Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Mikal Stapnes

Early Prediction of Cervical Cancer Using Matrix Factorization

Master's thesis in MTFYMA
Supervisor: Markus Grasmair
June 2020

NTNU
Norwegian University of Science and Technology

Mikal Stapnes

Early Prediction of Cervical Cancer Using Matrix Factorization

Master's thesis in MTFYMA
Supervisor: Markus Grasmair
June 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Acknowledgements

I want to thank the following people and organizations who have helped me throughout this journey:

My supervisors, Markus Grasmair and Valeriya Naumova, for their guidance, feedback, and motivation.

Severin Langberg, for our many useful discussions.

SimulaMet and the Cancer Registry of Norway, for allowing me to partake in this exciting project.

My friends and family, for everything.

Abstract

One percent of all Norwegian women will develop cervical cancer by the age of 75. It is the third most common cancer in women of ages 25 to 49. While the Nordic mass-screening programs for cervical cancer have a proven effect in reducing the incidence and mortality of the disease, it remains a challenge to minimize under-treatment and over-screening. In this thesis, we consider the early prediction of cervical cancer as a forecasting problem where the screening history of a single female is encoded as a sparse vector and used as a predictor. As the data is sparse, irregular, and heavily skewed, we employ state-of-the-art matrix factorization techniques along with temporal regularization to deduce robust trends in the data. We validate the method on synthetic data and identify critical challenges associated with our approach. The proposed classifiers are used to predict cervical cancer one year ahead in time in data collected by the Cancer Registry of Norway. Our results show that the classifiers meaningfully discern developing cases of cervical cancer and attain an AUC of 0.78. Finally, we propose future directions of the project.

Sammendrag

En av hundre norske kvinner kommer til å utvikle livmorhalskreft innen en alder av 75. Kreften er den tredje mest hyppige blant kvinner i alderen 25 til 49. De nordiske screeningprogramme for livmorhalskreft har en påvist effekt for å redusere forekomst og dødelighet av sykdommen. Likevel forblir det en utfordring å minimere underbehandling og overscreening. I denne oppgaven anser vi prediksjon av livmorhalskreft som et prognoseproblem. Vi oversetter en kvinnes screeninghistorikk til en glissen vektor og bruker denne som regressor for å predikere kvinnens fremtidige utvikling. Det resulterende datagrunnlaget er glissent, ujevnt observert og skjevfordelt. For å håndtere disse problemene implementerer vi toppmoderne metoder innen matrisefaktorisering. Vi validerer metoden på syntetiske data og kommenterer kritiske utfordringer knyttet til vår tilnærming. Vi bruker prediksjonsmodellene til å forutsi livmorhalskreft i data fra Kreftregisteret. Våre resultater viser at modellene identifiserer kvinner med økt risiko for sykdommen og oppnår en AUC på 0.78. Til slutt foreslår vi potensielle videre utviklinger av prosjektet.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cervical Cancer Screening

Cervical cancer ranks third among the most frequent types of cancer for Norwegian women of ages 25 to 49. It is estimated that 1.0% of Norwegian women will develop cervical cancer by the age of 75 [1]. To reduce the incidence and mortality of cervical cancer, the Norwegian Cervical Cancer Screening Programme (NCCSP) was launched in the late 1960s and is currently in effect. In recent years the NCCSP and its sibling programs in Denmark, Sweden, and Finland have been identified as key contributors to the low incidence and mortality of cervical cancer in the Nordic countries. It is estimated that in the absence of these programs, the prevalence of cervical cancer would be doubled [2].

Most cervical cancer screening programs, the NCCSP included, are characterized by population-wide regular screening. The recommended screening interval length, i.e., the time between two screenings, is typically homogeneous for all females. Considerable research has been devoted to investigating the reduction of incidence and mortality offered by a shortened interval. To preempt developing cases of cervical cancer, it should be short but not excessively so as this leads to over-screening and increased expenditure. Studies [3, 4] have found the triennial (every three years) screening interval to be a cost-effective compromise offering a high degree of protection.

The idea that all women should follow the same interval length is currently being challenged. It is proposed that a data-driven and personalized screening program can offer similar or increased protection at reduced over-screening and expenditure. The personalization should adapt the interval length to the female's probability of developing cervical cancer. The idea has gained increased traction as significant variation has been demonstrated in the protection offered by a fixed interval length across age-groups [5, 4, 6]. While a given interval length may offer sufficient protection for women between the ages of 40 to 69, the same length may be insufficient for women between the ages of 20 to 39. By using age and previous screening history as risk factors, it may be possible to

separate the Norwegian population into low- and high-risk groups.

1.2 The DeCipher Project

The maturity of the Nordic screening programs for cervical cancer creates an ideal environment for exploratory work in the development of a personalized screening program. Specifically, the wealth and quality of the data amassed by the NCCSP enable the use of data-driven methods. The development and testing of such methods is the goal of the ongoing DeCipher project. The project is funded by the Research Council of Norway and conducted in collaboration between SimulaMet, the Cancer Registry of Norway, Lawrence Livermore National Laboratory, and Karolinska Institutet [7].

> "We aim at developing a data-driven framework to provide a personalised time-dependent risk assessment of disease initiation and identify subgroups of individuals and possible biomarkers, which can lead to similar disease progression. The DeCipher results will allow for improvement of individuals' preventive cancer healthcare while reducing the cost of screening programs." [7]

As part of the DeCipher project, an introductory study was conducted to investigate if the PAtient reCord densiFIER (PACIFIER) algorithm [8] could be used as a tool in the early prediction of cervical cancer. While PACIFIER was designed for the domain of coronary heart disease and end-stage renal disease, the introductory study showed an indication that the algorithm is applicable also for cervical cancer [9]. In a direct continuation of this introductory study, we extend the ideas of PACIFIER into a full-fledged prediction algorithm for cervical cancer.

1.3 Encoding Screening Data

Central in this thesis is the use of a Norwegian population-based dataset comprised of 500.000 screenings gathered from 80.000 women. The data is collected by the Cancer Registry of Norway in the period 1992 to 2015 as part of the NCCSP and contains histological exams, cytological exams, and cervical cancer exams. All exams consist of a string representing the type of test and a resulting high-level diagnosis.

These strings, while informative to a medical practitioner, are intractable from a data science point of view. Histology and cytology are different means with the same intent: to check for the development of cervical cancer or precancerous changes to the cells of the cervix. The diagnoses of a histological exam and a cytological exam may differ yet still indicate the same status of the female. In our project, we are interested only in the status itself and not whether it was assessed through histology or cytology. By mapping the diagnoses of the different exams to a common set of integers, we derive a dataset

Table 1.1: Overview of the mapping from an exam diagnosis to a cervical cancer state. We map the diagnoses of the three exam types, histology (HIST), cytology (CYT) and cancer exam (CAN), to the cervical cancer state. Diagnoses without an encoding are removed from the dataset.

Exam type	Diagnosis	Grade	State
CYT	Normal	Normal	1
CYT	ASC-US	Low	2
CYT	LSIL	Low	2
CYT	ASC-H	High	3
CYT	AGUS/ACIS	High	3
CYT	HSIL	High	3
CYT	Cancer	Cancer	4
CYT	Metastasis	-	-
CYT	Unsatisfactory	-	-
HIST	NILM	Normal	1
HIST	CIN1	Low	2
HIST	CIN2	High	3
HIST	CIN3	High	3
HIST	ACIS	High	3
HIST	Unknown morphology	-	-
HIST	Unsatisfactory	-	-
CAN	Squamous cell carcinoma	Cancer	4
CAN	Adenocarcinoma	Cancer	4
CAN	Other cancers	Cancer	4

that is significantly easier to handle using a data-driven model. Simultaneously the simplified dataset remains representative of the prospect that the female will develop cancer in the upcoming future. The mapping was developed in collaboration with medical experts at the Cancer Registry of Norway and can be seen in its entirety in Table 1.1. The mapped variable is denoted the female's *cervical cancer state* or simply *state*.

Definition 1.1. The *cervical cancer state* / *state* $s_t \in \mathcal{S} = \{1, 2, 3, 4\}$ of a female is a condensed description of the female's overall health in relation to cervical cancer. The index t represents the female's age. The state $s_t = 1$ denotes normal, $s_t = 2$ denotes a low-grade state, $s_t = 3$ denotes a high-grade state and $s_t = 4$ denotes onset cancer.

We let a screening result represent a three-month interval such that the entire screening history of a female can be encoded as a vector. If the female had a single screening in the three months corresponding to a particular entry, we set it to be the mapped state of the screening. If the female had several, we set the entry to be the most severe. Conversely, if the female had none, the entry is set to zero and interpreted as missing. By using the same procedure for all females, the raw dataset is converted to an $N \times T$ matrix where N is the number of females, and T is the largest age difference between any two females measured in periods of three months.

It is challenging to deduce the disease development of a female from very few screenings. At the same time, our methods scale with the number of rows in the matrix. To test our algorithm in an optimistic environment and reduce the computational cost of running it, we restrict our attention to females for whom we have six or more screenings. In other words, we remove females with five or fewer screenings. The screening histories of the remaining 38001 Norwegian women are stored in a 38001×321 matrix referred hereafter to as "the Screening Dataset". Of the entries in the Screening Dataset, 97.12% are zero and interpreted as missing. Of the nonzero entries, 92.96% of the entries contain the normal state, 4.66% contain the low-grade state, 2.34% contain the high-grade state, and 0.04% contain the cancer state. We remark that the Screening Dataset is extremely sparse and highly imbalanced over the states. Also, the nonzero entries are mostly located at younger and middle ages.

1.4 Matrix Factorization

It is the extreme sparsity along the time-dimension that complicates the use of data-driven methods. Classical methods in time series analysis involve the extraction of derived features from the history of the object of prediction. In cervical cancer screening, this history consists of only a small number of unevenly spaced screenings. The sparsity is of such a degree that classical methods fail to extract most, if not all, the features. While this can be handled through the development of specialized feature-based methods, we instead use it as motivation to borrow inspiration from another rapidly emerging field.

Data sparsity arises in the commercial context of recommending products to customers based on their spending history. While the spending history is typically extremely sparse and irregular, the customer's preferences can be reconstructed by the association with other customers of similar spending patterns. Collaborative filtering is a class of methods developed to automatically predict the interests of a user by collecting preferences or taste information from many users. The commercial nature of the problem has inspired considerable interest.

Zhou et al. [8] recognized that collaborative filtering is not limited to commerce and that the concept of product preference is similar to that of phenotype in medicine. The correct treatment of a female can be inferred by the association with other females of similar disease progression. With modifications, the methods of collaborative filtering could also be applied in a medical context. Of particular interest is the class of methods named *matrix factorization* (MF) methods. These methods implement the assumption that the phenotypes of all females can be described as a weighting of a small number of phenotypic archetypes. The previously mentioned PACIFIER algorithm, which is an MF method, was applied in the early prediction of coronary heart disease and end-stage renal disease and found to outperform classical methods.

1.5 Thesis Overview

The purpose of this thesis is the development and testing of a prediction model based on matrix factorization for cervical cancer. We initiate our approach by defining the theoretical background for applying matrix factorization to cervical cancer screening data in Section 2. We demonstrate that the MF formulation can be derived as a relaxation of *matrix completion*. Interestingly, an equivalent model can be formulated in a probabilistic context. In Section 2.2 we reveal that by modelling the likelihood between the female's state and an underlying continuous-valued *latent risk*, the MF problem can be obtained as a result of maximum a posteriori (MAP) estimation.

Definition 1.2. The *latent risk* or simply *risk* $m_t \in \mathbb{R}$ of a female at time t determines the distribution of the female's cervical cancer state. The *latent risk profile* $\boldsymbol{m} \in \mathbb{R}^T$ describes the latent risk of a female at a uniformly spaced grid of time.

We make the fundamental assumption that the latent risk profiles of all women develop smoothly in time. To account for this, we in Section 2.3 extend the probabilistic model to enforce smooth temporal trends. For the first time, we introduce the smooth probabilistic matrix factorization (SPMF) and convolutional probabilistic matrix factorization (CPMF) models. We proceed by describing how the MAP estimate of the latent risk matrix can be found using the LMaFit [10] algorithm, which is an alternating minimization scheme, in Section 2.4.

To predict future cervical cancer state, we must further equip the models with some mechanism of classification. Section 2.5 describes how the screening

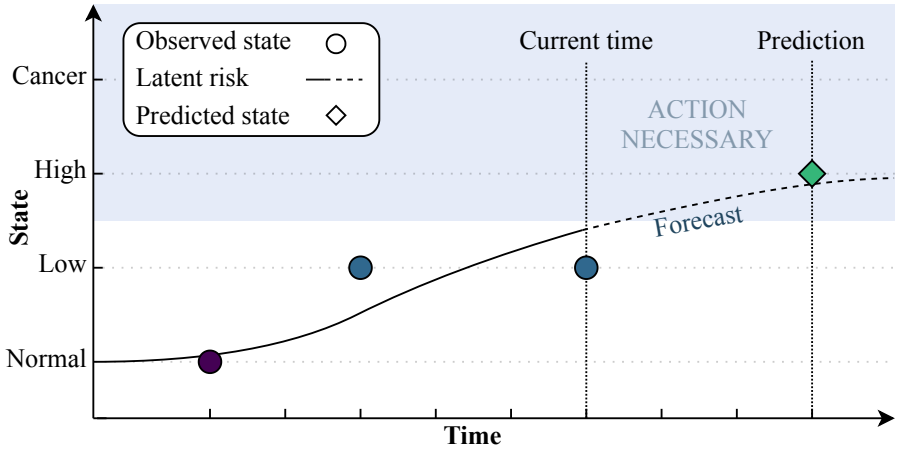


Figure 1.1: Illustration of the prediction scheme. The model associates the screening history of a female to specific latent risk profiles. In turn, these associations are used to forecast the future development of the female's latent risk and cervical cancer state. If a high-grade or cancer state is forecast, the female can be recommended to follow a shortened screening interval or some other action.

history of a female can be associated with the latent risk of a training population and how this association, in turn, can be used to forecast the future development of the female's latent risk and state. Figure 1.1 illustrates this process. Also, we reflect upon the relative severity of the cervical cancer states. Founded in our reflections, we argue that the problem can be reduced to a dichotomy where the females are either considered as sick or healthy concerning cervical cancer. We describe this new context and introduce the binary SPMF (B-SPMF) and binary CPMF (B-CPMF) classifiers.

In Section 3, we describe a method for generating synthetic data replicating the sparsity and irregularity of the Screening Dataset. We highlight essential characteristics of the Screening Dataset by comparison to the synthetic data. Moreover, we investigate the inner-workings of the training procedure of our classifiers and inspect whether the LMaFit method successfully reconstructs the latent risk matrix of synthetic data in Section 4.

Finally, we use the SPMF and CPMF classifiers to predict future cervical cancer state one year ahead in time. Section 5 describes our results using the classifiers on simulated data and the Screening Dataset. We further switch to the binary context and implement a model bias in the B-SPMF and B-CPMF classifiers to preempt developing cases of cervical cancer. Section 6 describes our results on simulated data and the Screening Dataset. Lastly, we summarize our findings and suggest directions for future development in Section 7.

The results of the SPMF and CPMF classifiers for the early prediction of cervical cancer, along with those of an approach based on geometric deep learning, have been submitted to an international conference [11].

