Maximum entropy methods, covariance completion and applications

Luca Barbiero

Department of Information Engineering

University of Padua

A thesis submitted for the

Bachelor degree in Information engineering

September 23rd, 2011

Supervisor: Michele Pavon

Day of the defense:

Signature from head of committee:

Contents

1 Introduction **1**

2 Preliminaries **3**

2.1 Random vectors 3

2.2 Multivariate normal distribution 4

2.3 Entropy 5

2.3.1 Information divergence 6

2.4 Lagrange multipliers 7

2.4.1 Lagrangian 8

3 Introduction to maximum entropy methods **9**

3.1 Heuristic 9

3.1.1 Boltzmann's dice 10

3.2 Formal approach 11

3.2.1 The minimun discrimination information principle 12

4 Covariance Selection **15**

4.1 The problem 15

4.2 A rule 16

4.3 Link to the general framework 17

4.3.1 Generalization to Matrix Completion Problems 17

5 Quasi-Newton methods **19**

5.1 Newton's step 19

5.1.1 Minimization and maximization problems 20

5.1.1.1 The multivariate case 20

5.2 Approximation 20
5.2.1 Entropy approach 21

References **23**

1

Introduction

The aim of this thesis is to give an insight on the motivations as well as the applications of the *maximum entropy methods* in Information Theory. Such techniques, altought assuming different aspects, all apply the dogmatic principle of maximum entropy introduced by the physicist Edwin Thompson Jaynes in 1957. After a brief recall of some mathematical tools in chapter 2, we'll introduce first heuristically, and then formally, the motivation of the entropic approach, also showing with a famous example that such approach is closely in agreement with nature: this is the true reason that motivates the wide spectrum of application it finds. Subsequently, in Chapter 4 we'll focus on an apparently disjoint context, that of matrix completion, referring to the work of the statistician Arthur P. Dempster who in 1972 with his "Covariance Selection" theory [4] gave rise to a whole stream of research in that field. We'll observe that despite the different formulation, Dempster's work is nothing but an application of the maximum entropy principle and what is even more interesting is that it opens the doors of a general matrix completion approach, regardless of the origin of need of a completion, that can come from 1) a lack of reliable information as well as 2) a goal of computational saving. Finally, remaining in the contex of matrix completion we'll threat a case of the second type, which in turn comes with a different appearence with respect to that presented at the end of the previous chapter but, again, applying the same original idea.

2

Preliminaries

We first recall some mathematical tools that will turn out to be useful for our aims.

2.1 Random vectors

Random vectors (rve) are the multivariate extension of random variables (rv). A rve $\mathbf{X} = (X_1, ..., X_n)^T \in \mathbb{R}^n$ is a map

$$\mathbf{X} : \Omega \to \mathbb{R}^n, \qquad \omega \mapsto \mathbf{X}(\omega) = (X_1(\omega), ..., X_n(\omega))^T. \tag{2.1}$$

The probability measure induced by $\mathbf{X}$ on $\mathbb{R}^n$

$$P(\mathbf{X} \in E), \qquad E \subset \mathbb{R}^n, \tag{2.2}$$

fully characterizes the rve $\mathbf{X}$ in a statistical sense. The distribution of $\mathbf{X}$ is in turn characterized from the multidimensional cumulative distribution function (CDF)

$$F_{\mathbf{X}}(x_1, ..., x_n) = P(X_1 < x_1, ..., X_n < x_n). \tag{2.3}$$

Analysing the CDF we can distinguish among discrete, continuous and mixed rve. For instance, consider the continuos case; in particular, the absolutely continuous CDFs are a subclass of the continuous CDFs that admit the propability density function (PDF) $f_{\mathbf{X}}(x_1, ..., x_n)$. In his points of continuity, the PDF is obtained from the CDF by derivation

$$f_{\mathbf{X}}(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 ... \partial x_n} F_{\mathbf{X}}(x_1, ..., x_n). \tag{2.4}$$

The usefulness of the PDF is that, when it exists, it reduces the calculus of probability to a multiple integration

$$P(\mathbf{X} \in E) = \int ... \int_E f_{\mathbf{X}}(x_1, ..., x_n) \ dx_1 ... dx_n. \tag{2.5}$$

The PDF doesn't always exist, as previously mentioned. In what follows, we'll make frequent use of the concept of mean vector and covariance matrix of a rve. The expected value of $\mathbf{X}$ is the vector in $\mathbb{R}^n$

$$E[\mathbf{X}] = (E[X_1], ..., E[X_n])^T \tag{2.6}$$

The covariance matrix (or simply covariance when it's clear that we're in a multidimensional context) is the matrix in $\mathbb{R}^{n \times n}$

$$\Sigma = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T], \tag{2.7}$$

in which the (i, j) element is $\sigma_{ij} = cov(X_i, X_j)$ i.e. the covariance between the ith and the jth component of the rve. Obviously, when $i = j$ we denote with σ_{ii} the variance of the ith component. The covariance matrix is symmetric (because $cov(X_i, X_j) = cov(X_j, X_i)$) and positive definite, in fact for every $\mathbf{a} \in \mathbb{R}^n$

$$\begin{aligned} \mathbf{a}^T \Sigma \mathbf{a} &= \mathbf{a}^T E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T] \mathbf{a} \\ &= E[\mathbf{a}^T (\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T \mathbf{a}] \\ &= var(\mathbf{a}^T \mathbf{X}) \geq 0 \end{aligned} \tag{2.8}$$

where we used the linearity of expectation.

