Lagrange dual function and obtain $h : \mathbf{R}^m \times \mathbf{S}^m \to \mathbf{R}$ given by

$$h(\lambda, \Lambda) = A(\lambda, \Lambda) - \lambda^T m - \mathbf{tr}\, \Lambda M$$

Minimum of h with respect to (λ, Λ) when

$$\frac{\partial h}{\partial \lambda} = \frac{\partial A}{\partial \lambda} - m = 0$$
$$\frac{\partial h}{\partial \Lambda} = \frac{\partial A}{\partial \Lambda} - M = 0$$

Partial Derivatives

$$\frac{\partial A}{\partial \lambda} = \frac{\sum_{x \in \mathcal{D}_x} x \exp\left(-E(x)\right)}{\sum_{x \in \mathcal{D}_x} \exp\left(-E(x)\right)} = \frac{\sum_{x \in \mathcal{D}_x} x \exp\left(-E(x) - A(\lambda, \Lambda)\right)}{\sum_{x \in \mathcal{D}_x} \exp\left(-E(x) - A(\lambda, \Lambda)\right)}$$
$$= \sum_{x \in \mathcal{D}_x} x p(x)$$
$$\frac{\partial A}{\partial \Lambda} = \frac{\sum_{x \in \mathcal{D}_x} x x^T \exp\left(-E(x)\right)}{\sum_{x \in \mathcal{D}_x} \exp\left(-E(x)\right)} = \frac{\sum_{x \in \mathcal{D}_x} x x^T \exp\left(-E(x) - A(\lambda, \Lambda)\right)}{\sum_{x \in \mathcal{D}_x} \exp\left(-E(x) - A(\lambda, \Lambda)\right)}$$
$$= \sum_{x \in \mathcal{D}_x} x x^T p(x)$$

We should match the moments!

To solve the above equations with respect to (λ, Λ) is not easy in general, and specifically not when the dimension of x is large.

Static Ranking of Web Pages

Let $V = \mathbf{N}_n$ be a set of web pages or vertices in a graph $G = (V, E)$, where the edge set $E \subset V \times V$ contains directed edges (i, j) describing that there is a link from web page i to web page j.

PageRank uses a Markov Chain model for which the probabilities of the out-links from a page are equal. Let p_{ij} be the transition probability that a user at page i goes to page j. These probabilities are given by

$$p_{ij} = \begin{cases} \frac{1}{d_i}, & (i,j) \in E \\ 0, & (i,j) \notin E \end{cases}$$

where d_i is the out-degree of node i, i.e., the number of edges $(i, j) \in E$ where $j \in V$.

The PageRank is defined as the stationary distribution of the Markov Chain, i.e. as the solution of $\pi^T = \pi^T P$, where $P \in \mathbf{R}^{n \times n}$ is the matrix of transition probabilities p_{ij}.

Alternative Model based on Network Flow

Let y_{ij} be the number of users following link $(i,j) \in E$ per unit time. Assume the web traffic is in a state of equilibrium so that the traffic out of a node is equal to the traffic in per unit time, i.e.

$$\sum_{j:(i,j)\in E} y_{ij} = \sum_{j:(j,i)\in E} y_{ji}, \qquad i \in V$$

Define

$$H_j = \sum_{i:(i,j)\in E} y_{ij}, \quad j \in V$$

which is the number of "hits" per unit time at node j. Also define

$$Y = \sum_{(i,j)\in E} y_{i,j} = \sum_{j\in V} H_j$$

which is the total number of hits per unit time.

Moment Constraints

Define the probabilities $p_{ij} = y_{i,j}/Y$, and we may then write the equilibrium condition as

$$\sum_{j:(i,j)\in E} p_{ij} = \sum_{j:(j,i)\in E} p_{ji}, \qquad i \in V$$

(These probabilities are not the transition probabilities discussed before. They are obtained by normalization with $\sum_{j:(i,j)\in E} p_{ij}$.) One solution to the above equations is

$$p_{ij} = \frac{H_i}{Yd_i}, \qquad (i,j) \in E$$

which agrees with the traditional PageRank model as described above after normalization.

There are many more solutions to the equilibrium condition, which can be interpreted as moment constraints.

Maximum Entropy for Static Ranking of Web Pages

The maximum entropy solution under the moment constraints is obtained from

$$\begin{aligned}
\text{maximize} \quad & -\sum_{(i,j)\in E} p_{ij} \ln p_{ij} \\
\text{subject to} \quad & \sum_{j:(i,j)\in E} p_{ij} = \sum_{j:(j,i)\in E} p_{ji}, \qquad i \in V \\
& \sum_{(i,j)\in E} p_{ij} = 1 \\
& p_{ij} \geq 0, \qquad (i,j) \in E
\end{aligned}$$

with variable p.

Entropy for Infinite Probability Distributions

For infinite probability distribution $p : \mathbf{R} \to \mathbf{R}_+$ define entropy as

$$H(p) = -\int_{-\infty}^{+\infty} p(x) \log p(x) dx$$

Normal Distribution

Let $\mathcal{D}$ be the space of real-valuled functions defined on $\mathbf{R}^n$ and consider $p \in \mathcal{D}$ subject to $p(x) \geq 0$, $\int_{\mathbf{R}^n} p(x)dx = 1$, $\int_{\mathbf{R}^n} xp(x)dx = m$ and $\int_{\mathbf{R}^n} xx^T p(x)dx = M$, where $m \in \mathbf{R}^n$ and $M \in \mathbf{S}_+$ are the first and second moments of the distribution, respectively. The maximum entropy problem is

$$\begin{aligned} &\text{maximize } H(p) && (14)\\ &\text{subject to } \int_{\mathbf{R}^n} xp(x)dx = m && (15)\\ &\int_{\mathbf{R}^n} xx^T p(x)dx = M && (16)\\ &\int_{\mathbf{R}^n} p(x)dx = 1 && (17)\\ &p(x) \geq 0, \quad x \in \mathbf{R}^n && (18) \end{aligned}$$

with variable $p \in \mathcal{D}$.

