Integrating inconsistent data in a probabilistic model*

Jiří Vomlel
vomlel@utia.cas.cz

Laboratory for Intelligent Systems
University of Economics,
Prague, Czech Republic

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague, Czech Republic

Abstract

In this paper we discuss *knowledge integration* as a process of building a joint probability distribution from an input set of low-dimensional probability distributions starting with an initial joint probability distribution. Since the solution of the problem for a consistent input set of probability distributions is known we concentrate on a setup where the input probability distributions are inconsistent. In this case the iterative proportional fitting procedure (IPFP), which converges in the consistent case, tends to come to cycles. We propose to use an algorithm, which we call GEMA (an abbreviation of Generalized Expectation Maximization Algorithm), that converges even in inconsistent case to a reasonable joint probability distribution. The important property of GEMA is that it can be efficiently implemented exploiting decomposability of considered distributions.

*This work was supported by the Grant Agency of the Czech Republic through grant nr. 201/02/1269

1 Introduction

In this paper we discuss the process of integrating contradicting data using the framework of classical probability. We define *knowledge integration* as a process of building a *joint probability distribution* Q from a set of low-dimensional probability distributions $\mathcal{P} = \{P_1, \ldots, P_k\}$, $k \in \mathbb{N}$ and from an *initial joint probability distribution* Q_0. To model a chosen domain we will use discrete random variables X_i indexed by natural numbers from $V = \{1, \ldots, n\} \subset \mathbb{N}$. Each low-dimensional probability distribution $P_j, j = 1, \ldots, k$ is defined on variables $\{X_\ell\}_{\ell \in E_j}, E_j \subseteq V$. *Knowledge integration* should yield a joint probability distribution $Q(X_1, \ldots, X_n)$ that would embody available knowledge about the chosen domain. The general problem of the construction of probability distributions with given marginals has a long history starting with the papers by Kellerer [6] and Vorobev [12].

Empirical observations of Dawes et al. [2] support the basic idea underlying the *knowledge integration* process. In their paper they compared clinical and actuarial approach to decision making. In the clinical approach to decision making, decisions are based on mental processes of the trained human judge. In the actuarial approach to decision making, a decision is made on a model of a judgment that uses the same information as the human judge in the clinical approach. Dawes et al. observed that in many decision making domains actuarial models are more accurate in making predictions about complex outcomes than are trained human judges.

Also, Osherson et al. [10] argue that a coherent probability model that is fitted as closely as possible to the incoherent judgments of an informant will often be closer to actual relative frequencies than the raw judgments of the informant.

If there exists a probability distribution $Q(X_1, \ldots, X_n)$ whose marginals equal to low-dimensional probability distributions $P_i \in \mathcal{P}, i = 1, \ldots, k$ then we say that probability distributions in set $\mathcal{P}$ are *consistent*. Otherwise we say that they are *inconsistent*.

Example 1 In Table 1 we give an example (borrowed from [5]) of an input set of three distributions $\mathcal{P} = \{P_1(X_1, X_2), P_2(X_2, X_3), P_3(X_1, X_3)\}$. If $\varepsilon =$

Table 1: Input set of distributions

$P_j, j = 1, 2$	$X_{j+1} = 1$	$X_{j+1} = 2$
$X_j = 1$	$\frac{1}{2} - \varepsilon$	ε
$X_j = 2$	ε	$\frac{1}{2} - \varepsilon$

P_3	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	ε	$\frac{1}{2} - \varepsilon$
$X_2 = 2$	$\frac{1}{2} - \varepsilon$	ε

$\frac{4}{20}$ then there exist probability distributions $Q(X_1, X_2, X_3)$ whose marginals equal to probability distributions P_1, P_2, and P_3 (e.g., one given in Table 2). Therefore the probability distributions in set $\mathcal{P}$ are consistent. If $\varepsilon = \frac{3}{20}$ then

Table 2: Joint probability distribution $Q(X_1, X_2, X_3)$

	$X_2 = 1$		$X_2 = 2$	
Q	$X_3 = 1$	$X_3 = 2$	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{3}{20}$
$X_1 = 2$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{3}{20}$

there does not exist any distribution $Q(X_1, X_2, X_3)$ whose marginals equal to probability distributions P_1, P_2, and P_3, i.e., the probability distributions in set $\mathcal{P}$ are inconsistent. $\diamond$