2.2 Multivariate normal distribution

A random vector $\mathbf{X} \in \mathbb{R}^n$ is said to have a multivariate normal distribution if

1. every linear combination of its components $Y = a_1 X_1 + ... + a_n X_n$ is normally distributed

2. there exists a random l-vector $\mathbf{Z}$ whose components are indipendent standard normal random variables, a n-vector μ and a $n \times l$ matrix A, such that $\mathbf{X} = A\mathbf{Z} + \mu$. In words, every multivariate normal distribution is an affine transformation of the so called normal standard multivariate.

Then, if the covariance matrix Σ is nonsingular, the PDF of $\mathbf{X}$ exists and can be expressed analitically as

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{2.9}$$

We remark that, like in the unidimensional case, the multivariate normal distribution is fully determined by its mean vector μ and covariance matrix Σ. Moreover, since a normal distribution can be made zero mean subtracting its mean (which can be derived by empirical experiments), we stress the fact that it's the covariance matrix Σ which characterizes the distribution, and observe that it's the inverse of the covariance matrix

$$\Sigma^{-1} = \begin{bmatrix} \sigma^{11} & \cdots & \sigma^{1n} \\ \vdots & \ddots & \vdots \\ \sigma^{n1} & \cdots & \sigma^{nn} \end{bmatrix} \tag{2.10}$$

that appears in the analytical expression of the distribution, were the σ^{ij} are its components.

2.3 Entropy

Entropy is a measure of randomness or, more precisely, unpredictability associated with random vectors (univariate random variables are special cases of rve). The higher the entropy, the smaller is our ability to predict events a priori: We say that high entropy means that we gain (on the average) high information when an outcome occours, hence we can think of this central concept as, in the end, a quantification of our ignorance about random phenomena. In particular, the case in which our ignorance about a rve is maximum is when its probability distribution is uniform over an interval i.e. every outcome is equally likely and we have no further information about them before the experiment.

Consider a discrete rve $\mathbf{X} \in \mathbb{R}^n$ (the continuous case being anologuos) with a finite sample space $\mathcal{X}$ of cardinality M and a valid probability mass function (PMF) for it, which we indicate here and in what follows for the ease of notation (for both continuous and discrete distributions), simply with p. A consistent entropy function $H(p)$ on the space of the probability distributions must satisfy the following properties:

1. if $\mathbf{X}$ is a.s. costants then $H(p) = 0$, otherwise $H(p) > 0$

2. if p^* is uniform over its alphabet, i.e. $p_i^* = \frac{1}{M} \quad i = 1...M$, then $p^* = argmax H(p)$, otherwise $H(p) < H(p^*)$

The above properties formalize the heuristic intuition we discussed previously. Now we introduce the analytical form of entropy proposed by C.E. Shannon in 1952.

$$H(p) = -\sum_{i=1}^{M} p_i \log p_i, \tag{2.11}$$

where $0 \log 0 = 0$ by definition. Note that entropy is associated with a PMF and does not depend on the sample space of the rve. This measure for the entropy of a distribution satisfies at the properties we stated, in particular it has a unique global maximum. Note that the base of the logarithm it's not important, provided it's greater than 1: in statistical mechanics, base e is used, instead in Information Theory base 2 is preferred (so the entropy of a fair coin is 1 bit, the unit measure of the information)

2.3.1 Information divergence

We present now a very powerful instrument, introduced by Kullback and Leibler in 1951 [6]. Condider two valid probability distributions (again we focus on discrete distributions: the continuous case can be treated substituting sums with integrals) p and q with the only restriction that the support of p is rigorously contained in the support of q

$$q_i = 0 \Rightarrow p_i = 0 \qquad \forall i \tag{2.12}$$

The information divergence, or relative entropy or KL-index of q from p is defined to be

$$\mathbb{D}(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{2.13}$$

Note that $\mathbb{D}(\cdot||\cdot)$ does not induce a metric in the space of probability distributions since it's not symmetric and, most important, it does not satisfy the triangular inequality. Nevertheless, it enjoyes two properties

1. $\mathbb{D}(p||q) \geq 0$,

2. $\mathbb{D}(p||q) = 0 \quad$ if and only if $\quad p = q$.