Lagrangian

Lagrangian $L : \mathcal{D} \times \mathbf{R} \times \mathbf{S}^n \times \mathbf{R} \to \mathbf{R}$ with

$$L(p, \lambda, \Lambda, \mu) = -\int_{\mathbf{R}^n} p(x) \ln p(x) dx + \lambda^T \left(m - \int_{\mathbf{R}^n} x p(x) dx \right) \tag{19}$$

$$+ \frac{1}{2} \mathbf{tr} \left(\Lambda \left(M - \int_{\mathbf{R}^n} x x^T p(x) dx \right) \right) \tag{20}$$

$$+ \mu \left(1 - \int_{\mathbf{R}^n} p(x) dx \right) \tag{21}$$

First variation of the Lagrangian:

$$\delta L = -\int_{\mathbf{R}^n} \delta p \left[\ln p + 1 + \lambda^T x + \frac{1}{2} x^T \Lambda x + \mu \right] dx$$

which should be non-positive for all δp when p is optimal. Hence expression in bracket must vanish and optimal distribution is

$$p(x) = e^{-1-\mu-\lambda^T x - \frac{1}{2} x^T \Lambda x} \geq 0$$

Lagrange Dual Function

Lagrange dual function $g : \mathbf{R} \times \mathbf{S}^n \times \mathbf{R} \to \mathbf{R}$ defined by

$$g(\lambda, \Lambda, \mu) = \int_{\mathbf{R}^n} e^{-1-\mu-\lambda^T x-\frac{1}{2}x^T \Lambda x} dx + \lambda^T m + \frac{1}{2}\,\mathbf{tr}\,\Lambda M + \mu \tag{22}$$

$$= \int_{\mathbf{R}^n} e^{-1-\mu+\frac{1}{2}\lambda^T \Lambda^{-1}\lambda-\frac{1}{2}\left(x+\Lambda^{-1}\lambda\right)^T \Lambda\left(x+\Lambda^{-1}\lambda\right)} dx \tag{23}$$

$$+ \lambda^T m + \frac{1}{2}\,\mathbf{tr}\,\Lambda M + \mu \tag{24}$$

and we determine the μ that minimizes it by setting the partial derivative of g with respect to μ equal to zero, which is equivalent to that $p(x)$ integrates to one.

Normalization

$$\int_{\mathbf{R}^n} p(x)dx = \int_{\mathbf{R}^n} e^{-1-\mu+\frac{1}{2}\lambda^T\Lambda^{-1}\lambda-\frac{1}{2}\left(x+\Lambda^{-1}\lambda\right)^T\Lambda\left(x+\Lambda^{-1}\lambda\right)}dx \quad (25)$$

$$= \frac{2^{n/2}e^{-1-\mu+\frac{1}{2}\lambda^T\Lambda^{-1}\lambda}}{\det\Lambda^{1/2}} \int_{\mathbf{R}} e^{-\bar{x}_1^2}d\bar{x}_1 \times \cdots \times \int_{\mathbf{R}} e^{-\bar{x}_n^2}d\bar{x}_n \quad (26)$$

$$= \frac{e^{-1-\mu+\frac{1}{2}\lambda^T\Lambda^{-1}\lambda}(2\pi)^{n/2}}{\sqrt{\det\Lambda}} = 1 \quad (27)$$

Hence $\mu = \frac{1}{2}\lambda^T\Lambda^{-1}\lambda - 1 - \frac{1}{2}\ln\frac{\det\Lambda}{(2\pi)^n}$ and we get

$$p(x) = \sqrt{\frac{\det\Lambda}{(2\pi)^n}}e^{-\frac{1}{2}\left(x+\Lambda^{-1}\lambda\right)^T\Lambda\left(x+\Lambda^{-1}\lambda\right)}$$

Back Substitution

Optimal μ into the Lagrange dual function defines $h : \mathbf{R}^n \times \mathbf{S}^n \to \mathbf{R}$ via

$$h(\lambda, \Lambda) = g\left(\lambda, \Lambda, \frac{1}{2}\lambda^T\Lambda^{-1}\lambda - 1 - \frac{1}{2}\ln\frac{\det\Lambda}{(2\pi)^n}\right) \tag{28}$$

$$= \frac{1}{2}\lambda^T\Lambda^{-1}\lambda - \frac{1}{2}\ln\frac{\det\Lambda}{(2\pi)^n} + \lambda^T m + \frac{1}{2}\mathbf{tr}\,\Lambda M \tag{29}$$

Minimizing h from:

$$\frac{\partial h}{\partial \lambda} = \Lambda^{-1}\lambda + m = 0 \tag{30}$$

$$\frac{\partial h}{\partial \Lambda} = -\frac{1}{2}\Lambda^{-1}\lambda\lambda^T\Lambda^{-1} - \frac{1}{2}\Lambda^{-1} + \frac{1}{2}M = -\frac{1}{2}mm^T - \frac{1}{2}\Lambda^{-1} + \frac{1}{2}M = 0 \tag{31}$$

Non-Natural Parameters

With $\Sigma = M - mm^T$ it holds that $\Lambda = \Sigma^{-1}$ and we have $\lambda = -\Sigma^{-1}m$ and

$$p(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)}$$

In case h is expressed in terms of (m, Σ) instead of in terms of (λ, Λ) it will not be a convex function.

Natural parameters make convex!