Next we introduce necessary notation. We will use small letter x_i to denote a value of variable X_i. The set of all possible values of variable X_i will be denoted $\mathbb{X}_i$. We will use X to denote multidimensional random variable $(X_1, \ldots, X_n)$, x will denote a value of X and $\mathbb{X}$ the set of all values of X. Let for a set $A \subseteq V$ the symbol X^A denote the multidimensional variable $\{X_i\}_{i \in A}$, x^A denote a value of X^A, and Q^A denote the marginal distribution of Q on variables X^A,i.e., $Q^A(x^A) = \sum_{x\prime : x'^A = x^A} Q(x')$. Recall, that each low-dimensional probability distribution $P_j, j = 1, \ldots, k$ from $\mathcal{P}$ is defined on a variables $X^{E_j}, E_j \subseteq V$.

Definition 1 Let P, Q be two probability distributions. If $\{x \in \mathbb{X}, P(x) > 0\} \supseteq \{y \in \mathbb{X}, Q(y) > 0\}$ then P is dominated by Q (denoted as $P \ll Q$).

Definition 2 I-divergence

$$I(P \parallel Q) = \begin{cases} \sum\limits_{x \in \mathbb{X},\, P(x) > 0} P(x) \log \dfrac{P(x)}{Q(x)} & \text{if } P \ll Q \\ +\infty & \text{if } P \not\ll Q \end{cases}$$

I-divergence geometry, introduced by Csiszár [1] is an effective instrument enabling us to study iterative procedures based on the operations of I-projections to a set of probability distributions $\mathcal{Q}$.

Definition 3 $Q^\star$ is I-projection of a probability distribution P to a set $\mathcal{Q}$ if for all $Q \in \mathcal{Q}$ it holds that

$$I(Q \parallel P) \ \geq \ I(Q^\star \parallel P) \ .$$

We will use $\pi(P, \mathcal{Q})$ to denote I-projection of P to $\mathcal{Q}$.

Let $\mathcal{S}_j$ denote the set of probability distributions having their marginal equal to probability distribution P_j, i.e., $\mathcal{S}_j = \{Q : Q^{E_j} = P_j\}$ and $\mathcal{S}$ denote the set of distributions having their marginals equal to probability distributions from the input set $\mathcal{P}$, i.e., $\mathcal{S} = \{Q : Q^{E_1} = P_1, \ldots, Q^{E_k} = P_k\}$.

I-projection will be the core operation used in both iterative methods discussed in Section 3. Let Q be a probability distribution such that $P_j \ll Q^{E_j}$. I-projection of Q to the set $\mathcal{S}_j$ can be computed [1] as

$$\pi(Q, \mathcal{S}_j)(x) = \begin{cases} 0, & \text{for } Q^{E_j}(x^{E_j}) = 0, \\ Q(x)\dfrac{P_j(x^{E_j})}{Q^{E_j}(x^{E_j})}, & \text{for } Q^{E_j}(x^{E_j}) > 0. \end{cases} \tag{1}$$

In the framework of probabilistic models *factorization* and *decomposition* are used to design computationally efficient representations of probability distributions.

Definition 4 Probability distribution Q factorizes with respect to set $\mathcal{F} = \{F_1, \ldots, F_\ell\}$ if there exist potentials $\psi_{F_i} : \mathbb{X}^{F_i} \mapsto \mathbb{R}, i = 1, 2, \ldots, \ell$ such that for all $x \in \mathbb{X}$

$$Q(x) = \prod_{F_i \in \mathcal{F}} \psi_{F_i}(x^{F_i}) .$$

Definition 5 Set $\mathcal{F}$ of subsets of V is decomposable if either it has exactly one element or it is the union of two disjoint decomposable sets $\mathcal{F}_1$ and $\mathcal{F}_2$ such that there exist $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$ so that:

$$(\cup_{F' \in \mathcal{F}_1} F') \cap (\cup_{F'' \in \mathcal{F}_2} F'') = F_1 \cap F_2 .$$

The rest of the paper will be organized as follows. In Section 2 we will define the I-aggregate - a criteria that measures the distance between the marginals of a joint distribution and the distributions of an input set. In Section 3 we describe IPFP and GEMA, an algorithm that can be regarded a modification of IPFP for inconsistent data. We will conclude the paper by a brief discussion of the properties of the proposed method and by presenting two experiments.