Put in another way, we can see the information divergence as a pseudo-distance of p from q, in some sense. The case in which q is a uniform distribution is interesting: if so, as seen before, q is characterized by having maximum entropy among all possible distributions with sample space of cardinality M and the smaller the divergence of q from p, the higher the entropy of p. In fact observe that

$$\mathbb{D}(p||q) = \sum_{i=1}^{M} p_i \log np_i = \log n + \sum_{i=1}^{M} p_i \log p_i = H_{max} - H(p) \tag{2.14}$$

It's easy to see that $\mathbb{D}(p||q) \to 0 \quad$ when $\quad H(p) \to H_{max}$.

2.4 Lagrange multipliers

The method of Lagrange multipliers provides a strategy for finding the maxima and minima of a function subject to constraints. Note that in this section we move out from the field of random phenomena to recall some results from Analysis, so that here $X \subset \mathbb{R}^n$ is an open set and $\Gamma \subset \mathbb{R}^n$ is a constraint defined to be

$$\Gamma = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{g}(\mathbf{x}) = \mathbf{b}\} \tag{2.15}$$

where $\mathbf{b} = (b_1, ..., b_m)$ is fixed and $\mathbf{g} : X \to \mathbb{R}^m$ is a C^1 function of components $\mathbf{g} = (g_1, ..., g_m)$ Let's recall briefly the main results in this field of Analysis

Definition 1. *A point $\mathbf{x}^* \in \Gamma$ is said to be a relative maximum (resp. minimum) constrained to Γ for a function $f : X \to \mathbb{R}^n$ if it exists a neighborhood $\mathcal{U}$ of $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ (resp. $f(\mathbf{x}^*) \leq f(\mathbf{x})$) $\forall \mathbf{x} \in \mathcal{U} \cap \Gamma$. A constrained maximum or minimum is also called constrained extreme.*

We're now ready to state the main result of this section. Recall that $\mathbf{x}^* \in \Gamma$ is said to be a regular point of 2.15 if $\nabla g(\mathbf{x}^*) \neq \mathbf{0}$.

Theorem (on Lagrange multipliers) 1. *Let $\mathbf{x}^* \in \Gamma$ be a regular point of Γ of constrained extreme for a function $f : X \to \mathbb{R}^n$ differentiable in $\mathbf{x}^*$. Then there exist $\lambda_1 ... \lambda_m \in \mathbb{R}$ such that*

$$\nabla f(\mathbf{x}^*) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*). \tag{2.16}$$

In particular, $\lambda_1 ... \lambda_m$ are called the Lagrange multipliers for the constrained extreme problem.

It follows that constrained maxima and minima must be sought between the irregular points of the constraint, and the regular ones which satisfy 2.16. In particular, if Γ is made only of regular points, the problem of constrained extreme consists in the solution of the $n+m$ system with $n+m$ unknowns

$$\begin{cases} \partial_{x_j} f(x_1, ..., x_n) = \sum_{i=1}^{m} \lambda_i \partial_{x_j} g_1(x_1, ..., x_n) & j = 1, ..., n \\ g_i(x_1, ..., x_n) = b_i & i = 1, ..., m \end{cases} \tag{2.17}$$

2.4.1 Lagrangian

The function $\mathcal{L} : X \times \mathbb{R}^m \to \mathbb{R}^n$ defined to be

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \langle \lambda, \mathbf{g}(\mathbf{x}) - \mathbf{b} \rangle = f(x_1, ..., x_n) - \sum_{i=1}^{m} \lambda_i [g_i(x_1, ..., x_n) - b_i] \tag{2.18}$$

is called lagrangian of the constrained extreme problem. The following results follows from the Lagrange multipliers theorem.

Corollary 1. *Let f be a C^1 function, having a local extreme contrained to Γ in $\mathbf{x}^*$ and let $\mathbf{x}^*$ a regular point of Γ. Then there exist $\lambda = (\lambda_1, ..., \lambda_n)$ such that $(\mathbf{x}^*, \lambda)$ is a free critical point for $\mathcal{L}$*

Proof. If $\mathbf{x}^* \in \Gamma$ is a local constrained extreme for f, then there exists $\lambda = (\lambda_1, ..., \lambda_n) \in \mathbb{R}^m$ such that $x_1^*, ..., x_n^*, \lambda_1^*, ..., \lambda_n^*$ are solutions of the system 2.17. Because

$$\begin{aligned} \partial_{x_j} \mathcal{L}(\mathbf{x}, \lambda) &= \partial_{x_j} f(x_1, ..., x_n) - \sum_{i=1}^{m} \lambda_i \partial_{x_j} g_1(x_1, ..., x_n) \\ \partial_{\lambda_i} \mathcal{L}(\mathbf{x}, \lambda) &= g_i(x_1, ..., x_n) - b_i \end{aligned} \tag{2.19}$$

this is equivalent in stating that $(\mathbf{x}^*, \lambda) \in X \times \mathbb{R}^m$ is a free critical point for $\mathcal{L}$. □

The mathematical usefulness of the lagrangian is now clear: it reconducts a constrained extreme problem to an unconstrained one.