Chapter 2

Matrix Factorization

The matrix factorization model can be derived using its roots in matrix completion. Matrix completion (MC) is the recovery of an underlying matrix M from only a subset Ω of its entries, denoted the matrix' *observation mask*. In this project, we consider real-valued and longitudinal matrices of dimensions N and T, $M \in \mathbb{R}^{N \times T}$. The observation mask is then a subset of all row-time combinations, $\Omega \subset \{0, 1, \ldots, N-1\} \times \{0, 1, \ldots, T-1\}$. The matrix is *observed* at indices $(i, j) \in \Omega$ and *missing* otherwise. In practice, any missing entry must be implemented as some numerical encoding. To correspond with previous literature on matrix completion this encoding is chosen as zero. A matrix representation of the observation mask is defined as

$$\text{mat}(\Omega) = \begin{cases} 1 & (i,t) \in \Omega \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

This allows the definition of a projection of any $N \times T$ matrix onto the observation mask as

$$\mathcal{P}_\Omega(X) = \text{mat}(\Omega) \circ X = \begin{cases} X_{it} & (i,t) \in \Omega \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where $\circ$ is the Hadamard product. The *density* of the observation mask is defined as

$$|\Omega| = \frac{1}{NT} \sum_{i=1}^{N} [\text{mat}(\Omega)]_{ij}. \tag{2.3}$$

2.1 Matrix Completion

To introduce the field of matrix completion we first consider the *observed matrix* Y to be a direct partial observation of the underlying matrix,

$$Y = \mathcal{P}_\Omega(M) \tag{2.4}$$

such that its entries are real-valued, $Y_{ij} \in \mathbb{R}$ for $(i,j) \in \Omega$. In this case the by far most common approach to matrix completion is a mathematical application of Occam's razor; we seek the lowest rank matrix that is in agreement with the observed data

$$\begin{aligned} \underset{M}{\arg\min} \quad & \operatorname{rank}(M) \\ \text{subject to} \quad & \mathcal{P}_\Omega(Y) = \mathcal{P}_\Omega(M). \end{aligned} \tag{2.5}$$

In the early advancements of matrix completion, this formulation was found to be NP-hard and therefore of little practical use. For the general case, all currently known algorithms that solve (2.5) require exponential time in the dimensions N and T in both theory and practice [12]. The pessimism induced by this result lasted until the mid-2000s, at which point it was shown that much simpler formulations were in many cases also capable of recovery if the underlying matrix is of low rank [13, 14]. In a landmark paper, Candes and Recht developed a nuclear norm minimization (NNM) formulation (2.6) that took into account that the observed entries may be perturbed by some noise and showed that if the noise level is small the error can be expected to be similarly small [15]. We restate the NNM formulation as

$$\begin{aligned} \underset{M}{\arg\min} \quad & \|M\|_* \\ \text{subject to} \quad & \|\mathcal{P}_\Omega(Y - M)\|_F \leq \rho, \end{aligned} \tag{2.6}$$

where the nuclear norm $\|X\|_* = \sum_k \sigma_k$ is the sum of the singular values. The Frobenious norm is defined as

$$\|X\|_F = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} X_{ij}^2}. \tag{2.7}$$

and the parameter $\rho \in \mathbb{R}$ determines the level of noise in the observations. The constrained formulation (2.6) is for some value $\alpha \in \mathbb{R}$ equivalent to the Lagrangian formulation

$$\underset{M}{\arg\min} \frac{1}{2} \|\mathcal{P}_\Omega(Y - M)\|_F^2 + \alpha \|M\|_* \tag{2.8}$$

The choice of the nuclear norm was not random; The NNM problem (2.6) is the tightest convex relaxation of the rank minimization problem [15]. The solution to (2.8) can be found using e.g. the Fixed Point Continuation (FPC) algorithm of [16]. Even though solvers of the NNM problem (2.8) are vastly more efficient than solvers of the rank minimization problem, they are still unfit for truly large-scale problems. The FPC algorithm and most other solvers of (2.6) bear the cost of repeated computation of the full or partial SVD of a $N \times T$ matrix [10]. To avoid this, the authors of [10, 17, 18] impose the additional assumption that M is of rank at most R and thus allows the rank decomposition $M = UV^T$ where $U \in \mathbb{R}^{N \times R}$ and $V \in \mathbb{R}^{T \times R}$. Inserted into (2.8) this yields another NNM problem

$$\underset{U,V}{\arg\min} \frac{1}{2} \|\mathcal{P}_\Omega(Y - UV^T)\|_F^2 + \alpha \|UV^T\|_*. \tag{2.9}$$

The purpose of assuming the decomposition $M = UV^T$ is that we can now use Lemma 1 of [13].

Lemma 1 [13]. *For any matrix X the following are all equal:*

1. *The minimum* $\min_{U,V} \|U\|_F\|V\|_F$, *subject to* $X = UV^T$
2. *The minimum* $\min_{U,V} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$, *subject to* $X = UV^T$
3. *The nuclear norm* $\|X\|_*$,

The lemma states that for the NNM problem (2.9) there exists the neighboring Matrix Factorization (MF) problem

$$\arg\min_{U,V} \|\mathcal{P}_\Omega(Y - UV^T)\|_F^2 + \alpha\left(\|U\|_F^2 + \|V\|_F^2\right). \tag{2.10}$$

The introduction of the rank decomposition in (2.9) creates a problem that is non-convex. As a result of this, numerical solvers of (2.10) may become stuck in local minima and a numerical solution should be allowed only if found to be reasonable. At the same time, (2.10) can be quickly solved even if N or T are large. Notice that letting either variable U or V be fixed, minimizing the objective

$$F(U,V) = \|\mathcal{P}_\Omega(Y - UV^T)\|_F^2 + \alpha\left(\|U\|_F^2 + \|V\|_F^2\right) \tag{2.11}$$

as a function of the remaining variable is a standard linear least-squares problem. The MF problem (2.10) can be efficiently solved by the alternating minimization procedure in which U, V are iteratively updated by

$$U^{(l+1)} = \arg\min_U F(U, V^{(l)})$$

and

$$V^{(l+1)} = \arg\min_V F(U^{(l+1)}, V).$$

Such a procedure monotonically decreases the objective (2.11), which is also bounded below by 0. As a consequence, the value of the objective is guaranteed to convergence. Note that providing a reasonable a priori upper bound R of the rank of M is in many practical applications challenging.

The use of an MF formulation in the recovery of large yet low-rank matrices is already a merited approach. The Incremented-Rank PowerFactorization (IRPF) algorithm, the Low-rank Matrix Fitting (LMaFit) and the Scaled Alternating Steepest Descent (ScaledASD) algorithms, all of which solve a formulation similar to (2.10), were found to produce similar or more accurate recovery than solvers of NNM formulations at significantly reduced running time [10, 17, 18]. In particular, The Low-rank Matrix Fitting (LMaFit) algorithm was found to outperform the SVD-based APGL of [19] and the Fixed Point Continuation with

Approximate SVD (FPCA) of [16] at a fraction of the CPU time for real and synthetic data. Recovery was similar to that of the APGL and FPCA algorithms, even in the case that R was overestimated to be $K = \text{floor}[1.50 \cdot \text{rank}(M)]$ [10].

The success and speed of these algorithms, LMaFit in particular, has motivated the use of an MF formulation in the medical domain. In a 2014 paper, Zhou et al. implemented an extension of the LMaFit algorithm for the densification of electronic medical records (EMRs). The feature vectors extracted from the densified EMRs were used in the early prediction of congestive heart failure and end-stage renal disease. When used as input to a secondary prediction model, the feature vectors extracted from the densified EMRs outperformed the feature vectors derived using a range of classical imputation methods [8]. In this medical setting, such a decomposition has the added benefit of implementing the intuitive idea that all females' disposition for illness can be expressed as a linear combination of at most R basic profiles (phenotypic archetypes). In other words, the MF approach can be used not only to gain additional insight into the individual females but also to discover patterns shared by the entire population. Building on the success of Zhou et al., we will implement a very similar model in the early prediction of cervical cancer.

2.2 Introducing a Probabilistic View

To introduce the classical derivation of the MF method, we had to assume that Y was real-valued. The Screening Dataset, on the other hand, contains either of the states

$$Y_{it} \in \mathcal{S} = \{1, 2, 3, 4\}, \quad (i, t) \in \Omega. \tag{2.12}$$

In the following section, we argue that the LMaFit algorithm can be applied to an observed matrix containing the integers one through four under certain assumptions on the latent structure of the data. To achieve this, we lend inspiration from the probabilistic matrix factorization (PMF) approaches of [20, 21]. In these, a probabilistic relation is assumed between the underlying matrix M and the observed matrix Y such that the unknown M can be recovered using maximum a posteriori (MAP) or Markov Chain Monte Carlo estimation techniques.

We imagine that the observed state matrix Y is a partially observed version of the *complete state matrix* S,

$$Y = \mathcal{P}_\Omega(S), \tag{2.13}$$

and that there exists an entrywise probabilistic relation between S and the latent risk matrix M.

(A1) We assume that the probability of female i being in a cervical cancer state at time t, S_{it}, is determined by the *sampled truncated Gaussian distribution*

$$p(S_{it} \mid M_{it}, \theta) \propto \exp[-\theta(S_{it} - M_{it})^2], \quad S_{it} \in \mathcal{S} \tag{2.14}$$

where θ is a distribution parameter.

(A2) We assume that the probability that female i attends a screening at time t is independent of previous screenings.

$$p(\Omega_{it} \mid \Omega_{i,t-1}, \ldots, \Omega_{i,1}, S_{i,t-1}, \ldots S_{i,1}) = p(\Omega_{it}) = q \tag{2.15}$$

where $q \in [0, 1]$.

The assumptions listed above are not expected to hold true in the Screening Dataset. Most importantly, we expect the observation of a cancer, high-grade or even low-grade state to alter the future screening attendance of the female. Even so, we retain these assumptions to produce an optimization problem that we can solve with the resources available. The likelihood in (A1) yields the Frobenius norm in the discrepancy term. As a result, the subproblems in the alternating minimization schemes remain LLS problems. By assuming (A2), the mask enters the posterior probability only as a constant factor. If, on the other hand, we had to account for the conditional dependence of screening attendance, the complexity of MAP estimation would be much greater.

Finally, we assume that the latent risk matrix M is of low rank and can be represented by the decomposition

$$M = UV^T. \tag{2.16}$$

The relation between the state matrix, the latent risk matrix and the decomposition U, V is illustrated in Figure 2.1. For cervical cancer, the time-component V can be interpreted as R basic disease trajectories. The coefficient matrix can be interpreted as a female-specific weighting of the trajectories in V.

Under (A1) the subsequent states of a female depend only on her latent risk profile and the observed mask. We compute the likelihood of the observed profile of a single female as

$$\begin{aligned} p(Y_i \mid U_i, V, \Omega_i, \theta) &= \prod_{t=1}^{T} [p(Y_{it} \mid [U_i V^T]_t, \theta)]^{\Omega_{it}} \\ &= \prod_{t \in \Omega_i} \exp\left[-\theta(Y_{it} - [U_i V^T]_t)^2\right], \end{aligned} \tag{2.17}$$

and similarly for a population of N independent females

$$\begin{aligned} p(Y \mid U, V, \Omega, \theta) &= \prod_{i=1}^{N} \prod_{t=1}^{T} [p(Y_{it} \mid [U_i V^T]_t)]^{\Omega_{it}} \\ &= \prod_{(i,t) \in \Omega} \exp\left[-\theta(Y_{it} - [UV^T]_{it})^2\right]. \end{aligned} \tag{2.18}$$

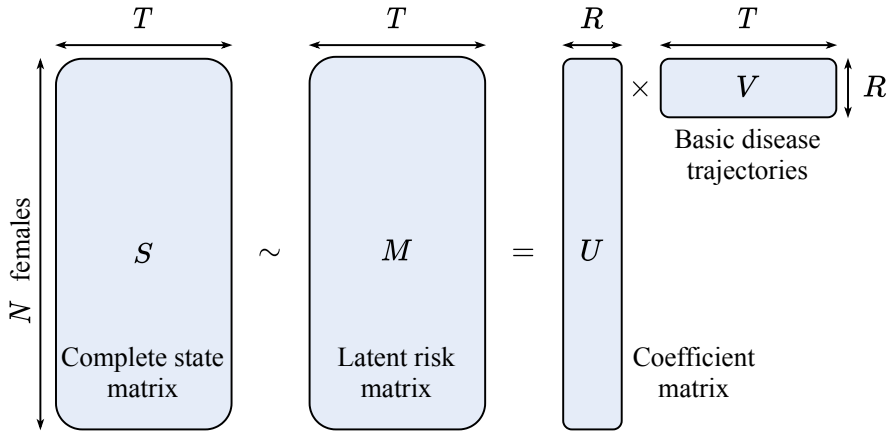


Figure 2.1: Illustration of the assumed relation between the state matrix, the latent risk matrix, the basic disease trajectories and the female coefficient matrix.

Using the deduced likelihood we derive the posterior probability of U, V conditioned on the observed matrix as

$$p(U, V \mid Y, \Omega, \theta) \propto p(Y \mid U, V, \Omega, \theta)\pi(U)\pi(V) \tag{2.19}$$

where $\pi(U)$ and $\pi(V)$ are the priors of U and V. To produce a formulation equivalent to (2.10), we first assume that the entries of U and V follow a Gaussian prior $U_{ij}, V_{ij} \sim \mathcal{N}(0, \sigma^2)$. Inserting these yields the posterior probability

$$\begin{aligned} p(U, V \mid Y, \Omega, \theta) \propto & \prod_{(i,t)\in\Omega} \exp\left[-\theta(Y_{it} - [UV^T]_{it})^2\right] \\ & \cdot \prod_{i=1}^{N}\prod_{k=1}^{K} \exp\left[-\frac{U_{ik}^2}{\sigma^2}\right] \prod_{t=1}^{T}\prod_{k=1}^{K} \exp\left[-\frac{V_{tk}^2}{\sigma^2}\right]. \end{aligned} \tag{2.20}$$

By standard means of MAP estimation we consider instead the logposterior

$$\begin{aligned} \ln p(U, V \mid Y, \Omega, \theta) \propto & \frac{\overline{\Omega}}{2} \ln\theta - \sum_{(i,t)\in\Omega} \theta(Y_{it} - [UV^T]_{it})^2 \\ & - (N\,R) \ln\sigma_U - \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{U_{ik}^2}{\sigma^2} \\ & - (T\,R) \ln\sigma_V - \sum_{t=1}^{T}\sum_{k=1}^{K} \frac{V_{tk}^2}{\sigma^2}, \end{aligned} \tag{2.21}$$

which we recognize in its simplified form

$$\ln p(U, Y \mid Y, \Omega, \theta) \propto -\theta\|\mathcal{P}_\Omega(Y - UV^T)\|_F^2 - \frac{1}{\sigma^2}\left(\|U\|_F^2 + \|V\|_F^2\right). \tag{2.22}$$

The MAP estimate of the decomposition is found by maximizing (2.22) with respect to U and V. We recognize this to be a matrix factorization problem equivalent to (2.10) with slightly modified regularization parameters. We have shown that, under certain assumptions, matrix factorization can safely be applied also to cervical cancer screening data.

2.3 Temporal Regularization

In the context of cervical cancer, we have prior belief that the R basic disease trajectories are temporally smooth. Large and sudden jumps in the latent risk of a female are considered unlikely and entries in the recovered M should be in some proximity of their neighbors along the time dimension. To induce such trends also in our latent risk matrix estimate we modify the assumed prior on $\pi(V)$.