2 Distance to a given input set

Assume that to every probability distribution P_j from input set $\mathcal{P}$ a non-negative weight w_j is assigned such that $\sum_{j=1}^{s} w_j = 1$. The weights can be understood as credibility weights of the corresponding probability distributions. If we have no preferences among the distributions we can set

$w_j = \frac{1}{s}, j = 1, \ldots, s$. Alternatively, the weights can be the ratio of the number of data vectors used to estimate the values of the corresponding probability distribution N_j and total number of data vectors N, i.e., $w_j = \frac{N_j}{N}$. Recall that by Q^{E_j} we denote marginal distribution of a joint probability distribution Q defined on variables $X_i, i \in E_j$, where $E_j \subseteq V$ is the index set of variables of $P_j \in \mathcal{P}$.

Definition 6 Let $\mathcal{P} = \{P_1, \ldots, P_k\}$ be the input set of distributions, Q be a joint probability distribution, and w_j be nonnegative weights summing up to 1. f-aggregate is the functional

$$\sum_{j=1}^{k} w_j \cdot f(P_j, Q^{E_j}) \ .$$

If $f(P_j, Q^{E_j}) = I(P_j \parallel Q^{E_j})$ then f-aggregate will be referred as the I-aggregate. It is obvious that in case of a consistent $\mathcal{P}$ minimization of the I-aggregate over the set of all probability distributions defined on $\mathbb{X}$ ensures that the resulting distribution Q has the required marginals, i.e., $Q \in \mathcal{S}$.

Example 2 Assume input set $\mathcal{P} = \{P_1(X_1, X_2), P_2(X_2, X_3), P_3(X_1, X_3)\}$ from Example 1. Take $\varepsilon = \frac{3}{20}$. In this case there is no distribution having distributions from $\mathcal{P}$ as its marginals. Therefore we search for a distribution minimizing the I-aggregate. We provide this distribution in Table 3. $\diamond$

Table 3: Joint probability distribution $Q(X_1, X_2, X_3)$

	$X_2 = 1$		$X_2 = 2$	
Q	$X_3 = 1$	$X_3 = 2$	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{6}$
$X_1 = 2$	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{6}$

Generally, the I-aggregate need not be minimized by a unique probability distribution. If there exists a probability distribution Q that minimize the I-aggregate then all probability distributions having the same marginals $Q^E, E \in \mathcal{E}$ minimize the corresponding aggregate as well. Furthermore, also distributions having different marginals $Q^E, E \in \mathcal{E}$ can be minimizers of I-aggregate with respect to a given input set $\mathcal{P} = \{P_1, \ldots, P_k\}$. We will denote the set of all probability distribution minimizing the I-aggregate by $\mathcal{T}$. In case of a *consistent input set* $\mathcal{P}$ it is obvious that $\mathcal{T} = \mathcal{S}$.

3 Iterative methods

In case of a consistent input set the knowledge integration task can be solved by means of the *Iterative Proportional Fitting Procedure* (IPFP), also known as the Iterative Proportional Scaling. IPFP dates back to 1940. It was proposed by W. E. Deming and F. F. Stephan [3] as a procedure for adjustment of frequencies in contingency tables.

In each step of the iterative proportional fitting procedure (IPFP) an I-projection of the distribution from previous step to a set $\mathcal{S}_j$ is computed. Projections are repeated until a convergence is reached[1]. A formal definition follows.