3

Introduction to maximum entropy methods

In Information Theory, the maximum entropy principle is a postulate which states that, in model fitting problems, when subject to known constraints (or incomplete information) the probability distribution that best represents the current state of knowledge is the one with the largest entropy.

3.1 Heuristic

For many decades it has been recognized through evidences in theoretic advancements as in applicative results that the notion of entropy defines a kind of measure on the space of the probability distributions, such that those of high entropy are in some sense preferable over others. The justification for this was stated in a variety of intuitive forms: higher entropy distribution represent more "disorder", they are "smoother", "more probable", "less predictable", "they assume less", according to Shannon's interpretation of entropy as an information measure. In all these keywords, the recurrent idea is that in a model fitting task, given some incomplete informations, it seems the best choiche to determine the model in a way that it allows the widest spectrum of behaviors compatible with the constraints, and this is precisely what we're accomplishing when we maximize entropy taking into account any constraints: we choose a model that describes the experimental evidences obtained, without (erroneously) unbalancing it on specific behaviors according to inexistent grounds: it is well know that tending to

maximum entropy means tending to the uniform distribution, that is over all that of complete ignorance.

3.1.1 Boltzmann's dice

Suppose that n dice are thrown on a table. We are faced with the task of determining the frequencies $p_i = \frac{n_i}{n}$ i.e. n_i, the number of dice showing face i. In absence of any experimental evidence (no contraints) we're led to choose a priori the uniform distribution, which assigns $p_i = 1/6, i = 1, ..., 6$. Indeed there's no reason to think that any face is more probable of any other or, put in another way, it would seems highly irrational to make any other estimate than the uniform one. Suppose now we're given the following experimental evidence: the total number of spots showing is $n\alpha$

$$\sum_{i=1}^{6} i n_i = n\alpha. \tag{3.1}$$

Note that from (3.1) it follows that

$$\sum_{i=1}^{6} i \frac{n_i}{n} = \alpha = E[X] \tag{3.2}$$

where X is the random variable which denote the number of spots shown by one dice. Consider the general case in which $E[X]$ differs from 3.5, the well known expected value of spots shown by a fair dice. Now the uniform distribution is not suitable to fit the model. One way to proceed is to count the number of ways that n dice can fall so that n_i dice show face i. There are

$$\binom{n}{n_1, ..., n_6} = \frac{n!}{n_1!...n_6!} \tag{3.3}$$

such ways, where (3.3) is the multinomial coefficient, which in combinatorics is the number of ways in which an n-elements set can be partitioned in 6 disjoint sets each having n_i $i = 1, ..., 6$ elements. This macrostate is indexed by $(n_1, ..., n_6)$ corresponding to (3.3) microstates, each one having probability $\frac{1}{6^n}$. We wish to maximize (3.3) in order to find the most probable macrostate, under the constraint (3.1). Using a crude

Stirling's approximation, $n! \approx (\frac{n}{e})^n$, we find that

$$\begin{aligned}
\binom{n}{n_1, ..., n_6} &\approx \frac{(\frac{n}{e})^n}{\prod_{i=1}^{6}(\frac{n_i}{e})^{n_i}} = \frac{\frac{n^n}{e^n}}{e^{-n}\prod_{i=1}^{6} n_i^{n_i}} = \frac{n^n}{\prod_{i=1}^{6} n_i^{n_i}} \\
&= \prod_{i=1}^{6}(\frac{n}{n_i})^{n_i} = \exp(\ln \prod_{i=1}^{6}(\frac{n}{n_i})^{n_i}) = \exp(\ln n^n \prod_{i=1}^{6} \frac{1}{n_i^{n_i}}) \\
&= \exp(n \ln n + \ln \prod_{i=1}^{6} \frac{1}{n_i^{n_i}}) = \exp(n \ln n - \sum_{i=1}^{6} n_i \ln n_i) \\
&= \exp(\sum_{i=1}^{6} n_i \ln n - \sum_{i=1}^{6} n_i \ln n_i) = \exp[\sum_{i=1}^{6} n_i(\ln n - \ln n_i)] \\
&= \exp(-\sum_{i=1}^{6} n_i \ln \frac{n_i}{n} = \exp[n(-\sum_{i=1}^{6} \frac{n_i}{n} \ln \frac{n_i}{n})] \\
&= \exp[nH(\frac{n_1}{n}, ..., \frac{n_6}{n})].
\end{aligned} \tag{3.4}$$

By the monotonicity of the exponential, under the constraint (3.1), maximizing (3.3) is almost equivalent to maximize $H(\frac{n_1}{n}, ..., \frac{n_6}{n})$ i.e. the entropy of the distribution to determine. Thus, the distribution of maximum entropy is the one that *can be realized in the greatest number of ways*: since the only constraint we have is the mean value of spot showing, determinig the frequencies (i.e. the PMF) taking into account such a constraint but maximizig the entropy is a very good idea because in so doing our model leaves open the wider set of behaviors. Moreover, for large n, the overwhelming majority of all possible distributions compatible with our information have entropy very close to the maximum and when $n \to \infty$ any frequency distribution other than the one of maximum entropy become highly atypical of those allowed by the constraints. This is the central results that come from Jaynes' Concentration Theorem in [1].