2.3.1 The SPMF Model

In the first model, referred to as *Smooth Probabilistic Matrix Factorization* (SPMF), we let the logprior

$$\ln \pi(U) + \ln \pi(V) \propto \lambda_1 \left(\|V\|_F^2 + \|U\|_F^2 \right) + \lambda_2 \|DV\|_F^2 \tag{2.23}$$

be proportional not only to the Frobenius norms of U and V but also to the Frobenius norm of the finite-difference approximation DV where

$$D = \begin{bmatrix} -1 & 1 & & & & \\ & -1 & 1 & & (0) & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & (0) & & & -1 & 1 \\ & & & & & 0 \end{bmatrix}. \tag{2.24}$$

This model will induce a constant level of smoothness along the time dimension of the basic profiles. Consider a scenario in which a net change along the time dimension in the basic trajectories is required to allow for the discrepancy term to be reduced. As the finite differences in the third term of (2.23) are squared directly, the model will spread this net change in V evenly along the time dimension to reduce the size of the logprior. Under this new prior the MAP estimate is found as the solution to

$$\underset{U,V}{\arg\min} \|\mathcal{P}_\Omega(Y - UV^T)\|_F^2 + \lambda_1 \left(\|U\|_F^2 + \|V\|_F^2 \right) + \lambda_2 \|DV\|_F^2. \tag{2.25}$$

2.3.2 The CPMF Model

In some cases, we want to center the net change in the basic profiles around certain time points of interest. The second model, referred to as *Convolutional*

Probabilistic Matrix Factorization (CPMF), implements this by including the linear mapping K in the logprior

$$\ln \pi(U) + \ln \pi(V) \propto \lambda_1 \left(\|V\|_F^2 + \|U\|_F^2 \right) + \lambda_2 \|KDV\|_F^2, \tag{2.26}$$

where K is defined by

$$K_{ij} = \exp(-|i-j|). \tag{2.27}$$

In this latter case, the finite differences are squared after being convolved along the time dimension. As a result, the finite-difference of a large and sudden jump in a basic profile will only be squared after being distributed along the time dimension. The convolution reduces the effect of the jump in the logprior. For that reason, the CPMF model will lend itself to a small number of sudden jumps in the basic profiles if this reduces the discrepancy term. The MAP estimates of U, V in the CPMF model are found as the solution to

$$\arg\min_{U,V} \|\mathcal{P}_\Omega(Y - UV^T)\|_F^2 + \lambda_1 \left(\|U\|_F^2 + \|V\|_F^2 \right) + \lambda_2 \|KDV\|_F^2. \tag{2.28}$$

Notice that (2.25) is the same as (2.28) if K is the identity mapping. An algorithm solving (2.28) for an arbitrary K can be used to implement both the SPMF and CPMF model. We refer to the process of estimating the latent risk matrix using either of the models as *training* the model.

2.4 Training the SPMF / CPMF Models

In this section, we describe in detail the solution of (2.28) using the IRPF [17] and LMaFit [10] algorithm and argue that the latter is computationally tractable. Both of these rely on an alternating minimization scheme in which U and V are updated iteratively.

2.4.1 The IRPF Algorithm

Implementing the IRPF algorithm for solving (2.28) we define the *objective*

$$F(U,V) = \|\mathcal{P}_\Omega(Y - UV^T)\|_F^2 + \lambda_1 \left(\|U\|_F^2 + \|V\|_F^2 \right) + \lambda_2 \|KDV\|_F^2, \tag{2.29}$$

such that U, V can be iteratively updated as the solutions to the subproblems

$$U^{(l+1)} = \arg\min_U F(U, V^{(l)}) \tag{2.30}$$

$$V^{(l+1)} = \arg\min_V F(U^{(l+1)}, V). \tag{2.31}$$

Notice that (2.30) is a linear least squares (LLS) problem with an optimality criterion

$$[\mathrm{mat}(\Omega) \circ \left(UV^{(l)T}\right)]V^{(l)} + \mu_1 U = [\mathrm{mat}(\Omega) \circ Y]\, V^{(l)T} \tag{2.32}$$

that is decoupled over the rows of U. This means that the solution of (2.32) can be computed row-wise by

$$U_{i,:}^{(l+1)} = V^{(l)T}\text{diag}(\text{mat}(\Omega)_{i,:})Y_{i,:}\left[V^{(l)T}\text{diag}(\text{mat}(\Omega)_{i,:})V^{(l)} + \lambda_1 I_R\right]^{-1}. \tag{2.33}$$

Similarly (2.31) is a LLS problem with optimality criterion

$$\begin{aligned}[\text{mat}(\Omega) \circ \left(VU^{(l+1)T}\right)]U^{(l+1)} + \lambda_1 V + \lambda_2 D^T K^T K D V \\ = \left[\text{mat}(\Omega) \circ Y\right] U^{(l+1)},\end{aligned} \tag{2.34}$$

which we recognize to be a Sylvester equation of the form $AX + XB = C$. The naive approach to solving a Sylvester equation is through vectorization. We define $\text{vec}(\cdot)$ to be the flattening of a matrix to vector form by stacking the columns successively and $\text{unvec}(\cdot)$ as its inverse operation. By defining the matrix

$$H_V = (U^{(l)T} \otimes I_T)\text{diag}(\text{vec}(\Omega))(U^{(l+1)} \otimes I_T) + I_R \otimes (\lambda_1 I_T + \lambda_2 D^T K^T K D) \tag{2.35}$$

the optimality criterion (2.34) can be rewrittten in vectorized form

$$H_V \,\text{vec}(V) \text{vec}\left[\mathcal{P}_\Omega(Y)^T U^{(l)}\right]. \tag{2.36}$$

This equation is explicitly solved for $V^{(l+1)}$ by

$$V^{(l+1)} = \text{unvec}\left(H_V^{-1} \text{vec}\left[(\text{mat}(\Omega)^T \circ Y^T)U^{(l)}\right]\right). \tag{2.37}$$

Unfortunately, the solution of the latter subproblem becomes problematic when applying the IRPF algorithm to the Screening Dataset. For the number of females, $N = 38001$ and the size of the time-dimension $T = 321$, the matrices involved in the computation of H_V are too large to fit in RAM for most modern-day computers. Even if it were not so, the solution of a $(RT) \times (RT)$ system at every iteration incurs a sizeable cost to the running time of the algorithm. While both problems can be mitigated by using a specialized solver, we instead use it as motivation for implementing the LMaFit algorithm.

2.4.2 The LMaFit Algorithm

As opposed to IRPF, the Low-rank Matrix Fitting (LMaFit) algorithm can easily be implemented without vectorization and is therefore more tractable in cases where N is large and the complexity of matrix-matrix operations dominate. This is achieved by the additional relaxation from (2.28) to the constrained problem

$$\begin{aligned}\underset{U,V,\Gamma}{\arg\min} \quad & \|\Gamma - UV^T\|_F^2 + \lambda_1\left(\|U\|_F^2 + \|V\|_F^2\right) + \lambda_2\|KDV\|_F^2 \\ \text{subject to} \quad & \mathcal{P}_\Omega(Y) = \mathcal{P}_\Omega(\Gamma).\end{aligned} \tag{2.38}$$

Observe that the projection onto the mask is applied only in the constraint. After relaxation the alternating minimization scheme now involves the solution to the three subproblems

$$\arg\min_{U} \|\Gamma^{(k)} - UV^{(k)T}\|_F^2 + \lambda_1\|U\|_F^2 \tag{2.39a}$$

$$\arg\min_{V} \|\Gamma^{(k)} - U^{(k+1)}V^T\|_F^2 + \lambda_1\|V\|_F^2 + \lambda_2\|KDV\|_F^2 \tag{2.39b}$$

$$\begin{aligned} \arg\min_{\Gamma} \quad & \|\Gamma - U^{(k+1)}V^{(k+1)T}\|_F^2 \\ \text{subject to} \quad & \mathcal{P}_\Omega(Y) = \mathcal{P}_\Omega(\Gamma). \end{aligned} \tag{2.39c}$$

While these may seem no simpler to solve than before, we can now leverage the similarity of these subproblems to those in PACIFIER [8]. In similar fashion as Zhou et al. we compute the eigenvalue decompositions

$$Q_1\Lambda^{(1)}Q_1^T = U^{(k+1)T}U^{(k+1)} + \lambda_1 I \tag{2.40a}$$

$$Q_2\Lambda^{(2)}Q_2^T = \lambda_2 D^T K^T K D, \tag{2.40b}$$

and assign

$$\Xi = Q_2^T(S^{(k)T}U^{(k+1)})Q_1 \tag{2.41a}$$

$$\widetilde{V}_{ij} = \frac{\Xi_{ij}}{\Lambda_{ii}^{(1)} + \Lambda_{jj}^{(2)}}. \tag{2.41b}$$

Then the subproblems (2.39a)-(2.39c) are explicitly solved by (2.42a)-(2.42c).

$$U^{(k+1)} = (\Gamma^{(k)}V^{(k)})\left(V^{(k)T}V^{(k)} + \lambda_1 I\right)^{-1} \tag{2.42a}$$

$$V^{(k+1)} = Q_2\widetilde{V}Q_1^T \tag{2.42b}$$

$$\Gamma^{(k+1)} = \mathcal{P}_{\Omega^c}(U^{(k+1)}V^{(k+1)T}) + \mathcal{P}_\Omega(Y) \tag{2.42c}$$

The matrix $V^{(k)T}V^{(k)}$ is symmetric positive semidefinite such that for $\lambda > 0$, the matrix $(V^{(k)T}V^{(k)} + \lambda_1 I)$ is positive definite and thus invertible. Using (2.42a), (2.42b) and (2.42c), we summarize an alternating minimization scheme for solving (2.38) in Algorithm 1. In this thesis we initialize the R basic trajec-

Algorithm 1: LMaFit [10]

Estimate rank as R;
Initialize $V^{(0)}$, $\Gamma^{(0)}$;
repeat
  Update $U^{(k+1)}$ using (2.42a);
  Update $V^{(k+1)}$ using (2.42b);
  Update $\Gamma^{(k+1)}$ using (2.42c);
until *convergence criterion*;

tories

$$V_{tr}^{(0)} = 1 + 3 \cdot \frac{r}{R}, \quad \forall t, r = 0, \ldots, R \tag{2.43}$$

as uniform over time but of increasing latent risk. The placeholder matrix is initialized simply as $\Gamma^{(0)} = Y$. The convergence criterion is chosen to be

$$\frac{\|U^{(k+1)}V^{(k+1)T} - U^{(k)}V^{(k)T}\|_F^2}{\|U^{(k+1)}V^{(k+1)T}\|_F^2} \leq \epsilon, \tag{2.44}$$

with $\epsilon = 10^{-4}$. Checking the convergence criterion is costly and therefore only done every 50th iteration. It is important to note that correct recovery of U, V depends on the rank estimate R and the regularization parameters.

2.5 Prediction

Recall that in formulating the regularized problem as a result of MAP estimation, we had to assume the sampled Gaussian likelihood (A1) and the independence of screening participation (A2). In the following, we explicate how these assumptions can be used to derive a prediction scheme for the future cervical cancer state. Also, we argue that early prediction of cervical cancer can be reduced to a binary problem and derive a prediction scheme in this latter context.

2.5.1 Predicting Cervical Cancer State

Initially we view the early prediction of cervical cancer state as a multiclass classification problem. Let $\boldsymbol{y} \in (\{0\} \cup \mathcal{S})^T$ denote the screening history of a given female encoded as described in Section 1.3 and let $s_t \in \mathcal{S}$ denote the cervical cancer state of the female at a future time t. Under (A1) and (A2), the conditional probability of s_t given $\boldsymbol{y}$ can be computed as

$$\begin{aligned} p(s_t \,|\, \boldsymbol{y}, \theta) &\propto \int_{\boldsymbol{m}} p(s_t \,|\, \boldsymbol{m}, \theta) \cdot p(\boldsymbol{m} \,|\, \boldsymbol{y}, \theta) d\boldsymbol{m} \\ p(s_t \,|\, \boldsymbol{y}, \theta) &\propto \int_{\boldsymbol{m}} p(s_t \,|\, \boldsymbol{m}, \theta) \cdot p(\boldsymbol{y} \,|\, \boldsymbol{m}, \theta) \cdot \pi(\boldsymbol{m})\, d\boldsymbol{m}. \end{aligned} \tag{2.45}$$

In the derivation of (2.28), we assumed a prior on $\pi(m)$. Even after inserting the assumed form of the prior and simplifying the integral, it is of such high dimensionality that its computation is intractable. Instead, we can imagine that we have at our disposal a training set with an observed matrix $Y^{(\text{train})}$ consisting of the encoded screening histories

$$\left\{ Y_{1,:}^{(\text{train})}, Y_{2,:}^{(\text{train})}, \ldots, Y_{N,:}^{(\text{train})} \right\}.$$

These do not contain but are sampled from the same population as $\boldsymbol{y}$. Training the SPMF / CPMF model on this set yields the latent risk matrix estimate

$\hat{M}^{(\text{train})}$. We can then view the rows of the estimated latent risk matrix of the training set as samples from the prior distribution $\pi(\boldsymbol{m})$. Moreover, we can approximate the conditional probability (2.45) as the integral over the empirical prior defined by the samples

$$\hat{p}(s_t \mid \boldsymbol{y}, \theta) \propto \sum_{i=1}^{N} p(s_t \mid \hat{M}_{i,:}^{(\text{train})}, \theta) \cdot p(\boldsymbol{y} \mid \hat{M}_{i,:}^{(\text{train})}, \theta). \tag{2.46}$$

We finish the conditional probability estimate by inserting for the sampled Gaussian likelihood

$$\hat{p}(s_t \mid \boldsymbol{y}, \theta) \propto \sum_{i=1}^{N} \exp[-\theta(s_t - \hat{M}_{it}^{(\text{train})})^2]\, p(\boldsymbol{y} \mid \hat{M}_{i,:}^{(\text{train})}) \tag{2.47}$$

where $s_t \in \mathcal{S}$. Using the probability estimates, we can predict the future cervical cancer state of the given female by

$$\hat{s}_t = \arg\max_{s_t} \hat{p}(s_t \mid \boldsymbol{y}, \theta) \cdot a_{s_t} \tag{2.48}$$

where a_{s_t} is a bias term. We refer to the Smooth Matrix Factorization (SPMF) classifier as the scheme in which the latent risk matrix of the training set is estimated using the SPMF regularization model and predictions are computed using (2.51). Similarly we refer to the Convolutional Matrix Factorization (CPMF) classifier as the scheme in which the training set is estimated using the CPMF regularization model and predictions are computed using (2.51). The prediction parameter θ and the regularization parameters λ_1,λ_2 and R are referred to as the *hyperparameters* of the classifiers.

2.5.2 The Difficulty of The Multiclass Bias

The medical nature of the problem implies that wrongfully predicting females of a high-grade or cancer state to the low-grade or normal state is significantly worse than the other way around. To avoid under-screening, we can set a higher bias towards the high-grade and cancer states. However, for the multiclass problem, this begs the question in what manner these should be prioritized. While it is unfortunate that the algorithm fails to predict a high-grade or cancer state, it is difficult to argue how harmful a missed cancer is in comparison to a missed high-grade state.

To circumvent this challenge, we can narrow the scope of what we want to predict. We are most interested in predicting the sudden transmission from a state indicating the female is healthy to a state indicating that the female is sick. Basing ourselves on discussion conducted as part of the DeCipher project, we argue the following: Both the normal and low-grade states indicate that the female is healthy. Females diagnosed with the low-grade state are expected to regress to normal. Conversely, both the high-grade and cancer states indicate that the female is sick. The high-grade state implies an immediate risk of

developing cervical cancer even though it has not fully developed. With the intent of predicting whether action will be required at a future time, we can reduce the number of states to two: healthy and sick.

2.5.3 The Binary Prediction Model

We adopt a simplified view of the early prediction of cervical cancer and refer to females of the normal or low-grade state as *healthy* and women of the high-grade or cancer state as *sick*. For a given female, we create the variable $b_t \in \{0, 1\}$ encoding the binary disease status of the female at time t,

$$b_t = \begin{cases} 1, & s_t \in \{3, 4\} \\ 0, & s_t \in \{1, 2\}. \end{cases} \tag{2.49}$$

We compute binary probability estimates

$$\hat{p}_b(b_t \,|\, \boldsymbol{y}, \theta) = \begin{cases} \hat{p}(3 \,|\, \boldsymbol{y}) + \hat{p}(4 \,|\, \boldsymbol{y}), & b_t = 1 \\ \hat{p}(1 \,|\, \boldsymbol{y}) + \hat{p}(2 \,|\, \boldsymbol{y}), & b_t = 0, \end{cases} \tag{2.50}$$

by mapping over the state probability estimates of (2.46). The subscript b is included to differentiate these estimates from those in the multiclass context. As before we use the probability estimates to derive binary predictions

$$\hat{b}_t = \begin{cases} 1, & \hat{p}_b(1 \,|\, \boldsymbol{y}, \theta) \geq \delta \\ 0, & \text{otherwise}, \end{cases} \tag{2.51}$$

where $\delta \in [0, 1]$ is a bias towards the sick state. We refer to the Binary Smooth Matrix Factorization (B-SPMF) classifier as the scheme in which the latent risk matrix of the training set is estimated using the SPMF regularization model and predictions are computed using (2.51). Similarly we refer to the Binary Convolutional Matrix Factorization (B-CPMF) classifier as the scheme in which the latent risk matrix of the training set is estimated using the CPMF regularization model. The binary classifiers permit the same hyperparameters as in the multiclass context.