Definition 7 Let $Q_{(0)}(X_1, \ldots, X_n)$ be an initial probability distribution such that for all its marginals $Q_{(0)}^{E_j}$ it holds that $P_j \ll Q_{(0)}^{E_j}$ for $j = 1, \ldots, k$. IPFP is following algorithm:

$$\boxed{\begin{array}{l} \text{for } i = 0, 1, 2, \ldots \\ \quad \text{for } j = 1, \ldots, k \\ \qquad Q_{(i \cdot k + j)} \;=\; \pi(Q_{(i \cdot k + j - 1)}, \mathcal{S}_j), \end{array}}$$

Remark If the input set contains k strictly positive distributions and the initial probability distribution is strictly positive then we can see (by using formula 1) that each step of IPFP corresponds to

$$Q_{(i \cdot k + j)}(x) \quad = \quad Q_{(i \cdot k + j - 1)}(x) \cdot \frac{P_j(x^{E_j})}{Q_{(ki + j - 1)}^{E_j}(x^{E_j})} \quad .$$

For the case of consistent discrete probability distributions Csizsár [1] proved that IPFP converges to the I-projection of probability distribution $Q_{(0)}$ to the set of all distributions having the required marginals $\mathcal{S}$, i.e., to $\pi(Q_{(0)}, \mathcal{S})$. This is an important property of IPFP since it means that resulting distribution $Q^\star$ is the nearest distribution (measured by I-divergence) to the original distribution Q_0 from all distributions in $\mathcal{S}$.

In case of an *inconsistent* input set of probability distributions, IPFP tends to come in cycles. Nevertheless, we wish to get some representative probability distribution even in case when the input set is inconsistent. We endeavor to find iterative procedures that converge even for *inconsistent* input sets, the marginals of the resulting distribution should be somehow close to

[1] k consecutive steps $i \cdot k + 1, \ldots, (i + 1) \cdot k$ will be referred as a *cycle*.

the input distributions and independent of the ordering of the input low-dimensional distributions. The application to a *consistent* input sets should lead to the same distribution as IPFP. Assume that the initial probability distribution $Q_{(0)}$ factorizes with respect to $\mathcal{F} = \{F_1, \ldots, F_\ell\}$. We aim at a distribution $Q^\star$ that factorizes with respect to $\mathcal{F}$ and that minimizes the I-aggregate, i.e., $Q^\star \in \mathcal{T}$.

In [7] an algorithm (called Pentis) for solving the inconsistent marginal problem was proposed. It finds a quasi-optimal solution minimizing f-aggregate, where

$$f(P_j, Q^{E_j}) \;=\; \sum_{x^{E_j}} \left| P_j(x^{E_j}) - Q^{E_j}(x^{E_j}) \right| \;.$$

In [5] several modifications of IPFP were proposed. They were further analyzed in [11]. Next we will describe a procedure that we call GEMA. It is a special case of the generalized EM-algorithm [9] applied to our problem.

Each cycle of GEMA has two different phases. During the *first phase* I-projections of the distribution from the previous cycle to sets $\mathcal{S}_j$ are computed. Subsequently, the weighted average of the I-projections is computed. For the second phase we need to compute marginal distributions from the weighted average. Observe, that these distributions are necessarily consistent since they are all marginals of a joint probability distribution. The *second phase* consists of ℓ consecutive steps of IPFP starting with the distribution from the previous cycle of GEMA and consequently fitting the marginal distributions computed in the previous phase. A formal definition follows.

Definition 8 Let $Q_{(0)}$ be an initial joint probability distribution satisfying the same conditions as the initial joint probability distribution in Definition 7. GEMA is following algorithm:

```
for i = 0, 1, 2, ...
    Q̄_(i·ℓ) = Σ_{j=1}^k w_j · π(Q_(i·ℓ), S_j);  //  S_j = {Q : Q^{E_j} = P_j}
    for n = 1, ..., ℓ
        compute the marginal Q̄^{F_n}_(i·ℓ) of Q̄_(i·ℓ);
    for n = 1, ..., ℓ
        Q_(i·ℓ+n) = π(Q_(i·ℓ+n−1), S'_n);  //  S'_n = {Q : Q^{F_n} = Q̄^{F_n}_(i·ℓ)}
```