3.2 Formal approach

The formal framework of any maximum entropy method (ME) was introduced by Jaynes in [3] as follows. We discuss the univariate case for the ease of the treatment, without loss of generality. Consider a rv X, its sample space $\mathcal{X}$ and the three entities:

1. a valid probability distribution $p = \{p_i\}_{i=1,...,n}$, $\qquad \sum_{i=1}^{n} p_i = 1$;
2. a consistent entropy measure, for example that of Shannon $H(p) = -\sum_{i=1}^{n} p_i \ln p_i$;

3. a set of linear constraints $\sum_{i=1}^{n} p_i g_r(x_i) = a_r, \qquad r = 1, ..., m.$

Notice that althought widely used, Shannon's entropy measure is not the only one: what is really important is that we take a consistent measure for entropy, as discussed in (2.3). Furthermore, we remark that the constraints must be linear: they are usually moment constraints. We are faced with a constrained extreme problem (see 2.4) in which we have to maximize entropy (i.e. a function) subject to a set of linear contraints (that with a multidimensional notation we called Γ in (2.4)) The lagrangian (2.18) of the problem is:

$$\mathcal{L}(p_1, .., p_n, \lambda_0, .., \lambda_m) = -\sum_{i=1}^{n} p_i \ln p_i - (\lambda_0 - 1)[\sum_{i=1}^{n} p_i - 1] - \sum_{r=1}^{m} \lambda_r [\sum_{i=1}^{n} p_i g_r(x_i) - a_r] \quad (3.5)$$

maximizing $\mathcal{L}$, i.e. imposing

$$\begin{cases} \partial_{p_i} \mathcal{L} = -(\ln p_i + 1) - \sum_{r=1}^{m} \lambda_r g_r(x_i) - (\lambda_0 - 1) = 0 & i = 1, ..., n \\ \partial_{\lambda_r} \mathcal{L} = -(\sum_{i=1}^{n} p_i g_r(x_i) - a_r) = 0 & r = 1, ..., m \end{cases} \quad (3.6)$$

we obtain that

$$p_i = \exp[-(\lambda_0 + \lambda_1 g_1(x_i) + ... + \lambda_m g_m(x_i))] \qquad i = 1, ..., n \quad (3.7)$$

while the equations on the partial derivatives in λ_r simply lead back to the constraints. In order to determine the Lagrange multipliers, we substitute (3.7) into the contraints equations to get the m + 1 (nonlinear) equation in m + 1 unknows system:

$$\begin{cases} e^{\lambda_0} = \sum_{i=1}^{n} exp[-(\lambda_1 g_1(x_i) + ... + \lambda_m g_m(x_i))] \\ a_r e^{\lambda_0} = \sum_{i=1}^{n} g_r(x_i) exp[-(\lambda_1 g_1(x_i) + ... + \lambda_m g_m(x_i))] & r = 1, ..., m \end{cases} \quad (3.8)$$

that can find a solution via numerical methods. Again, we remark that the continuous case can be treated symply by substituting sums with integrals: no convergence problems arise, since entropy is a bounded, smooth functtion.

3.2.1 The minimun discrimination information principle

The minimun discrimination information principle (MDI) from Kullback extends the framework introduced by Jaynes. Suppose we substitute the entropy measure as second entity with the Information divergence (2.13). Now we seek a constrained minimum instead of a maximum but what is really interesting is that now we have a fourth entity

in the new framework: the distribution q. From the MDI point of view, ME seeks to determine that distribution p, out of those that satisfy the constraints, for which $\mathbb{D}(p||u)$ is a minimum, with u denoting the uniform distribution. Kullback's MDI extends this concept. It seeks to minimize the relative entropy $\mathbb{D}(p||q)$, which means it seeks to determine the distribution p that satisfies the constraints and is closest to a given distribution q. This fourth entity, say a "settable reference distribution" of maximum entropy in absolute makes MDI more flexible than Jaynes' ME and allows, as we will see, interesting applications in contexts that seems not to have so much in common with probability distributions.

4

Covariance Selection

We discuss now the covariance selection theory introduced by Dempster in [4].