Now the bias term is a single variable and more easily interpreted. A focal point of this thesis is that in the proposed binary classifiers, the bias can be chosen freely to match the intended behavior of the classifier.

Chapter 3

Simulation

The Screening Dataset, as illustrated in Figure 3.1, is extremely sparse, integer-valued, and its nonzero observations are irregularly spaced along the time dimension. This poses a significant challenge in the training procedure of the classifiers. When applied to real data, we do not know the ground truth latent risk matrix and cannot ensure that the models behave as intended. To investigate the convergence of the algorithm, we first construct a simulation model that can produce data subject to the same challenges as the Screening Dataset but for which we know the ground truth.

In this section, we describe the Discrete Gaussian Distribution (DGD) and Hidden Markov (HMM) simulation models for generating the complete state matrix. We design the former as part of this thesis with the probabilistic matrix factorization classifiers in mind. The DGD model adheres to (A1) and samples the states from the underlying latent risk through the sampled Gaussian distribution. The second model is a model developed by Soper et al. [22] and previously trained on data collected by the NCCSP. It was designed to produce complete state matrices replicating the rate of transmission from normal to pre-cancerous states in the Nordic population.

Furthermore, we describe a model for simulating screening attendance, i.e., simulating the observation mask. The model should reproduce the irregularity and sparsity of the Screening Dataset. Combining the mask simulation model with either the DGD or HMM, we generate synthetic observed state matrices by simulating the complete state matrix and projecting it onto the simulated observation mask.

3.1 Discrete Gaussian Distribution

In the DGD simulation model, the ground truth latent risk matrix is chosen as a weighting of R predetermined basic disease trajectories. We choose $R = 5$ and determine the basic disease trajectories by

$$V_{tk} = \exp\left[-10^{-3}(t - c_k)^2\right], \tag{3.1}$$

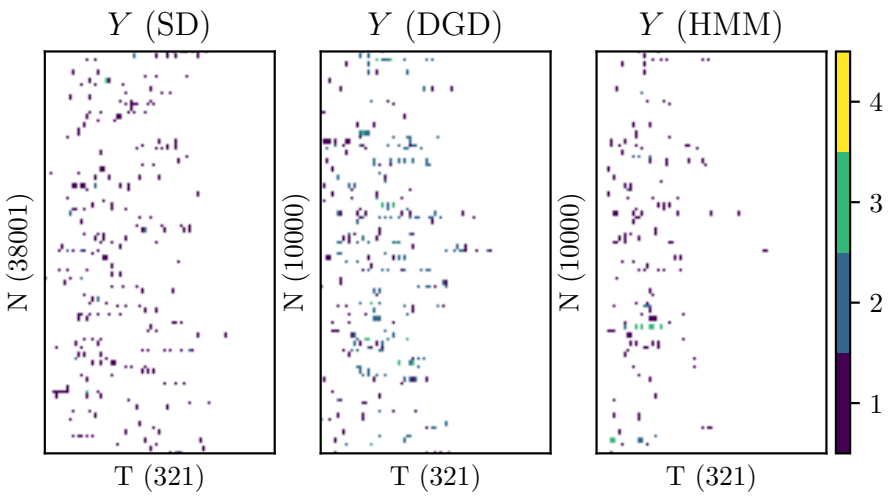


Figure 3.1: Visualization of the Screening Dataset and the observed state matrices of the DGD and HMM models. The matrices are uniformly undersampled along with the patient- and time-dimension to produce 100 rows and 200 columns. A colored square represents an observed screening whereas a white square represents a missing entry.

where $\boldsymbol{c} = (70, 95, 120, 145, 170)$. The value 10^{-3} is chosen such that the basic disease trajectories undergo a moderately narrow peak of high latent risk. In addition the trajectories are temporally smooth. The entries of the coefficient matrix U are sampled from the standard exponential distribution, $U_{ik} \sim \text{Exp}(1)$. To satisfy (A1), the synthetic complete state matrix is entrywise sampled from the latent risk matrix through the sampled Gaussian likelihood

$$p(S_{it} \mid M_{it}, \theta) \propto \exp\left[-\theta(S_{it} - M_{it})^2\right], \quad S_{it} \in \mathcal{S}, \tag{3.2}$$

where the kernel parameter $\theta = 2.5$. This value was chosen to yield a proportion of the high-grade and cancer states similarly small as in real screening data. The likelihood is discrete and we sample the states using inverse transform sampling.

3.2 Hidden Markov Model

The Hidden Markov (HMM) model differs from the DGD model in that there exists no latent risk matrix M. Instead, the complete state matrix is sampled using a set of transmission probabilities. The model was developed by Bradel et al. and is described [22]. Therefore, we provide only a high-level overview of the model. The complete state matrix S is sampled row-wise. For a single female, we sample the initial state $s_0 \in \mathcal{S}$ with the probabilities of Table 5

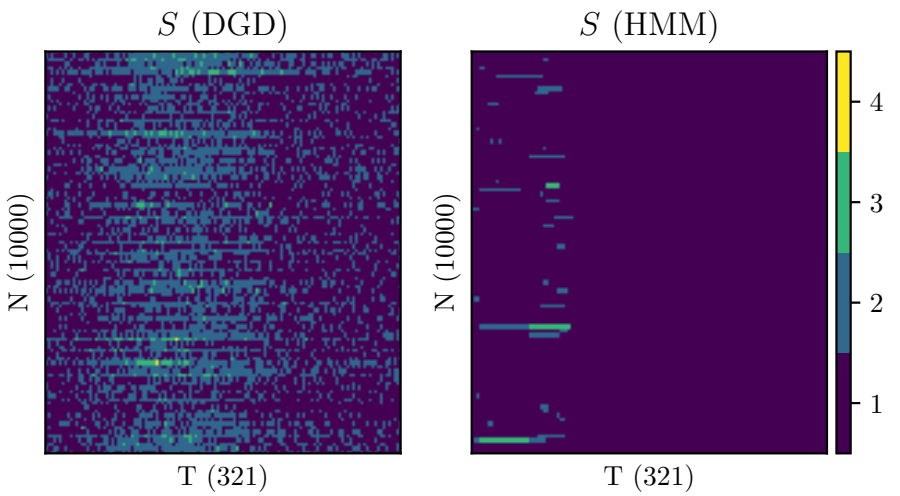


Figure 3.2: Visualization of the complete state matrices of the DGD and HMM models. The matrices are uniformly undersampled along the patient- and time-dimension to produce 100 rows and 200 columns.

in [22]. Further, we sample the time spent in this state is a random variable with cumulative distribution determined by Proposition 1 of [22]. Upon exiting the initial state, we sample the state s_t from the discrete distribution determined by the intensities in Table 6 and Table 7 in [22], which depend on the female's age and state. We now consider the second state of being the initial state and repeating the given steps. The entire process is repeated until the accumulated time interval exceeds T, at which point we have generated the complete state vector of a single female. This process is repeated N times to generate the complete state matrix.

Observe from Figure 3.2 that the DGD and HMM models produce complete state matrices of a different nature. In the former, the females frequently transition from normal to low-grade and vice versa. On average, the sampled states correspond to the latent risk from which they were sampled. In HMM data, such transitions occur only once or twice in the entire lifetime of a simulated female.

3.3 Simulating Screening Attendance

As the final step in the generation we simulate the observation mask Ω. To satisfy (A2) this can be done using

$$P((i,j) \in \Omega) = \epsilon, \tag{3.3}$$

where $\epsilon \in [0,1]$ is the intended sparsity of the resulting mask. However this assumes that the probability of a patient choosing to undergo a screening is

constant. In other words, it is independent of past screening outcomes and the time spent since the last screening. This offends our intuition and we instead argue that past screenings affect the patient's future screening participation. To simulate this behavior we devise a discrete random process where the screening probability is determined by the outcome of the last remembered screening,

$$P\left((i,t+1)\in\Omega\right)=\begin{cases} p_0, & & t-t'>\nu \\ p_1, & Y_{it'}=1, & t-t'\leq\nu \\ p_2, & Y_{it'}=2, & t-t'\leq\nu \\ p_3, & Y_{it'}=3, & t-t'\leq\nu \\ p_4, & Y_{it'}=4, & t-t'\leq\nu. \end{cases} \tag{3.4}$$

The index

$$t'=\underset{\tilde{t}\leq t}{\arg\max}\,(i,\tilde{t})\in\Omega, \tag{3.5}$$

is the time of the last observed entry, ν is a memory parameter of the patients and $p_s\in[0,1]$ is the probability of undergoing a screening given the result of last remembered screening. The screening probability is determined solely by the most recent remembered screening result. In the case that the patient had no screenings within the memory period $[t-\nu,t-1]$, the probability is set to a base probability $p_0\in[0,1]$. In this thesis we let the memory parameter $\nu=10$ and the probabilities

$$\begin{aligned} p_0&=0.01\cdot\xi \\ p_1&=0.03\cdot\xi \\ p_2&=0.08\cdot\xi \\ p_3&=0.12\cdot\xi \\ p_4&=0.04\cdot\xi. \end{aligned} \tag{3.6}$$

be constants multiplied by the global sparsity parameter $\xi\in[0,\frac{1}{0.12}]$. In the Screening Dataset we found the probability of screening to be increased after an observed low- or high-grade state. The parameters of (3.6) where chosen as we observed them to recreate this effect in the simulated data. Using this parametrization, the sparsity of the mask can be varied by altering the global sparsity parameter ξ. At the same time, the simulated behavior remains largely the same.

In the illustration of the Screening Dataset in Figure 3.1, we observe that after some time-index, there are no further observations for a specific female. This phenomenon, referred to as *censoring*, is caused by a multitude of reasons. We recall that the NCCSP recommends regular screening only for women of the ages 25 to 69. The censoring affects the training of the models as the future development of that female is then completely unknown. We extend the simulation model to incur the same loss of information in the synthetic data. We sample censoring times from a Beta-binomial with parameters $\alpha=4.57$ and $\beta=5.47$, which we obtained by fitting the distribution to the censoring times in the Screening Dataset using maximum likelihood. Figure 3.3 shows the empirical distribution of the Screening Dataset and the fitted distribution. All

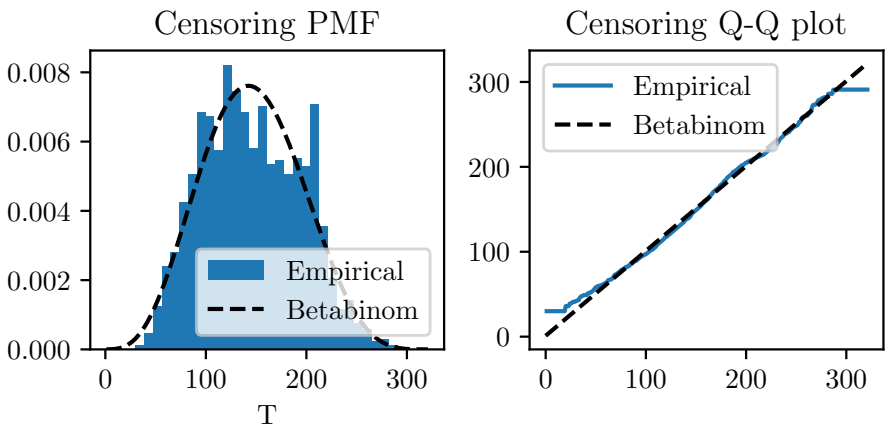


Figure 3.3: Empirical PMF of censoring in the Screening Dataset along with the PMF of the fitted Betabinomial distribution. The two distributions are compared using a quantile-quantile (Q-Q) plot.

Table 3.1: The distribution of cervical cancer states in the Screening Dataset and synthetic data generated by the HMM and DGD models. The frequencies are listed as percentages.

Dataset \ State	1	2	3	4
SD	92.964	4.656	2.344	0.036
DGD	43.881	51.587	4.454	0.078
HMM	86.540	8.734	4.724	0.002

observations after the censoring time $t_i^{(\mathrm{ct})}$ of a particular female are removed from the simulated mask. We note that synthetic datasets are generated with censoring unless explicitly specified otherwise.

3.4 Comparison

Figure 3.1 shows a visual comparison of DGD data, HMM data, and real screening data. Observe that the mask simulation model successfully reproduces the sparsity and irregularity of the Screening Dataset. As the probability of screening is higher after previously observing a low- or high-grade state, there is a clustering behavior around such observations. The censoring effect is also apparent in the simulated data.

Table 3.1 shows the distribution over the states of the three types of data. While the HMM and Screening Data are dominated by the normal state, the DGD contains an equally large proportion of the low-grade state. In addition the three datasets differ somewhat in the proportion of the high-grade and cancer

Table 3.2: The rates of transmission between the states in the Screening Dataset and synthetic data generated by the HMM and DGD models. The rates are listed as percentages normalized by the "From" state / row.

SD

From \ To	1	2	3	4
1	95.820	3.225	0.931	0.024
2	62.166	27.900	9.858	0.076
3	46.339	10.970	42.239	0.451
4	40.000	2.222	55.556	2.222

DGD

From \ To	1	2	3	4
1	58.600	40.720	0.678	0.002
2	31.158	63.740	5.084	0.018
3	5.331	59.182	34.497	0.989
4	0.439	14.474	68.860	16.228

HMM

From \ To	1	2	3	4
1	98.709	1.176	0.114	0.001
2	11.157	87.278	1.565	0.000
3	0.866	2.307	96.802	0.025
4	100.000	0.000	0.000	0.000

state.

The simulated data differ also in the rate of transition between the states. Table 3.2 the rate of transition to a given state categorized and normalized by the current state. Across all datasets, there is a tendency to remain in the normal state. It is interesting that females of a low- or high-grade state in the Screening Dataset also display a tendency to regress directly to the normal state. We do not observe this trend in the simulated datasets; in both HMM and DGD data, females of a high-grade or cancer state are more likely to transition into the low-grade state.

In this section, we have described two methods for generating sparse, integer-valued, and irregularly observed data. We were able to replicate the sparsity and clustering behavior of the observation mask. However, the generated synthetic data differ from real data in distribution and transitionary behavior.

Chapter 4

Reconstructing the Latent Risk Matrix

In this section, we investigate if we reconstruct the latent risk matrix by training the SPMF and CPMF models. To measure reconstruction at a specific point we define the *pointwise absolute error* (PAE)

$$[\text{PAE}]_{it} = |M_{it} - [UV^T]_{it}| \tag{4.1}$$

as the difference in absolute value between the estimated latent risk and ground truth at the point in question. To measure reconstruction across an entire dataset, we define the *reconstruction mean-squared error* (recMSE)

$$\text{recMSE} = \frac{\|\mathcal{P}_{\Omega^c}(M - UV)\|_F^2}{(1 - |\Omega|)}. \tag{4.2}$$

as the sum of all squared differences between the latent risk estimate and ground truth latent risk at the unobserved entries. The recMSE is a measure of how well the model extrapolates the temporal trends of the observed entries into points of time where we have no information of the patients. Both measures can only be computed if the ground truth latent risk matrix M is known. For that reason, we investigate reconstruction only for simulated DGD data.

4.1 Convergence

We simulate the fully observed state matrix S according to the DGD model with $N = 10000$ and simulate screening attendance with global sparsity parameter $\xi = 0.6$ both with and without censoring. Finally, we train the SPMF and CPMF models on the resulting state matrix Y.