for $i = 0, 1, 2, \ldots$
$\quad \overline{Q}_{(i\cdot\ell)} \;=\; \sum_{j=1}^{k} w_j \cdot \pi(Q_{(i\cdot\ell)}, \mathcal{S}_j)$; // $\mathcal{S}_j = \{Q : Q^{E_j} = P_j\}$
$\quad$ for $n = 1, \ldots, \ell$
$\quad\quad$ compute the marginal $\overline{Q}^{F_n}_{(i\cdot\ell)}$ of $\overline{Q}_{(i\cdot\ell)}$;
$\quad$ for $n = 1, \ldots, \ell$
$\quad\quad Q_{(i\cdot\ell+n)} \;=\; \pi(Q_{(i\cdot\ell+n-1)}, \mathcal{S}'_n)$; // $\mathcal{S}'_n = \{Q : Q^{F_n} = \overline{Q}^{F_n}_{(i\cdot\ell)}\}$

Following theorem assures that GEMA can be implemented efficiently.

Theorem 1 *Assume that the initial distribution $Q_{(0)}$ factorizes with respect to $\mathcal{F} = \{F_1, \ldots, F_\ell\}$ and satisfies the conditions of Definition 7. If for a decomposable set $\mathcal{C} = \{C_1, \ldots, C_m\}$, $C_i \subseteq V, i = 1, \ldots, m$ it holds that*

- *for each $P_j(X^{E_j}) \in \mathcal{P}$ there exists $C_i \in \mathcal{C}$ such that $E_j \subseteq C_i$ and*
- *for each $F_n \in \mathcal{F}$ there exists $C_i \in \mathcal{C}$ such that $F_n \subseteq C_i$*

then all distributions computed during a finite number of cycles of GEMA factorize with respect to the decomposable set $\mathcal{C}$.

If $Q_{(0)}$ factorizes with respect to $\mathcal{F}$ then all probability distributions $Q_{(i \cdot \ell + n)}$, $n = 1, \ldots, \ell, i < \infty$ factorize with respect to $\mathcal{F}$ as well. Observe that we do not need to have $\overline{Q}_{(i \cdot \ell)}$ represented explicitelly, we only need its marginals $\overline{Q}^{F_n}_{(i \cdot \ell)}, n = 1, \ldots, \ell$. We can compute them, for example, using the junction tree propagation with $C_i \in \mathcal{C}$ being the nodes of the junction tree. Thus the first phase of GEMA can be implemented similarly as suggested by Lauritzen in [8]. Since the second phase of GEMA is equivalent to one cycle of IPFP we can use the space-saving implementation of IPFP proposed by Jiroušek [4]. Consequently, the computational complexity of one cycle of GEMA is proportional to $\sum_{C_i \in \mathcal{C}} |\mathbb{X}^{C_i}|$.

It can be shown [11] that minimization of the I-aggregate corresponds to maximum likelihood estimation from missing data. Therefore the convergence properties of the EM-algorithm studied by Wu in [13] are useful in our setup as well.

We tested the behavior of IPFP and GEMA on *consistent* and *inconsistent* versions of input set $\mathcal{P} = \{P_1, P_2, P_3\}$ defined in Table 1. The plots in Figures 1 and 2 describe development of one probability value (vertical axis) of joint probability distribution $Q_{(i)}$ for one combination of values of random variables. Horizontal axis corresponds to iteration i . We can see that in the *consistent* case both methods converge to the same value. In the *inconsistent* case IPFP oscillates while GEMA converges. For other experiments see [11].

4 Conclusions

The proposed algorithm - GEMA satisfies the original intention to design a method that is an extension of IPFP for inconsistent input set. GEMA can be used to find a distribution that minimizes the I-aggregate. Since for a given input set there can be a set of the minimizers of the I-aggregate we still need an additional characterization of the limit distributions. An important property of GEMA is that it can be implemented efficiently.

Acknowledgments

This paper benefited from personal communication with Radim Jiroušek, Gernot Kleiter, and Fero Matúš.