4.1 The problem

Suppose we are faced with the task of fitting a model known to be described by a multivariate normal distribution (2.9). Recall that the normal distribution has the welcome property to be fully determined by its second order description, i.e. its mean vector and covariance matrix, but actually only by the second one by reducing it to a zero mean distribution. So the fitting procedure consists in determining the covariance structure

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix} \tag{4.1}$$

i.e. the set of parameters σ_{ij} $i, j = 1, ..., n$. Tipically, we have a sample of m n-variate observations $\mathbf{x}_1, ..., \mathbf{x}_m$ and so an estimated $n \times n$ sample covariance matrix S derived using the formula

$$S = \frac{1}{m} \sum_{l=1}^{m} (\mathbf{x}_l - \bar{\mathbf{x}})^T (\mathbf{x}_l - \bar{\mathbf{x}}) \tag{4.2}$$

where

$$\bar{\mathbf{x}} = \frac{1}{m} \sum_{l=1}^{m} \mathbf{x}_l. \tag{4.3}$$

However, the computational ease with which the set of parameters can be estimated should not lead us to obscure the unwisdom of such estimation from limited data. Hence, we identify a subset of parameters whose reliability we trust from the data

and look for a valid completion of the covariance structure. The insight that underlies Dempster's covariance selection is the principle of parsimony in parametric model fitting, which suggests that parameters should be introduced only when the data indicate they are required. Note that in (2.9) what appears is not the covariance matrix Σ but its inverse Σ^{-1} so that *parameters reduction* may resonably be attempted by setting certains σ^{ij} to 0. Parameters reduction involves a tradeoff between benefits and costs: annihilating a substantial number of parameters the amount of noise in a fitted model due to estimation error is significantly reduced but, on the other hand, errors of misspecification are introduced because the null values are incorrect: every decision to fit a model involves an implicit balance between these two kinds of errors.

4.2 A rule

Let I be a subset of the index pairs (i, j) with $1 \leq i \leq j \leq n$ and J the set of remaining pairs. Think about J as the set of entries whose reliability we trust and I the complementary set of parameters. The formal rule that concretizes the insight given in the previous section is the following.

Rule 1. *Choose $\hat{\Sigma}$ to be the positive definite symmetric matrix such that S and $\hat{\Sigma}$ are identical for index pairs $(i, j) \in J$ while $\hat{\Sigma}^{-1}$ is identically 0 for index pairs $(i, j) \in I$.*

This choice, which we name Dempster's completion, may at first look less natural than setting the unspecified elements of Σ to zero. It has nevertheless considerable advantages compeared to the latter [4]. Dempster established the following far reaching result.

Theorem 1. *Assume that a symmetric, positive-definite completion of Σ exists. Then there exists a unique Dempster's Completion Σ^0. This completion maximizes the entropy*

$$H(p) = -\int_{\mathbb{R}^n} \log(p(\mathbf{x}))p(\mathbf{x})d\mathbf{x} = \frac{1}{2}\log(\det\Sigma) + \frac{1}{2}n(1 + \log 2\pi) \tag{4.4}$$

among zero-mean Gaussian distributions having the prescribed elements $\sigma_i j, (i, j) \in J$.

Thus, Dempster's Completion Σ^0 solves a *maximum entropy* problem, i.e., maximizes entropy under linear constraints [7].

4.3 Link to the general framework

Dempster's covariance selection revisits from a different point of view the former work of Jaynes. In fact, instead of determining a probability distribution solving a constrained extreme problem, he thought in terms of parameters reduction, but the underlying idea is the same: given incomplete information on the model, a good way of fitting it is that to leave open the wider spectrum of possible behaviors. This target is accomplished in both cases even if they appear not to have so much in common (actually, it seems that maximum entropy is a consequency in Dempster's work instead of a goal). But this is not the case. In fact, it can be easily seen that the incomplete information on the covariance structure is nothing but a set of linear constraints on the distribution, while the fact that it was assumed a priori for the distribution to be a (multivariate) normal one is not restrictive as it can be shown that if the linear constraints are the second order description (mean vector and covariance matrix) the maximum entropy distribution is normal [5]. Finally, the fact that Rule 1 leads to the maximum entropy normal distribution follows from Theorem 1 which summarizes Dempster's Statistical Theory.

4.3.1 Generalization to Matrix Completion Problems

Dempster's Covariance Selection is in conclusion just one, although if really important, task of matrix completion. Here the *original problem* is the unwisdom affecting collected data: this is the reason for which we start with a subset of entries of the matrix and need to find a valid completion. As we will see in the following chapter, this entropic approach is well suited in other matrix completion contexts. We'll focus on a different *original problem*, that of reducing a significant computational burden. Observe in Theorem 1 that maximizing entropy of a normal distribution is equivalent, apart from constant factors and considering the monotonicity of the logarithm, to extremizing $\det \Sigma$: we can think about every symmetric, positive-definite matrix as the covariance structure of a multivariate normal distribution and apply Rule 1 to it. Furthermore, M. Pavon and A. Ferrante proved in [7] that symmetry and positive-definiteness are not necessary since the constrained extremization of the determinant only involves the positive part of the matrix. Hence such approach can be extended really to every matrix, also in the rectangular case.