Figure 4.1 shows recMSE as a function of iteration number in Algorithm 1. We note that the training procedures of the two SPMF and CPMF models are

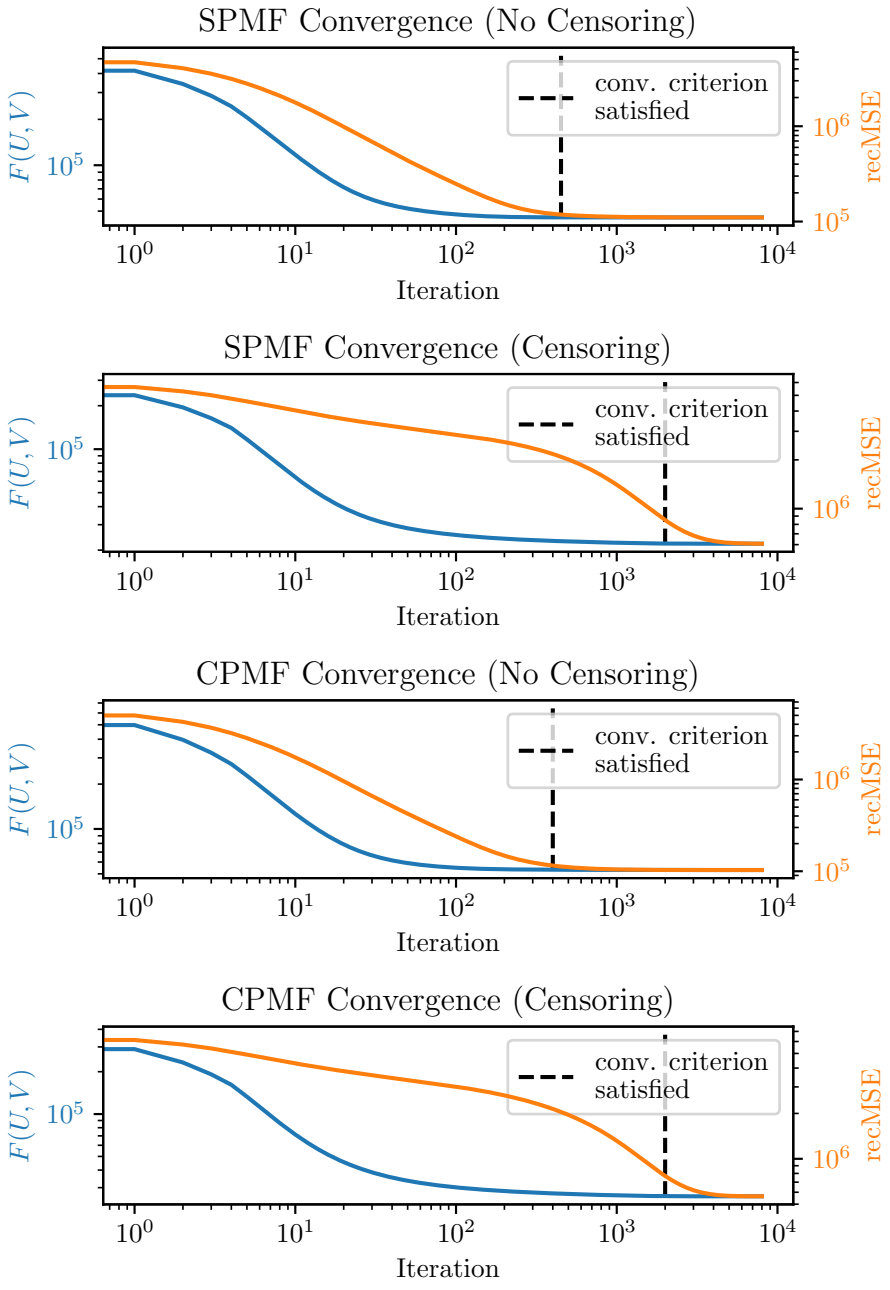


Figure 4.1: Convergence in the training procedure of the SPMF and CPMF models for DGD data. The objective and recMSE are plotted on separate log-log scales as a function of iteration number in Algorithm 1.

identical. Therefore the following applies to both. We observe that for data simulated without censoring, the objective and recMSE converge to their respective lower bounds within the first three hundred iterations. The convergence criterion is satisfied at the exact point where subsequent iterations yield no visible decrease in the recMSE. Interestingly, introducing censoring in the simulated data changes the convergence behavior. For censored data, the training can be divided into two phases. In the first phase, the objective rapidly decreases, whereas the recMSE only undergoes a moderate decrease. This continues until the algorithm has run for 1000 iterations. In the second phase, the convergence rate of the recMSE increases despite there being no visible decrease in the objective. This implies that when training the models on the Screening Dataset, convergence cannot be determined from the reduction in the objective. To avoid early termination, the criterion in (2.44) is instead based on the norm of the difference between subsequent estimates of the latent risk matrix. Even so, applying this rule to DGD data with censoring terminates the algorithm despite subsequent iterations yielding a small but significant decrease in the recMSE.

4.2 The Effect of Data Sparsity

The sparsity of the Screening Dataset is a central theme in this thesis. With that in mind, it is interesting to study if the sparsity of the observed state matrix Y affects the recMSE obtained. Figure 4.2 shows the PAE of the estimated latent risk matrices of the CPMF model trained on DGD data with varying sparsity parameters. We first consider the sparsest example ($\overline{\Omega} = 0.04$). Even though the algorithm has converged, there remains a considerable difference between the ground truth and the reconstruction. This is expected as certain rows contain almost no nonzero observations. From only a few entries, the algorithm is not able to estimate the coefficient matrix. We increase the density ($\overline{\Omega} = 0.08$) and observe the PAE to decrease in a localized fashion. Some rows now contain additional screenings, and the algorithm can use these to estimate the corresponding latent risk profiles correctly. Other rows display no change even though the dataset, as a whole, contains more nonzero entries. Finally, as the density becomes high ($\overline{\Omega} = 0.33$), estimation is accurate for all but some rows. The censoring effect still complicates reconstruction for higher ages, but the effect is weakened.

From the examples above, we suspect that the model's ability to reconstruct the latent risk matrix depends continuously on the density of the data. Figure 4.3 confirms this and shows that the rate of convergence is sublinear for DGD data. Interestingly, the CPMF model outperforms the SPMF model across all densities. We recall the true basic disease trajectories of the DGD data in Section 3.1. The flexibility of the CPMF allows the estimated basic disease trajectories to undergo a net change over a short time interval. While this was implemented to accommodate the spurious jumps in the Screening Dataset, it turns out to be amenable also to the sharp peaks in the true basic disease trajectories of DGD data. The regularization towards such peaks in the CPMF

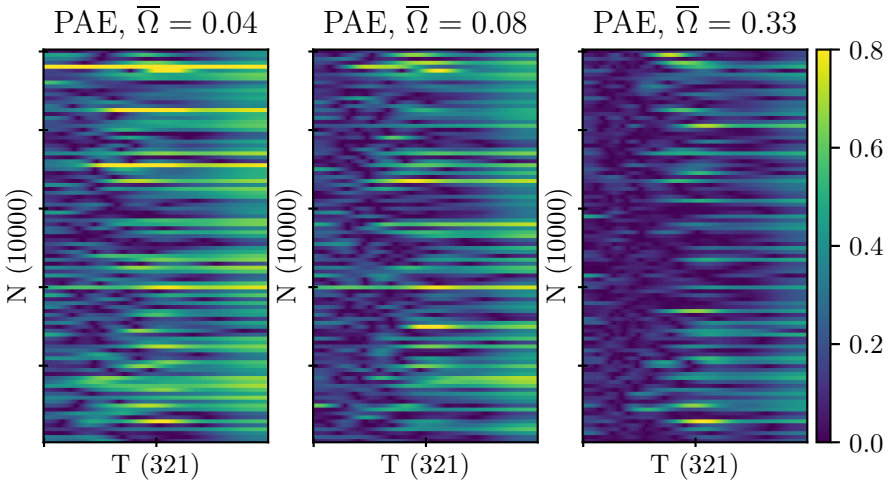


Figure 4.2: Pointwise absolute error (PAE) of reconstruction attained by training the CPMF model on three examples of DGD data. The data is simulated with $N = 10000$ and $\xi \in \{0.6, 1.0, 3.0\}$ (left to right). The model is trained with hyperparameters $R = 5, \lambda_1 = 1, \lambda_2 = 1000$.

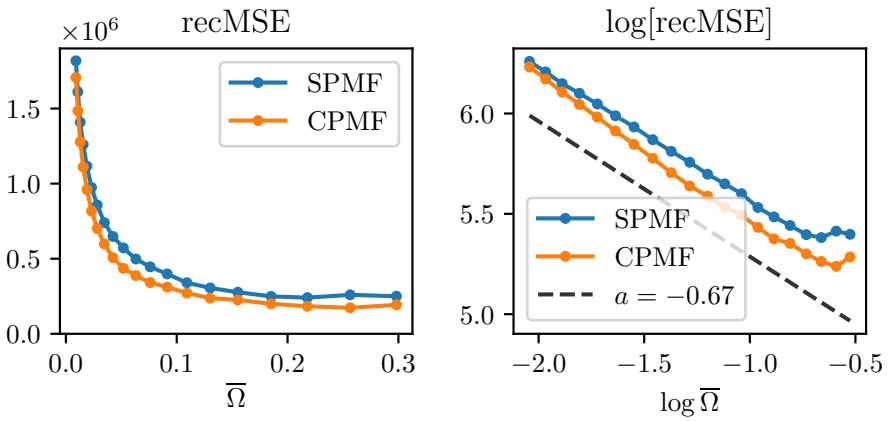


Figure 4.3: recMSE of the SPMF and CPMF models as a function of the density of the DGD data. The data is simulated with $N = 10000$ and $\xi \in [0.2, 2.0]$. The model is trained with hyperparameters $R = 5, \lambda_1 = 1, \lambda_2 = 1000$.

enables the model to achieve more accurate reconstruction.

In this section, we have demonstrated that the SPMF and CPMF models successfully reconstruct the latent risk matrix from the sparse, integer-valued, and irregularly observed state matrices generated by the DGD simulation model. We found the censoring effect described in Section 3.3 to negatively affect the convergence of the algorithm and demonstrated that the models' ability to reconstruct depends on the sparsity of the input in a continuous fashion.

Chapter 5

Predicting Cervical Cancer State

Early prediction of cervical cancer state can be viewed as a multiclass classification problem. In the following section, we use the proposed SPMF and CPMF models described in Section 2 to perform such classification. The performance of the models is evaluated on the simulated datasets and the Screening Dataset.

5.1 Preprocessing

As the prediction scheme requires a pretrained latent risk matrix estimate, predicting cervical cancer state is a two-step process. In the first step, the SPMF / CPMF models are trained using a training set to produce a latent risk matrix estimate. The trained model is then used to assign the most likely future state for a prediction population. Neither the simulated datasets nor the Screening Dataset are preassigned to a training/prediction population, but we can create such a division using a row-wise split

$$Y = \begin{bmatrix} Y^{(\text{train})} \\ Y^{(\text{pred})} \end{bmatrix} \tag{5.1}$$

where $Y^{(\text{train})} \in (\{0\} \cup \mathcal{S})^{N_1 \times T}$, $Y^{(\text{pred})} \in (\{0\} \cup \mathcal{S})^{N_2 \times T}$ and $N_1 + N_2 = N$. To replicate a scenario in which the prediction algorithm is of diagnostic usefulness, its input should consist only of past observed states. To create such input we must from the prediction set extract a *regressor matrix*, defined by

$$[Y^{(\text{regressor})}]_{ij} = \begin{cases} [Y^{(\text{pred})}]_{ij} & j < t_i - \Delta t \\ 0 & j \geq t_i - \Delta t. \end{cases} \tag{5.2}$$

where $\Delta t \in \mathbb{N}$ is the size of the smallest prediction window. In this thesis this is set to $\Delta t = 4$. Observe that the rows of the regressor matrix $Y^{(\text{regressor})}$ is nonzero only prior to the prediction window. This ensures that when predicting

the future cervical cancer state for a given patient we rely only on the subset of that patient's screening history recorded at least Δt periods before the time of prediction. By setting $\Delta t = 4$ we replicate the scenario in which we want to predict the cervical cancer state of a patient at least one year ahead in time.

To assess the accuracy of the classification procedure we extract also the *true labels* as

$$y_i = Y_{it_i}^{(\text{pred})}. \tag{5.3}$$

such that the predictions can be compared to the true labels. In the extraction of both the regressor matrix and the true labels, the time of prediction t_i is chosen as the last observed entry for that patient. The probability estimates

$$\hat{p}(y_i \,|\, Y_i^{(\text{regressor})}, \theta) \tag{5.4}$$

computed using (2.46) are used to assign class probability estimates for patient i at time t_i. Basing ourselves in the arguments made in Section 2.5.2 we in this multiclass context choose to forego a bias and compute predictions simply as the estimated maximum posterior probability state

$$\hat{y}_i = \arg\max_{y_i \in \mathcal{S}} \hat{p}(y_i \,|\, Y_i^{(\text{regressor})}, \theta). \tag{5.5}$$

We create the split (5.1) using two strategies. The first is used for simplicity and ease of interpretation and consists of choosing $N_1 = \lceil 0.8 \cdot N \rceil$, i.e., choosing the training set to be the first 80% of the rows in the dataset. In the second strategy, we implement 5-fold *cross-validation* for increased accuracy in the measured performance. This divides the dataset evenly into five partitions, assigns four of the partitions to the training set, and assigns the remaining to the prediction set. We then train the model on the training set and evaluate it on the prediction set. We increment the index of the prediction partition sequentially and repeat the entire process for a total of five runs. Finally, we average the performance measures attained on the five partitions to yield a single result.

5.2 Metrics

To evaluate the success/failure of the multiclass classifiers we make use of the visual *confusion matrix* tool and two scoring metrics. A confusion matrix details the number of patients assigned to the different classes categorized by the class of the true label. The row of an entry specifies the true class and the column specifies the predicted class. For the classification of cervical cancer state, an example is shown in Table 5.1. The confusion matrix provides a complete view of the classification results. To score the overall success/failure of the classification schemes, we extract from the confusion matrix two informative summary metrics. The *accuracy* of a classification scheme

$$\text{ACC} = \frac{\sum_k T_k}{\sum_k T_k + \sum_k \sum_l F_{kl}}, \tag{5.6}$$

Table 5.1: A template confusion matrix.

		Predicted			
		1	2	3	4
True	1	T_1	F_{12}	F_{13}	F_{14}
	2	F_{21}	T_2	F_{23}	F_{24}
	3	F_{31}	F_{32}	T_3	F_{34}
	4	F_{41}	F_{42}	F_{43}	T_4

is classical and describes the ratio of the number of total observations that were correctly predicted. A perfect algorithm will achieve an accuracy of ACC $= 1$ and thus we may seek an algorithm that is as close to this as possible. However, the accuracy lends itself to trivial solutions when the classes are imbalanced. We saw in Section 3 that of the nonzero entries in the Screening Dataset, 93% of the entries described a normal state. Thus even the trivial solution of predicting the future cervical cancer state to be normal for all patients regardless of screening history can be expected to achieve an accuracy of 0.93. To rely not only on the accuracy we also implement the Gorodkin R_K statistic [23]

$$R_K = \frac{\sum_k \sum_l \sum_m T_k F_{lm} - F_{kl}F_{mk}}{\sqrt{\sum_k (\sum_l F_{kl})(\sum_{k' \neq k} \sum_{l'} F_{k'l'})} \sqrt{\sum_k (\sum_l F_{lk})(\sum_{k' \neq k} \sum_{l'} F_{l'k'})}}, \quad (5.7)$$

which is a generalization of the Matthews correlation coefficient to the multiclass problem. The R_K statistic is designed to favor non-trivial solutions also in the case of heavily imbalanced data.

5.3 Baselines

We compare the performance of the SPMF and CPMF classifiers to a set of easily implementable *baseline* models.