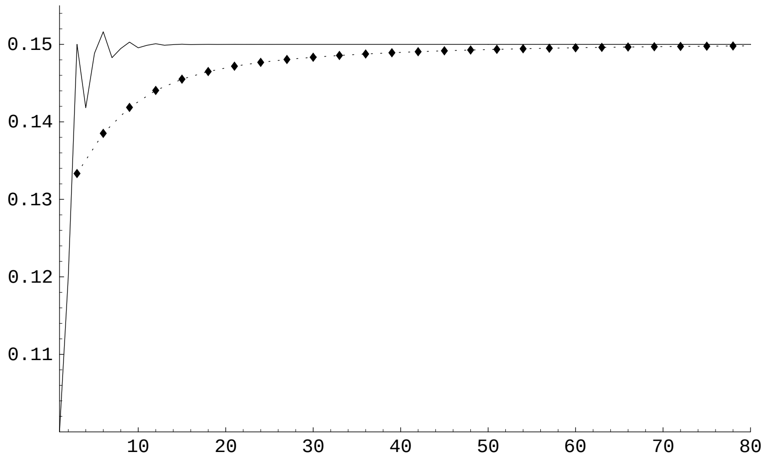

Figure 1: IPFP (full line) and GEMA (dotted line) applied to *consistent input set* ($\varepsilon = \frac{4}{20}$).

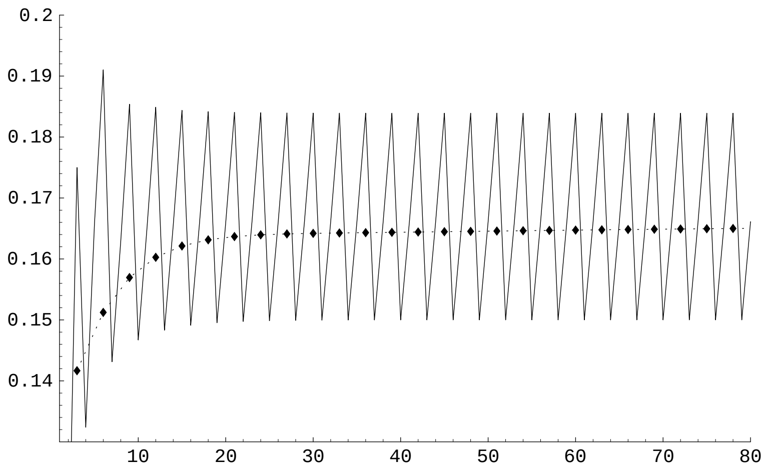

Figure 2: IPFP (full line) and GEMA (dotted line) applied to *inconsistent input set* ($\varepsilon = \frac{3}{20}$).

References

[1] I. Csiszár. *I-divergence geometry of probability distributions and minimization problems.* *Ann. Prob.*, 3(1):146–158, 1975.

[2] R. M. Dawes, D. Faust, and P. E. Meehl. Clinical versus actuarial judgment. *Science*, 243:1668–1674, 1989.

[3] W. E. Deming and F. F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, 11:427–444, 1940.

[4] R. Jiroušek. Solution of the marginal problem and decomposable distributions. *Kybernetika*, 27(5):403–412, 1991.

[5] R. Jiroušek and J. Vomlel. Inconsistent knowledge integration in a probabilistic model. In *Proc. of Workshop "Mathematical Models for handling partial knowledge in A.I."*, New York, 1995. Plenum Publ. Corp.

[6] H. G. Kellerer. Verteilungsfunktionen mit gegeben Marginalverteilungen. *Z. Wahrsch. Verw. Gebiete*, 3:247–270, 1964. In German.

[7] O. Kříž. Inconsistent marginal problem on finite sets. In V. Beneš and J. Štěpán, editors, *Distributions with given marginals and moment problems*, pages 235–242. Kluwer Academic Publishers, 1997.

[8] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19:191–201, 1995.

[9] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* Wiley series in probability and statistics. John Wiley & Sons, Inc., 1997.

[10] D. Osherson, E. Shafir, D. H. Krantz, and E. E. Smith. Probability bootstrapping: Improving prediction by fitting extensional models to knowledgeable but incoherent probability judgements. *Organizational Behavior and Human Decision Processes*, 69(1):1–8, January 1997.

[11] J. Vomlel. *Methods of Probabilistic Knowledge Integration.* PhD thesis, Czech Technical University, Faculty of Electrical Engineeging, 1999.

[12] N. N. Vorobev. Consistent families of measures and their extensions. *Theory of Probability and Applications*, 7:147–163, 1962.

[13] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.