5

Quasi-Newton methods

In numerical analysis, Newton's method is an algorithm for finding successively better approximations to the roots of a smooth, real valued function. The idea of the method is as follows: one starts with an initial guess which is reasonably close to the true root, then the function is approximated by its tangent line (which can be computed using the tools of calculus), and one computes the x-intercept of this tangent line (which is easily done with elementary algebra). This x-intercept will typically be a better approximation to the function's root than the original guess, and the method can be iterated.

5.1 Newton's step

Consider for the ease of exposition the unidimensional case. Let's $X \subset \mathbb{R}$ a compact set, $f : X \to \mathbb{R}$ a differentiable function that takes values in $\mathbb{R}$. Suppose we have some current approximation for the position of one root, say x_n. Then the formula for a better approximation x_{n+1} is derived as follows from the definition of the derivative

$$f'(x_k) = \frac{\Delta y}{\Delta x} = \frac{f(x_k) - 0}{x_k - x_{k+1}} \quad k \geq 0 \tag{5.1}$$

Then by use of simple algebra we get

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad k \geq 0 \tag{5.2}$$

We should start with some arbitrary initial value x_0: the closer to the root, the better. In absence of any intuition about where the zero might lie, we could spread out different

initial possibilities in a reasonably small interval appealing to the intermediate value theorem.

5.1.1 Minimization and maximization problems

Newton's method can be easily extended to maxima and minima problems: actually, it's sufficient to ask for f to be twice differentiable, and look for the roots of its first derivative, according to Fermat's theorem on stationary points

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(k_n)} \qquad k \geq 0 \tag{5.3}$$

5.1.1.1 The multivariate case

In a multivariate context (by far the most interesting case, where we'll concentrate in the next section), i.e. $X \subset \mathbb{R}^n, f : X \to \mathbb{R}$ and under the hypotesis $f \in C^2$, (5.3) becomes

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [Hf(\mathbf{x}_k)]^{-1} \nabla f(\mathbf{x}_k) \qquad k \geq 0 \tag{5.4}$$

where $\nabla f(\mathbf{x}_k)$ and $Hf(\mathbf{x}_k)$ are respectively the gradient and the Hessian matrix of f at $\mathbf{x}_k$.

5.2 Approximation

In the execution of the algorithm, the most expensive part (computationally speaking) is finding, storing and inverting the Hessian. Quasi-Newton methods seek to approximate the Hessian matrix (or its inverse) for the kth step by accumulating information from the preceding steps using only first derivatives (or they finite-difference approximation) [8]. Consider the second order Taylor expansion

$$f(\mathbf{x}_k + \Delta\mathbf{x}_k) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T \Delta\mathbf{x}_k + \frac{1}{2}\Delta\mathbf{x}_k Hf(\mathbf{x}_k)\Delta\mathbf{x}_k, \qquad \Delta\mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k. \tag{5.5}$$

Taking the gradient on both sides respect to $\Delta\mathbf{x}_k$, we get

$$\nabla f(\mathbf{x}_k + \Delta\mathbf{x}_k) \approx \nabla f(\mathbf{x}_k) + Hf(\mathbf{x}_k)\Delta\mathbf{x}_k. \tag{5.6}$$

Let B_k be an approximation of $Hf(\mathbf{x}_k)$ (B_0 is usually taken to be the identity). In QN one employs the Newton's step (5.4) with $Hf(\mathbf{x}_k) := B_k$ imposing in view of (5.6) the secant equation

$$\nabla f(\mathbf{x}_k + \Delta\mathbf{x}_k) = \nabla f(\mathbf{x}_k) + B_k\Delta\mathbf{x}_k. \tag{5.7}$$

In more than one dimension, the secant equation is under determined. Various methods are used to find a symmetric B_{k+1} closest (according to some metric) to the current approximation B_k and satisfying (5.7). The underlying idea in all QN is that of avoiding to calculate the Hessian for every Newton's step, approximating it by rank one (or even rank two) updates specified by gradient evaluations. Historically, remarkable examples of QN are the DFS formula from Davidon–Fletcher–Powell (the first updating scheme proposed), BFGS from Broyden–Fletcher–Goldfarb–Shanno and the SR1 (Symmetric Rank 1) method.

5.2.1 Entropy approach

Consider now the case where f is a strongly convex function, i.e.