5.3.1 The Forward Fill Baseline

The *Forward Fill* (FF) baseline is defined by

$$\bar{y}_i = Y_{il_i}^{(\text{regressor})} \quad (5.8)$$

where $l_i < t_i - \eta$ is the index of the last nonzero entry in the regressor vector $Y_i^{(\text{regressor})}$. In this manner, FF predicts the future cervical cancer state to be equal to the state observed at the last screening. The forward fill scheme is well-founded in the context of cervical cancer as it assumes the state of the patient will remain in place and be the same at a later time. Recall from Table 3.2 that we observed this often also to be the case. Still, the FF is of little value in practice; If a patient has recently been screened to the high-grade or cancer

state, then the scenario that we want to prevent has already occurred. Rather it is of interest to predict the sudden transmission from a normal or low-grade state to a high-grade or cancer state before it is realized. The FF will never predict such a development.

5.3.2 The Oracle Baseline (DGD)

For the data simulated by the DGD model, we can use the ground truth latent risk matrix M to formulate an additional baseline. The *Oracle baseline*

$$\tilde{y}_i = \arg\max_{y_i} p(y_i \mid M_{it}^{(\text{pred})}, \theta) = \lceil M_{it}^{(\text{pred})} - 0.5 \rceil. \tag{5.9}$$

is aptly named as it can only be used in the artificial example in which the latent risk matrix M_{it} is known. The Oracle predicts the highest posterior probability state under (A1) and (A2).

5.4 Prediction of State in DGD data

We simulate the fully observed state matrix S as in Section 3.1 with $N = 10000$. Moreover, we simulate screening attendance as in Section 3.3 with global sparsity parameter $\xi = 0.6$. We preprocess the resulting observed state matrix Y as described in Section 5.1.

The confusion matrices of the SPMF, CPMF, FF and Oracle classifiers are shown in Table 5.2-5.5. The SPMF and CPMF classifiers display similar results; both successfully predict a large portion of the females of a normal or low-grade states. At the same time, both misclassify most females of a high-grade state to low-grade. We recall the parameter choices required to reproduce the characteristics of the Screening Dataset using the DGD simulation model. To yield only a small number of the high-grade or cancer states, the vast majority of the latent risk profiles were chosen to be centered around a low risk. As a result, the few high-grade or cancer states will have originated from latent profiles of low risk. In other words, the classifiers wrongly predict the low-grade state as this is actually most likely. We confirm this by comparing the performance to that of the Oracle classifier. The MF classifiers perform almost as well as the Oracle, and also the latter misses most of the high-grade and cancer states.

Let now the global sparsity parameter ξ used in the simulation of screening attendance vary. This alters the density of the observed state matrix Y and allows us to investigate in what manner the density of the dataset affects the predictive performance of the SPMF and CPMF classifiers. We observe from Figure 5.1 that the accuracy and R_K of the classifiers increase with increasing density. The increase is rapid for lower densities, whereas it is moderate for larger densities. This behavior illustrates a point of significant interest. The density of the Screening Dataset lies in the region where the SPMF / CPMF classifiers experience a rapid performance increase. If the data we have available for the early prediction of cancer is enlarged, even just slightly, the accuracy with

Table 5.2: Confusion matrix of the SPMF classifier applied to synthetic DGD data generated with $N = 10000$ and $\xi = 0.6$. The SPMF algorithm was trained using $R = 5, \lambda_1 = 1, \lambda_3 = 1000$, and predictions were computed using $\theta = 2.5$.

		Predicted			
		1	2	3	4
True	1	821	306	0	0
	2	209	468	2	0
	3	4	20	3	0
	4	1	2	1	0

Table 5.3: Confusion matrix of the CPMF classifier applied to synthetic DGD data generated with $N = 10000$ and $\xi = 0.6$. The CPMF model was trained using $R = 5, \lambda_1 = 1, \lambda_3 = 1000$, and predictions were computed using $\theta = 2.5$.

		Predicted			
		1	2	3	4
True	1	818	309	0	0
	2	198	477	4	0
	3	3	21	3	0
	4	1	2	1	0

Table 5.4: Confusion matrix of the FF baseline applied to synthetic DGD data generated with $N = 10000$ and $\xi = 0.6$.

		Predicted			
		1	2	3	4
True	1	741	382	4	0
	2	203	443	32	1
	3	3	18	6	0
	4	1	0	3	0

Table 5.5: Confusion matrix of the Oracle baseline applied to synthetic DGD data generated with $N = 10000$ and $\xi = 0.6$.

		Predicted			
		1	2	3	4
True	1	847	280	0	0
	2	155	524	0	0
	3	1	22	4	0
	4	0	1	3	0

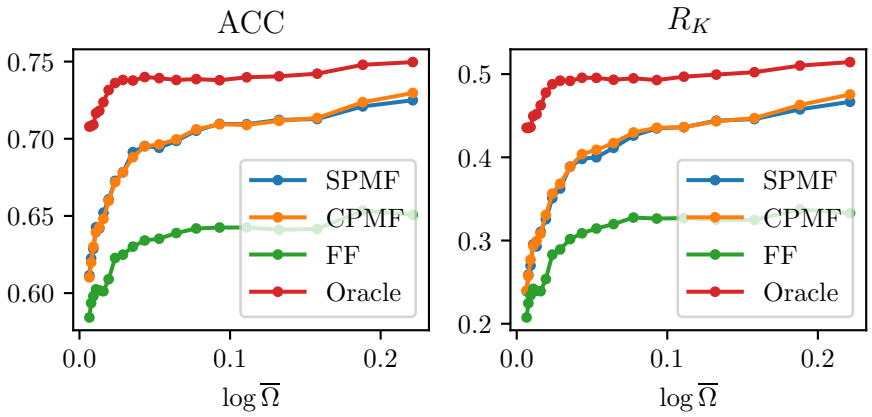


Figure 5.1: Accuracy and R_K of the SPMF classifier as a function of density in synthetic DGD data. The data is simulated with $N = 10000$ and $\xi \in [0.31, 3.16]$. The model is trained with hyperparameters $R = 5, \lambda_1 = 10, \lambda_2 = 100$ and predictions are computed using the hyperparameter $\theta = 2.5$.

which we can do so may improve significantly. We recall that when encoding the raw screening data, a time resolution of three months was chosen. In future extensions of this project, it is a compelling thought to increase the time step to six months or even a year.

5.5 Predicting of State in HMM data

We simulate the fully observed state matrix S as in Section 3.2 with $N = 10000$ and simulate screening attendance as in Section 3.3 with global sparsity parameter $\xi = 0.6$. Moreover, we preprocess the resulting observed state matrix Y as described in Section 5.1.

Table 5.6 and Table 5.7 show the confusion matrices of the SPMF and FF classifiers. We see that the SPMF algorithm successfully predicts females across all states. This may lead us to believe that the model has successfully estimated the future development of the females and assigns the most likely future state. Unfortunately, this need not be the case. We recall from Section 3.2 that the data simulated according to the HMM model were characterized by an extreme tendency to remain in the same state as was observed at the most recent screening. Thus the only trend that the SPMF algorithm needs capture is that the state will largely remain in place. By comparing the results to the FF baseline, we discover that the FF performs at least as well for the normal, high-grade, and cancer state and better for the low-grade state.

We argue that one should at this point question the validity of using the SPMF / CPMF classifiers on data generated by the HMM model. The clas-

Table 5.6: Confusion matrix of the SPMF classifier applied to synthetic HMM data generated with $N = 10000$ and $\xi = 0.6$. The SPMF algorithm was trained using $R = 5, \lambda_1 = 1, \lambda_3 = 1000$, and predictions were computed using $\theta = 5.5$.

		Predicted			
		1	2	3	4
True	1	1504	14	3	0
	2	45	51	1	0
	3	2	2	20	0
	4	0	0	0	0

Table 5.7: Confusion matrix of the FF baseline applied to synthetic HMM data generated with $N = 10000$ and $\xi = 0.6$.

		Predicted			
		1	2	3	4
True	1	1498	20	3	0
	2	29	67	1	0
	3	1	1	22	0
	4	0	0	0	0

sifiers are founded on the premise that a female's probability of developing cervical cancer follows the weighting of a small number of basic trajectories. If the classifiers are to outperform the FF baseline, these trajectories need to follow some development from low- to high risk or vice versa; the trajectories should at some point undergo a transition. While the HMM simulation model is implemented to include such age-dependence, this transitionary behavior is lost when the simulated screening attendance masks the complete state matrix. If one observes no or little a priori evidence that there is transitionary behavior in the dataset, one should also challenge the use of the SPMF and CPMF models. Based on these arguments, we refrain from conducting additional experiments on data simulated according to the HMM model.

5.6 Prediction of State in Screening Data

The Screening Dataset is preprocessed as described in Section 5.1. Table 5.8, Table 5.9 and Table 5.10 show the confusion matrices of the SPMF, CPMF and FF classifiers. The three classifiers behave very similarly, and all assign the vast majority of the patients to the normal state. Unlike for DGD data, none of the classifiers consistently predict females of the low-grade, high-grade, or cancer states. We recall from Section 3.4 that the rates of transition between the four states in the Screening Dataset are different from those in DGD data.

Table 5.8: Confusion matrix of the SPMF classifier applied to the Screening Dataset. The SPMF was trained using $R = 5, \lambda_1 = 1, \lambda_3 = 1000$, and predictions were computed using $\theta = 5.5$.

		Predicted			
		1	2	3	4
True	1	6583	99	16	0
	2	82	10	0	0
	3	32	3	2	0
	4	11	2	0	0

Table 5.9: Confusion matrix of the CPMF classifier applied to the Screening Dataset. The CPMF was trained using $R = 5, \lambda_1 = 1, \lambda_3 = 1000$, and predictions were computed using $\theta = 5.5$.

		Predicted			
		1	2	3	4
True	1	6588	94	16	0
	2	82	10	0	0
	3	29	8	0	0
	4	11	2	0	0

Table 5.10: Confusion matrix of the FF baseline applied to the Screening Dataset..

		Predicted			
		1	2	3	4
True	1	6582	71	44	1
	2	78	11	3	0
	3	28	5	4	0
	4	8	5	0	0

While simulated females in DGD data largely transition to neighboring states, the females in screening data transition to neighboring but also directly to and from the normal state. The normal state is the most likely scenario even for females of increased latent risk. Therefore the SPMF and CPMF classifiers, unless otherwise coerced, predict the normal state.

It is relevant to investigate whether the poor results are a consequence of the hyperparameters used. Ideally, we would evaluate the model on the Screening Data over an exhaustive range of possible combinations to fully eliminate the possibility that the models can be tuned to improve performance. However, the classifiers use four hyperparameters, and due to restrictions on time and computational resources, we limit ourselves to exploring them in a two-and-two fashion. We study model performance as a function of the parameters θ, R when the regularization parameters λ_1, λ_2 are fixed and similarly as a function of λ_1, λ_2 when θ, R are fixed.

Let initially $\lambda_1 = 1, \lambda_2 = 1000$. Table 5.11 lists the performance of the SPMF and CPMF models for a selection of the hyperparameters θ, R. Interestingly, we observe that the accuracy of the algorithm increases as θ decreases. For a small θ, the likelihood term in (2.47) is large even when the regressor profile and the estimated latent risk profile differ significantly. This has a middling effect; the prediction for a certain patient is affected strongly by all the estimated latent risk profiles in the training set. As the Screening Dataset is dominated by the normal state, this draws the predictions towards the normal state for all patients. If we use accuracy to deduce the best choice of hyperparameters, we select a trivial classifier. We instead turn our attention to the R_K statistic. This is greatly improved as $\theta = 0.5$ is increased to $\theta = 10.5$. For both models, the R_K statistic increases as the rank estimate is increased from $R = 1$ to $R = 5$. However, the statistic remains constant when the rank is further incremented. As the running time of the training procedure depends on the rank estimate, it should be kept small to reduce the computational cost of training the model.

Let now $\theta = 5.50$ and $R = 5$ be fixed. Figure 5.2 displays the performance of the SPMF classifier as a function of the regularization parameters. The R_K statistic is maximized for $\lambda_1 = 0.51$ and $\lambda_2 = 3831$. Neighboring hyperparameters offer similar performance. The regular spacing of the level curves indicates that the performance is unimodal, which would imply that it cannot be further improved by tuning the regularization. Figure 5.3 displays the performance of the CPMF classifier. The R_K statistic is maximized for $\lambda_1 = 0.51$ and $\lambda_2 = 562$. The level curves are slightly shifted but similar to those of the SPMF classifier. Also for the CPMF classifier the level curves lead us to believe that the performance is unimodal around the maximizer.

In this section, we have seen that neither the SPMF nor the CPMF classifiers improve upon the FF baseline in the prediction of cervical cancer state for the Screening Dataset. The classifiers behave in a foreseeable manner in response to variations in the hyperparameters, but we found no combination leading to performance exceeding that of the forward fill strategy.

Table 5.11: Accuracy and R_K of the SPMF and CPMF classifiers applied to the Screening Dataset as a function of the hyperparameter θ, R.

Metric		SPMF				
	$r \setminus \theta$	0.5	3.0	5.5	8.0	10.5
ACC	1	0.974	0.953	0.949	0.947	0.946
	3	0.973	0.956	0.950	0.947	0.945
	5	0.974	0.962	0.958	0.954	0.952
	7	0.975	0.965	0.962	0.959	0.958
	9	0.975	0.966	0.963	0.961	0.959
R_K	1	0.021	0.061	0.070	0.070	0.071
	3	0.034	0.085	0.092	0.097	0.099
	5	0.024	0.091	0.111	0.119	0.119
	7	0.022	0.097	0.110	0.114	0.119
	9	0.025	0.096	0.113	0.121	0.123
Metric		CPMF				
	$r \setminus \theta$	0.5	3.0	5.5	8.0	10.5
ACC	1	0.974	0.953	0.949	0.948	0.947
	3	0.973	0.956	0.950	0.947	0.945
	5	0.974	0.960	0.957	0.954	0.953
	7	0.974	0.961	0.957	0.955	0.954
	9	0.974	0.961	0.957	0.955	0.954
R_K	1	0.022	0.060	0.070	0.069	0.070
	3	0.029	0.088	0.094	0.102	0.105
	5	0.026	0.090	0.102	0.120	0.122
	7	0.028	0.092	0.101	0.119	0.123
	9	0.024	0.091	0.100	0.119	0.124
Metric		FF				
ACC		0.9644				
R_K		0.137				

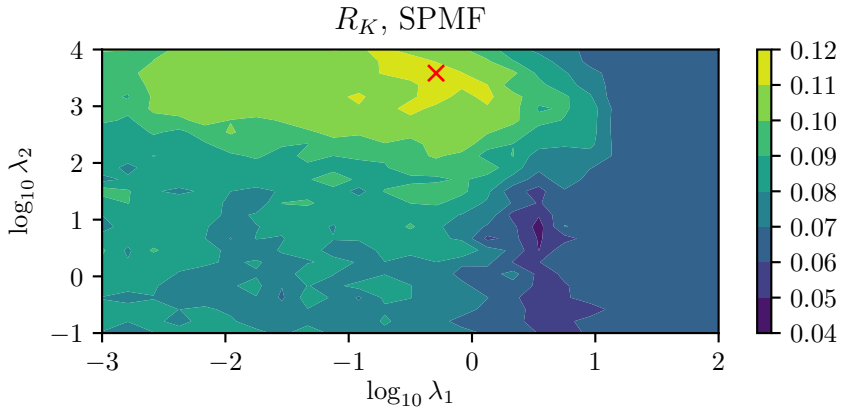


Figure 5.2: R_K of the SPMF classifier attained for the Screening Dataset as a function of the hyperparameters λ_1, λ_2. The model is trained with $R = 5$ and predictions are computed using $\theta = 5.5$.