$$Hf(\mathbf{x}_k) > \alpha I_n, \qquad \exists \alpha > 0, \quad \forall k > 0, \tag{5.8}$$

in this case, B_k should be positive definite. Recall from section 2.3.1 the definition of relative entropy and its interpretation. In the case of two multivariate normal distributions p, q with covariance matrixes respectively P, Q (2.13) has a close form

$$\begin{aligned}
\mathbb{D}(p||q) &= \int \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \\
&= \int \log\{\frac{|P|^{-1/2}}{|Q|^{-1/2}} \exp[-\frac{1}{2}\mathbf{x}^T(P^{-1} - Q^{-1})\mathbf{x}]\} p(\mathbf{x}) d\mathbf{x} \\
&= \int \log |PQ^{-1}|^{-1/2} + [-\frac{1}{2}\mathbf{x}^T(P^{-1} - Q^{-1})\mathbf{x}]\} p(\mathbf{x}) d\mathbf{x} \\
&= \frac{1}{2}\log |P^{-1}Q| + \int \frac{1}{2}\mathbf{x}^T(Q^{-1} - P^{-1})\mathbf{x} p(\mathbf{x}) d\mathbf{x} \\
&= \frac{1}{2}[\log |P^{-1}Q| + \int tr(Q^{-1} - P^{-1})\mathbf{x}\mathbf{x}^T p(\mathbf{x}) d\mathbf{x}] \\
&= \frac{1}{2}[\log |P^{-1}Q| + tr(Q^{-1} - P^{-1}) \int \mathbf{x}\mathbf{x}^T p(\mathbf{x}) d\mathbf{x}] \\
&= \frac{1}{2}[\log |P^{-1}Q| + tr(Q^{-1} - P^{-1})P] \\
&= \frac{1}{2}[\log |P^{-1}Q| + tr[(Q^{-1}P) - I_n]] \\
&= \frac{1}{2}[\log |P^{-1}Q| + tr(Q^{-1}P) - n].
\end{aligned} \tag{5.9}$$

Notice that $\mathbb{D}(p||q)$ uniquely depends on the covariance matrixes P, Q and so, with an abuse of notation, we introduce

$$\mathbb{D}(P||Q) = \frac{1}{2}[\log |P^{-1}Q| + tr(Q^{-1}P) - n]. \tag{5.10}$$

that can be thought as a (pseudo) metric between (symmetric and positive definite) matrixes.
This result gives rise to an important application of MDI of section 3.2.1, which we know to be a refinement of the original ME. Consider the minimization problem

$$\min \mathbb{D}(B_{k+1}||B_k) \tag{5.11}$$

subject to the linear constraint

$$B_{k+1}\Delta\mathbf{x}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k). \tag{5.12}$$

Here we are faced with the task of finding the *nearest* matrix B_{k+1}, i.e. the update of the current approximation of the Hessian, according to the generalized entropic approach of section 4.3.1, using the current approximation B_k and satisfying the linear constraint given by the secant equation (5.12). The lagrangian of the problem is

$$\mathcal{L}(B_{k+1}, \lambda_{k+1}) = \frac{1}{2}[\log|B_{k+1}^{-1}B_k| + tr(B_k^{-1}B_{k+1}) - n] + \lambda_{k+1}^T[B_{k+1}\Delta\mathbf{x}_k - \nabla f(\mathbf{x}_{k+1}) + \nabla f(\mathbf{x}_k)] \tag{5.13}$$

Imposing $\delta\mathcal{L}(B_{k+1}, \lambda_{k+1}, \delta B) = 0$ for all δB we get

$$(B_{k+1})^{-1} = B_k^{-1} + 2\Delta\mathbf{x}_k\lambda_{k+1}^T. \tag{5.14}$$

This is the step on which it's possible to construct iterative schemes to update cyclically B_{k+1}^{-1} and λ_k. Note that in (5.14) $(B_{k+1})^{-1}$ is a rank one update of B_k^{-1}, just like any *conventional* QN.
The maximum entropy approach shows in this application all its versatility: we're not considering a model fitting task but an optimization one. We should not forget anyway that what underlies (5.10) is a (pseudo) metric defined on the space of probability distributions and the matrixes involved in $\mathbb{D}(\cdot||\cdot)$ must be thought as the covariance matrixes of multivariate normal distributions: not by chance at the beginning of this section we posed as condition for the function f to be strongly convex in the region of interest, this allows the Hessian to gain positive-definiteness, in addition to symmetry, that's a property held by every Hessian matrix.

References

[1] Edwin T. Jaynes, *On The Rationale of Maximum-Entropy Methods*, Proceedings of the IEEE, vol. 70, 1982 11

[2] H.K. Kesavan, J.N. Kapur, *The Generalized Maximum Entropy Principle*, IEEE transactions on systems, man, and cybernetycs, vol. 19, 1989.

[3] Edwin T. Jaynes, *Information Theory and Staistical Mechanics*, Physical Reviews, vol. 106, 1957. 11

[4] A. P. Dempster, *Covariance Selection*, Biometrics, vol. 28, 1972. 1, 15, 16

[5] Thomas M. Cover, Joy A. Thomas, *Elements of information theory*, Wiley, 1938 17

[6] Kullback , Leibler, *On Information and Sufficiency*, Annals of Mathematical Statistics 22, 1951 6

[7] M. Pavon, A. Ferrante, *Matrix Completion à la Dempster by the principle of Parsimony*, IEEE Transactions on Information Theory, 2011 16, 17

[8] P.E. Gill, W. Murray, *Numerical methods for constrained optimization*, Academic Press, 1974 20