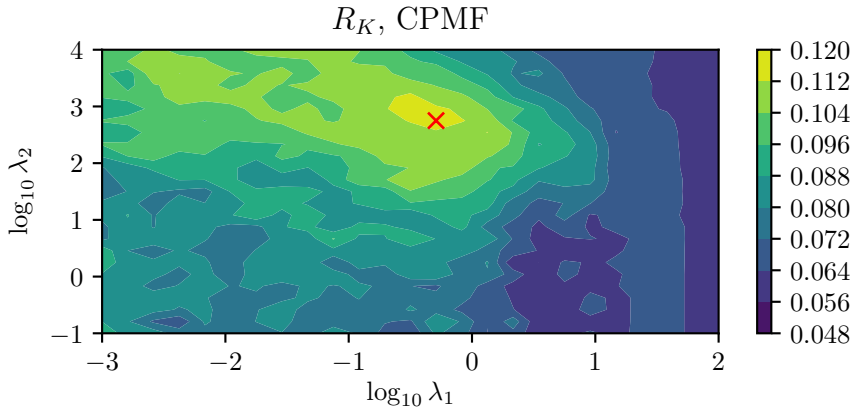


Figure 5.3: R_K of the CPMF classifier attained for the Screening Dataset as a function of the hyperparameters λ_1, λ_2. The model is trained with $R = 5$ and predictions are computed using $\theta = 5.5$.

Chapter 6

Binary Prediction of Cancer

As described in Section 2.5.3 we now reduce the problem to a dichotomy. In the binary context, the female is classified either as healthy or sick. In all following examples we preprocess the observed state matrix of the relevant dataset using the steps described in Section 5.1. The exception to this being that we redefine the true labels as

$$b_i = \begin{cases} 1, & Y_{it_i}^{(\text{pred})} \in \{3,4\} \\ 0, & Y_{it_i}^{(\text{pred})} \in \{1,2\}. \end{cases} \tag{6.1}$$

We predict the binary state of the females in the relevant dataset with the B-SPMF and B-CPMF classifiers described in Section 2.5.3. Specifically, we predict the positive binary state if the binary probability estimate

$$\hat{b}_i = \begin{cases} 1, & \hat{p}_b(1 \mid Y_i^{(\text{regressor})}) \geq \delta \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

exceeds the chosen bias $\delta \in [0,1]$. We will vary the bias throughout the following section to illustrate interesting properties of our proposed classifiers.

6.1 Metrics

To evaluate the success of the classifiers we again make use of the binary confusion matrix. This is similarly defined as in the multiclass context but with only two classes. An example is shown in Table 6.1. Two important metrics can be extracted from the binary confusion matrix. The *sensitivity* (SENS) of a binary classifier

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6.3}$$

is defined as the number of females correctly predicted as sick divided by the number of females who were sick at the time of prediction. As an example let $\text{SENS} = 0.40$. This implies that the classifier correctly predicted 40% of the sick

Table 6.1: A template binary confusion matrix.

		Predicted	
		0	1
True	0	TN	FP
	1	FN	TP

females. Conversely, the classifier wrongly predicted the remaining 60% to be healthy. To reduce the incidence and mortality of cervical cancer in the screening population, the sensitivity of the algorithm should be as large as possible. The *specificity* (SPEC)

$$\text{SPEC} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{6.4}$$

is another important metric defined as the number of females correctly predicted as healthy divided by the number of females who were healthy at the time of prediction. As an example let SPEC = 0.90. This implies that the classifier correctly predicted 90% of the healthy females. The remaining percentage of females were wrongfully predicted to be sick. To reduce over-treatment and program expenditure, the specificity of the binary classification algorithm should be as large as possible.

The essential feature of the probabilistic classifiers is that we can tune the bias δ of (6.2) to yield a compromise between sensitivity and specificity. We encourage the classifier to predict the females to be sick by setting the bias parameter to be small. This is expected to yield a high sensitivity but reduce the specificity of the algorithm. By letting the bias vary, we can study the number of sick females we can correctly specify for any given specificity. We conduct this using the Receiver Operating Characteristic (ROC) curve, which plots the sensitivity of the classifiers as a function of their specificity. In practice, there typically exists a constraint demanding that the specificity is above a specific limit. In this thesis, we assume this limit to be 0.75. By following the ROC curve, we find the maximum number of developing cases of cancer we can preempt under the specified constraint.

When evaluating the performance of our classifiers for a particular selection of hyperparameters, we use a metric that is independent of the choice of bias. The Area Under Curve (AUC) is found as the area under the ROC curve. The metric ranges from 0.5 to 1, where the former describes random guessing, and the latter describes a perfect classifier.

6.2 Baseline

The previous baselines are slightly modified for use in the binary context.

6.2.1 The Binary Forward Fill Baseline

We define the *Binary Forward Fill* baseline (B-FF) by

$$\overline{\mathrm{IR}}_i = \begin{cases} 1, & Y_{il_i}^{(\text{regressor})} \geq 2 \\ 0, & Y_{il_i}^{(\text{regressor})} < 2, \end{cases} \tag{6.5}$$

where as before $l_i < t_i - \eta$ is the index of the last nonzero entry in the regressor vector. The B-FF predicts all patients who had a low-grade state at their last screening to be positive.

6.2.2 The Binary Oracle Baseline (DGD)

The Oracle baseline is modified to define the *Binary Oracle* baseline (B-Oracle)

$$\widetilde{\mathrm{IR}}_i = \begin{cases} 1, & p(3 \,|\, M_{it}^{(\text{pred})}) + p(4 \,|\, M_{it}^{(\text{pred})}) \geq \eta, \\ 0, & p(3 \,|\, M_{it}^{(\text{pred})}) + p(4 \,|\, M_{it}^{(\text{pred})}) < \eta \end{cases} \tag{6.6}$$

where $\eta \in [0, 1]$ is also a bias parameter. As in the multiclass context, the B-Oracle baseline predicts treatment as if the latent risk of the patient were fully known and therefore predicts the binary outcome with highest posterior probability under (A1) and (A2).

6.3 Binary Prediction in DGD Data

We produce DGD data by simulating the fully observed state matrix S as in Section 3.1 with $N = 10000$ and simulating screening attendance as in Section 3.3 with global sparsity parameter $\xi = 0.6$. Furthermore, we extract the training matrix, the regressor matrix and the true labels are extracted as described above.

We initially assign a weak bias towards the sick state. Table 6.2-Table 6.5 show the results of the B-SPMF and the B-CPMF classifiers along with those of the B-FF and B-Oracle baselines. By shifting the predictions in favor of the sick state we increase the number of females predicted to be sick. As a direct result the sensitivity is increased and the specificity is decreased. This particular bias value is chosen as it leads the matrix factorization classifiers to achieve the same specificity as the B-FF. Similarly the bias of the B-Oracle is chosen to match the specificity of the other two. If we force the restriction that the four classifiers produce the same specificity, the B-SPMF, B-CPMF, and B-Oracle achieve a much higher sensitivity than the B-FF baseline. In other words if the classifiers are competing under similar terms, the matrix factorization classifiers outperform the B-FF baseline and are only slightly outperformed by the B-Oracle baseline.

It is interesting to see whether we can successfully predict more positive cases if we further decrease the bias. Figure 6.1 shows the sensitivity and specificity of the B-SPMF, B-CPMF, and B-Oracle classifiers as a function of their respective bias parameters. As the bias is increased from zero, the specificity rapidly

Table 6.2: Binary confusion matrix of the B-SPMF classifier applied to DGD data with bias $\delta = 0.15$. We trained the B-SPMF classifier using $R = 5, \lambda_1 = 1, \lambda_2 = 1000$ and computed predictions using $\theta = 2.5$.

		Predicted	
		0	1
True	0	1769	37
	1	17	14

Table 6.3: Binary confusion matrix of the B-CPMF classifier applied to DGD data with bias $\delta = 0.15$. We trained the B-SPMF classifier using $R = 5, \lambda_1 = 1, \lambda_2 = 1000$ and computed predictions using $\theta = 2.5$.

		Predicted	
		0	1
True	0	1767	39
	1	16	15

Table 6.4: Binary confusion matrix of the B-Oracle classifier applied to DGD data with bias $\eta = 0.20$ applied to synthetic DGD data.

		Predicted	
		0	1
True	0	1768	38
	1	16	15

Table 6.5: Binary confusion matrix of the B-FF baseline applied to DGD data.

		Predicted	
		0	1
True	0	1769	37
	1	22	9

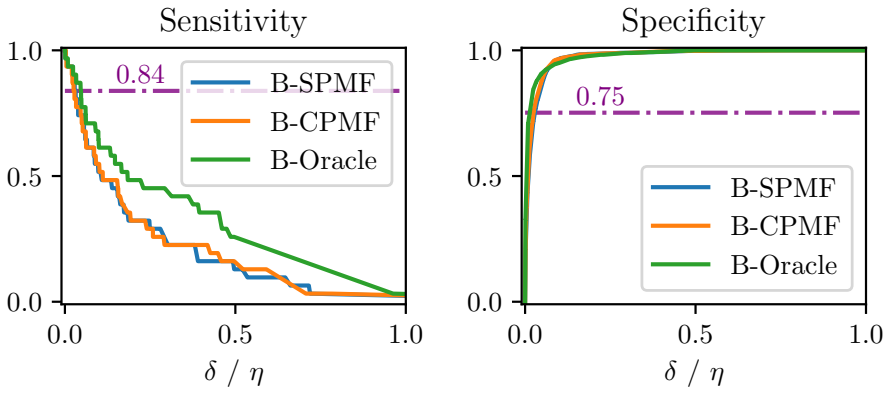


Figure 6.1: Sensitivity and specificity of the B-SPMF and B-Oracle classifiers attained on synthetic DGD data as a function of the bias parameters. The data was simulated with $N = 10000$ and $\xi = 0.6$. We trained the B-SPMF classifier hyperparameters $R = 5, \lambda_1 = 10, \lambda_2 = 100$ and computed predictions using $\theta = 2.5$.

increases. The sensitivity, on the other hand, does not decrease correspondingly fast. This implies that by choosing the bias appropriately, we can retain a high specificity while at the same time achieving high sensitivity. The possible compromises between sensitivity and specificity are found along the ROC curves of the classifiers shown in Figure 6.2. If we assume that constraints demand a specificity above 0.75, we could use the B-SPMF to attain a sensitivity of 0.84. The B-SPMF and B-CPMF display similar sensitivity for all values of the specificity, but the latter attains a marginally higher AUC. The B-Oracle baseline outperforms both classifiers.

We have observed that the bias parameter of the matrix factorization classifiers can be modified to predict females of a sick state in DGD data. Even though the B-SPMF and B-CPMF classifiers receive as input a regressor history that is considerably sparse and irregular, the methods performed nearly as well as the B-Oracle.

6.4 Binary Prediction in Screening Data

Finally, we investigate if the results of the previous section carry over to the Screening Dataset. This would imply that by implementing a matrix factorization classifier and carefully selecting the bias, we can predict developing cases of cervical cancer one year ahead in time.

We extract the training matrix, regressor matrix and true labels as described above. Initially we set the bias parameter $\delta = 0.0025$. Table 6.6-6.8 show the confusion matrices of the B-SPMF and B-CPMF classifiers along with those of the B-FF baseline. For this particular choice of bias, both matrix factorization classifiers behave similarly as the B-FF baseline. The specificity of the three

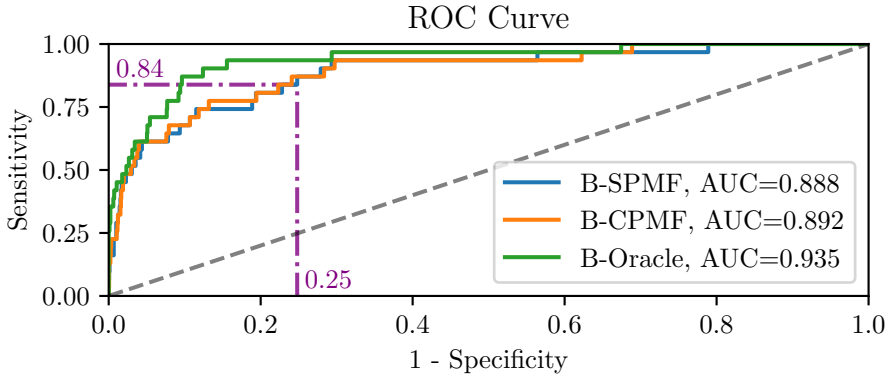


Figure 6.2: ROC curves of the B-SPMF and B-Oracle classifiers attained on synthetic DGD data simulated with $N = 10000$ and $\xi = 0.6$. We trained the B-SPMF classifier with hyperparameters $R = 5, \lambda_1 = 10, \lambda_2 = 100$ and computed predictions using $\theta = 2.5$.

Table 6.6: Binary confusion matrix of the B-SPMF classifier applied to the Screening Dataset data with bias $\delta = 0.0025$. We trained the B-SPMF classifier using $R = 5, \lambda_1 = 1, \lambda_2 = 1000$ and computed predictions using $\theta = 5.5$.

		Predicted	
		0	1
True	0	7300	235
	1	50	16

Table 6.7: Binary confusion matrix of the B-CPMF classifier applied to the Screening Dataset using a bias $\delta = 0.0025$. We trained the B-CPMF classifier with $R = 5, \lambda_1 = 1, \lambda_2 = 1000$ and computed predictions using $\theta = 5.5$.

		Predicted	
		0	1
True	0	7360	175
	1	52	14

Table 6.8: Binary confusion matrix of the B-FF baseline applied to the Screening Dataset.

		Predicted	
		0	1
True	0	7386	149
	1	49	17

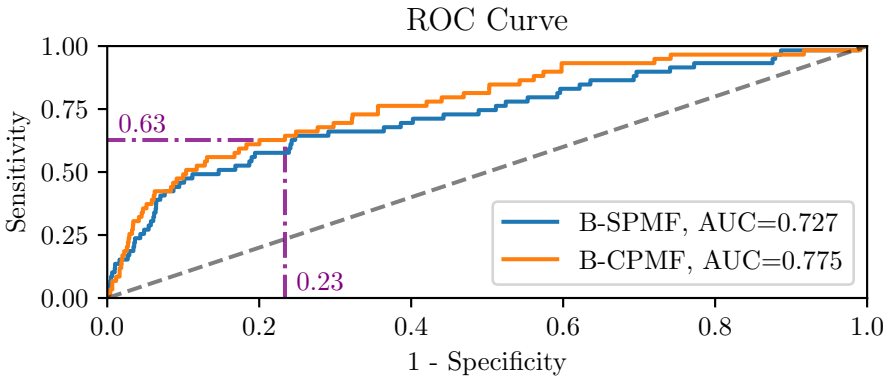


Figure 6.3: ROC curves of the B-SPMF and B-CPMF classifiers attained on the Screening Dataset. Both classifier are trained with hyperparameters $R = 5, \lambda_1 = 1, \lambda_2 = 1000$ and predictions are computed using the hyperparameter $\theta = 5.5$.

is high, and most of the healthy females are correctly predicted to be healthy. Conversely, the sensitivity is low, and many of the sick females are wrongfully predicted to be sick. Unlike for DGD data, the proposed classifiers do not achieve a higher sensitivity if the specificity is matched to that of the B-FF. As in the multiclass context, the transitionary behavior in the Screening Data obfuscates the temporal trends in the data.

However, the essential feature of the matrix factorization classifiers is that we can choose the bias to yield an arbitrary sensitivity. The problem of the forward fill strategy is that it is constrained to a single level of specificity/sensitivity. There are no parameters we can tune to make the B-FF classifier more sensitive to the sick state. Therefore, it is of lesser importance that the classifiers do not outperform the B-FF baseline for a single fixed specificity.

We consider the ROC curves of the B-SPMF and B-CPMF classifiers in Figure 6.3. The purple line indicates that if we allow the specificity to be as low as 0.77, we can attain a sensitivity of 0.63. In other words, if we accept that we will, on average, subject 25% of the healthy population to increased screening, then we can, on average, assign 63% of the sick population to similarly increased screening. To decide what increased screening entails is outside the scope of this thesis. However, we do conclude that using the females' previous screening history, the B-SPMF, and B-CPMF classifiers meaningfully discern females who should be subjected to such and females for whom this is unnecessary.

Judging by Figure 6.3, the B-CPMF appears to outperform the B-SPMF classifiers consistently. However, we recall from Section 5.1 that a single ROC curve is computed using a split containing only 20% of the data. To compare the two classifiers thoroughly, we compute their AUC using cross-validation for several possible combinations of the hyperparameters.

Let initially $\lambda_1 = 1$ and $\lambda_2 = 1000$. Table 6.9 shows the AUC of the B-SPMF and B-CPMF classifiers for a selection of the hyperparameters θ and R. The

Table 6.9: AUC of the B-SPMF and B-CPMF classifiers applied to the Screening Dataset as a function of the hyperparameters θ, R. Both classifiers are trained with hyperparameters $\lambda_1 = 1$ and $\lambda_2 = 1000$.

Metric		B-SPMF				
	$r \setminus \theta$	0.5	3.0	5.5	8.0	10.5
AUC	1	0.751	0.760	0.760	0.759	0.760
	3	0.716	0.773	**0.782**	**0.782**	0.779
	5	0.733	0.773	0.773	0.766	0.757
	7	0.735	0.773	0.773	0.765	0.756
	9	0.735	0.772	0.769	0.757	0.745
		B-CPMF				
	$r \setminus \theta$	0.5	3.0	5.5	8.0	10.5
AUC	1	0.749	0.760	0.760	0.760	0.760
	3	0.727	0.775	0.784	**0.786**	**0.786**
	5	0.735	0.778	0.784	0.784	0.781
	7	0.735	0.777	0.782	0.780	0.776
	9	0.735	0.777	0.782	0.780	0.776

AUC of the B-SPMF classifiers is maximized for $R = 3$ and its performance actually decreases if the rank estimate is increased beyond this point. The additional regularization afforded by the low rank restricts the model from being overfitted in the training procedure and thus leads to a higher AUC. If using the kernel parameter value $\theta = 8.0$, the difference in AUC for rank estimate $R = 3$ and $R = 9$ is a considerable 0.025. The AUC of the B-CPMF is maximized for $R = 3$, but for higher ranks the decrease in AUC is not as drastic as for the B-SPMF. In other words, the convolutional model is more resistant to overestimation in the rank parameter. We recall that small changes in the basic disease trajectories weigh disproportionately heavy in the regularization of the CPMF model. As a result these small changes are removed entirely if they do not significantly contribute to the reduction of the discrepancy term. We hypothesize that this reduces the model's susceptibility to overfitting.

It may seem odd that the models achieve an AUC of 0.760 even for a rank estimate $R = 1$, in which case the model deduces only a single basic disease trajectory. However, the model is at liberty to decide U, the females' coefficients. Even though there is only a single trajectory, this trajectory may be scaled differently from female to female. As a result, the females' estimated latent risk profiles may be entirely different.

Let now $\theta = 5.5$ and $R = 5$ be fixed. Figure 6.4 displays the AUC of the B-SPMF classifier as a function of λ_1 and λ_2. The AUC is maximized for $\lambda_1 = 2.15$ and $\lambda_2 = 10000$ and remains high for a wide selection of neighboring parameters. Figure 6.5 similarly displays the AUC of the B-CPMF classifier, which is maximized for $\lambda_1 = 2.15$ and $\lambda_2 = 1468$. Also in this case the AUC

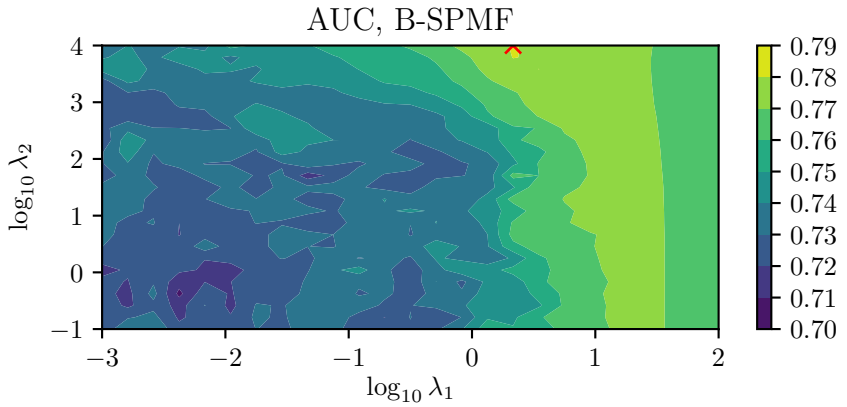


Figure 6.4: AUC of the B-SPMF classifier attained for the Screening Dataset as a function of the hyperparameters λ_1, λ_2. The model is trained with $R = 5$ and predictions are computed using $\theta = 5.5$.

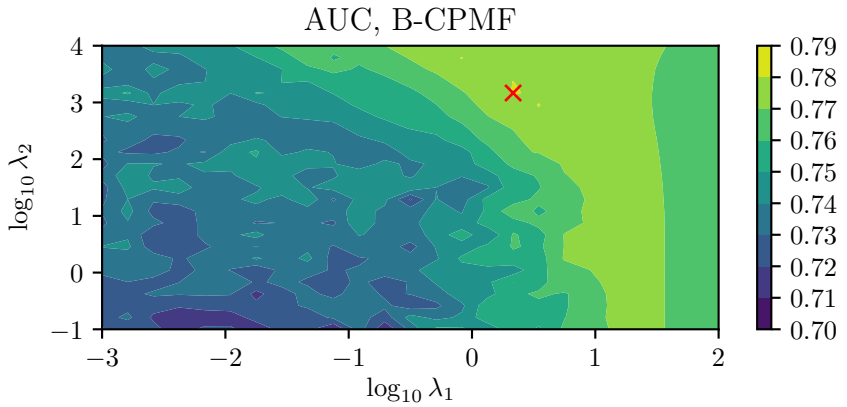


Figure 6.5: AUC of the B-CPMF classifier attained for the Screening Dataset as a function of the hyperparameters λ_1, λ_2. The model is trained with $R = 5$ and predictions are computed using $\theta = 5.5$.

remains high for a wide selection of neighboring parameters. It appears that the level curves of the B-CPMF classifiers are identical but shifted towards smaller values of λ_2 in comparison to the B-SPMF classifier. In other words, the convolutional model requires less temporal regularization to attain its highest AUC. The difference in maximal performance attained by the models is too small to be nuanced.

Chapter 7

Closing Remarks

This is a pioneering work. By implementing matrix factorization in cervical cancer screening data, we derive robust temporal trends in the latent risk of the Norwegian population. We adopt a binarized view of the early prediction of cervical cancer and develop the B-SPMF and B-CPMF classifiers, which we use to successfully predict cervical cancer one year ahead in time. To the best of our knowledge, no similar work has previously been done.

The two matrix factorization classifiers display a similar ability to the binary disease state. However, the latter exhibits a desirable increased resistance to overfitting. If the rank of the dataset is overestimated, predictive performance in the B-CPMF classifier is only slightly decreased. Therefore, proper tuning of the model is a more manageable task.

Throughout this thesis, we have discussed several shortcomings of our model assumptions. Recall that we were not able to match the distribution over the states in the synthetic DGD data with that in the Screening Dataset. Specifically, if we matched the proportions of high-grade and cancer state in DGD data to real data, the proportion of the normal state would no longer match. We argue that the mismatch between the DGD simulation model and the real data testify that the sampled Gaussian kernel is not the best choice of likelihood. At the same time, implementing another likelihood removes the Frobenius-norm in the discrepancy term in (2.28). As a consequence, the LMaFit algorithm cannot be applied. We speculate that it will be difficult to replicate the computational efficiency of the LMaFit when finding the MAP estimate under an alternative likelihood.

We believe increased performance can be achieved by refining the priors assumed on the basic disease trajectories and the coefficient matrix. In an extension of this thesis, total variation (TV) regularization penalty can be used to promote piecewise constant basic disease trajectories. This resonates with the idea that the latent risk of a female will stay constant throughout specific life periods. The escalation in latent risk may result from a sudden change in lifestyle and, as such, best be represented by a piecewise constant function rather than a smooth development. Currently, we allow the latent risk profile

of a female to be a weighting of all the basic disease trajectories. Intuitively we argue that a female should be assigned only to a single or few trajectories to correspond with the concept of phenotype. To induce sparsity in the coefficient matrix, one can replace the Frobenius-norm penalty by the ℓ_1 penalty. This is implemented in PACIFIER but was foregone for practical concerns. If the Frobenius penalty on the coefficient matrix is replaced, the subproblem will need to be solved iteratively. As a consequence, the computation complexity of the algorithm as a whole will increase. We believed it to be critical that the model could be rapidly trained in this initial phase of the project and, therefore, implemented the more straightforward Frobenius-norm penalty.

Finally, a more accurate risk assessment of cervical cancer may require more data. We have seen that making predictions for cervical cancer screening data is incredibly challenging. Even more challenging is to do so using only past screening results. Cervical cancer is associated with several physical risk factors, notably HPV infection [24] and smoking [25, 26]. By factoring in these variables, we can potentially predict Norwegian females' sudden transmission to the high-grade or cancer states even more accurately.

Bibliography

[1] Cancer Registry of Norway. Cancer in norway 2018: Cancer incidence, mortality, survival and prevalence in norway. *Cancer in Norway*, 2019.

[2] S Vaccarella, S Franceschi, G Engholm, S Lönnberg, S Khan, and F Bray. 50 years of screening in the nordic countries: quantifying the effects on cervical cancer incidence. *British Journal Of Cancer*, 111:965 EP –, 07 2014.

[3] IARC Working Group On Evaluation Of Cervical Cancer Screening Programmes. Screening for squamous cervical cancer: Duration of low risk after negative results of cervical cytology and its implication for screening policies. *British Medical Journal (Clinical Research Edition)*, 293(6548):659–664, 1986.

[4] P Sasieni, J Adams, and J Cuzick. Benefit of cervical screening at different ages: evidence from the uk audit of screening histories. *British journal of cancer*, 89(1):88–93, 2003.

[5] Stefan Lönnberg, Ahti Anttila, Tapio Luostarinen, and Pekka Nieminen. Age-specific effectiveness of the finnish cervical cancer screening programme. *Cancer Epidemiology and Prevention Biomarkers*, 21(8):1354–1361, 2012.

[6] Peter Sasieni, Alejandra Castanon, and Jack Cuzick. Effectiveness of cervical screening with age: population based case-control study of prospectively recorded data. *BMJ*, 339, 2009.

[7] SimulaMet. Decipher: Data-driven framework for personalised cancer screening. Unpublished, 2019.

[8] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 135–144, New York, NY, USA, 2014. ACM.

[9] Jérôme Bonnin. Internship report. Unpublished, 2019.

[10] Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, Dec 2012.

[11] Severin Langberg, Mikal Stapnes, Jan Nygård, Mari Nygård, Markus Grasmair, and Valeriya Naumova. Towards a data-driven system for personalised time-dependent risk assessment of cervical cancer. Submitted, 2020.

[12] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[13] Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 713–719, New York, NY, USA, 2005. Association for Computing Machinery.

[14] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009.

[15] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010.

[16] Shiqian Ma, Donald Goldfarb, and Lifeng Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1):321–353, Jun 2011.

[17] J. P. Haldar and D. Hernando. Rank-constrained solutions to linear matrix equations using powerfactorization. *IEEE Signal Processing Letters*, 16(7):584–587, July 2009.

[18] Jared Tanner and Ke Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417 – 429, 2016.

[19] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010.

[20] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

[21] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 931–940, 2008.

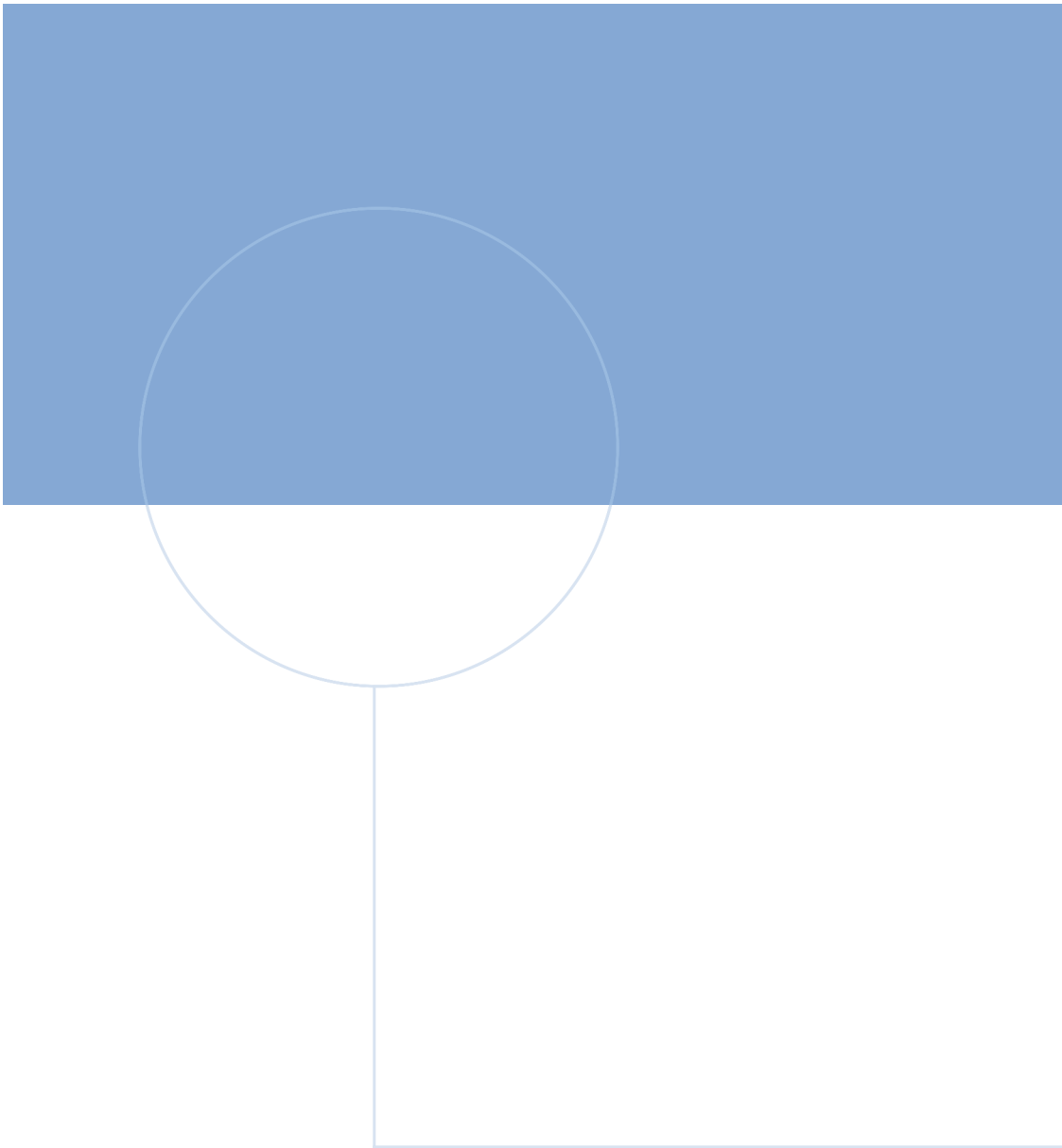
[22] Braden C. Soper, Mari Nygård, Ghaleb Abdulla, Rui Meng, and Jan F. Nygård. A hidden markov model for population-level cervical cancer screening. *Statistics in Medicine*, 2020.

[23] J. Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367 – 374, 2004.

[24] Jan M. M. Walboomers, Marcel V. Jacobs, M. Michele Manos, F. Xavier Bosch, J. Alain Kummer, Keerti V. Shah, Peter J. F. Snijders, Julian Peto, Chris J. L. M. Meijer, and Nubia Muñoz. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *The Journal of Pathology*, 189(1):12–19, 1999.

[25] WARREN WINKELSTEIN JR. Smoking and cervical cancer—current status: a review. *American Journal of Epidemiology*, 131(6):945–957, 1990.

[26] Martyn Plummer, Rolando Herrero, Silvia Franceschi, Chris JLM Meijer, Peter Snijders, F Xavier Bosch, Silvia de Sanjosé, and Nubia Muñoz. Smoking and cervical cancer: pooled analysis of the iarc multi-centric case–control study. *Cancer Causes & Control*, 14(9):805–814, 2003.



NTNU

Norwegian University of
Science and Technology