Factor Analysis and Nuisance Attribute Projection Revisited

Lukáš Machlica, Zbyněk Zajíc

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia,
Pilsen, Czech Republic

machlica@kky.zcu.cz, zzajic@kky.zcu.cz

Abstract

In the paper Factor Analysis (FA) and Nuisance Attribute Projection (NAP) are reviewed, analyzed and compared. Since nowadays FA become a part of most state-of-the-art recognition systems (used e.g. in the concept of i-vectors or PLDA models) it is of relevance to examine different insights into the problem. NAP was chosen as a counterpart to FA as an advanced PCA like method often utilized in speaker recognition systems along with FA. It is demonstrated how can be both FA and NAP expressed as solutions of Least Squares (LS), the consequences of the LS formulation are discussed, and it is shown in what extent do solutions of NAP and FA overlap.

Index Terms: NAP, FA, PCA, least squares

1. Introduction

In last years Factor Analysis (FA) based techniques gained on popularity. Progressive methods as Joint Factor Analysis (JFA) [1], closely related i-vectors [2] or Probabilistic Linear Discriminant Analysis (PLDA) [3] are all based on FA. FA was integrated to the task of speaker recognition when supervectors (SVs), mostly based on GMM parameters, were introduced. Since SVs are of substantially high dimension methods to reduce their dimension and/or bind the parameters in a supervecetor were requested. First methods dealing with dimensionality reduction in the sense of subspace estimation were Nuisance Attribute Projection (NAP) [4] and JFA [5]. While NAP was suited for Support Vector Machines (SVMs) and addressed the channel compensation, JFA was focusing also on the within-speaker variability. Both within and between speaker subspaces were estimated *jointly* at the same time. Lately, in [6] it was shown how to decouple the estimation of both subspaces. At first the between speaker subspace is estimated, and subsequently the within speaker covariance is decomposed and the within speaker subspace is determined. We will focus on the latter case when analyzing FA. In fact the estimation algorithm used in JFA to handle supervectors differs from the standard FA algorithm in that it puts weights on distinct dimensional blocks of supervectors.

The paper aims to review NAP and FA, and show the background of both methods in a different light. At first, NAP will be described in Section 2 and presented as a standard Least Squares (LS) problem. Rather than the JFA algorithm the FA algorithm will be addressed in Section 3 and it will also be related to LS. Finally, both approaches will be compared and the equivalence of NAP and FA solutions will be analyzed in Section 4.

2. Nuisance attribute projection (NAP)

NAP was suited for the concept of SVMs and supervectors [4]. It reduces the influence of the channel variability projecting out the supervector dimensions that are mostly vulnerable to changes of operating conditions. Lots of speakers recorded in several operating conditions have to be collected (several sessions of each speaker have to be available). For each session $h = 1, \dots, H_s$ of each speaker $s = 1, \dots, S$ a D dimensional vector $\boldsymbol{x}_{sh}$ is extracted (e.g. a supervector or an i-vector). Let $\boldsymbol{X}_s = [\boldsymbol{x}_{s1}, \dots, \boldsymbol{x}_{sH_s}]$ be the s^{th} speaker's data matrix with H_s vectors ordered in columns, and let $\boldsymbol{X} = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_S]$ be the overall data matrix.

The objective function to be minimized, introduced in [4], has the form

$$J_{\text{NAP}}(\boldsymbol{P}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} W_{ij} ||\boldsymbol{P}(\boldsymbol{x}_i - \boldsymbol{x}_j)||^2 \qquad (1)$$

where $N = \sum_s H_s$ is the number of input vectors, $\boldsymbol{W} = [W_{ij}]$ is a $N \times N$ symmetric matrix of zeros and ones, $W_{ij} = 1$ if both vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent the same speaker and W_{ij} is zero otherwise, and $\boldsymbol{P}$ is a $D \times D$ projection matrix of low rank D_p. The projection matrix $\boldsymbol{P}$ will be assumed in the form

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{F}_{\perp} \boldsymbol{F}_{\perp}^{\text{T}}, \qquad (2)$$

where columns of the $D \times D_p$ matrix $\boldsymbol{F}_{\perp}$ span the subspace which we are going to project out, and in addition $\boldsymbol{F}_{\perp}^{\text{T}} \boldsymbol{F}_{\perp} = \boldsymbol{I}$. Thus, columns of $\boldsymbol{F}_{\perp}$ are orthonormal, otherwise the projection matrix would have the form $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{F}(\boldsymbol{F}^{\text{T}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\text{T}}$ (assuming that $\boldsymbol{F}$ has full column rank D_p, otherwise the inversion $(\boldsymbol{F}^{\text{T}}\boldsymbol{F})^{-1}$ has to be replaced by a generalized inversion [7]). Note that the objective (1) does not depend on the choice of the base of the subspace, it depends only on the subspace generated by this base. Hence the orthonormal restriction does not violate the generality of the solution of (1).

Equation (1) can be rewritten as

$$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} W_{ij}||\boldsymbol{P}(\boldsymbol{x}_i-\boldsymbol{x}_j)||^2 =$$
$$=\sum_{s=1}^{S}\sum_{i=1}^{N_s-1}\sum_{j=i+1}^{N_s}\boldsymbol{e}_{ij}^{\mathrm{T}}\boldsymbol{X}_s^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}_s\boldsymbol{e}_{ij}$$
$$=\sum_{s=1}^{S}\mathrm{tr}(\boldsymbol{P}\boldsymbol{X}_s(\sum_{i,j}\boldsymbol{e}_{ij}\boldsymbol{e}_{ij}^{\mathrm{T}})\boldsymbol{X}_s^{\mathrm{T}})$$
$$=\mathrm{tr}(\boldsymbol{P}\sum_{s=1}^{S}H_s\boldsymbol{X}_s\boldsymbol{J}_s\boldsymbol{X}_s^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{P}\boldsymbol{X}\boldsymbol{J}\boldsymbol{X}^{\mathrm{T}}),\quad(3)$$

where properties $\boldsymbol{P}^2=\boldsymbol{P}$, $\boldsymbol{P}=\boldsymbol{P}^{\mathrm{T}}$ of the projection matrix (2) were used, $\boldsymbol{e}_{ij}$ is a zero vector with +1 in it's i^{th} entry and -1 in it's j^{th} entry ($\boldsymbol{X}_s\boldsymbol{e}_{ij}=\boldsymbol{x}_{si}-\boldsymbol{x}_{sj}$), and $\boldsymbol{J}_s=\sum_{i,j}\boldsymbol{e}_{ij}\boldsymbol{e}_{ij}^{\mathrm{T}}=\boldsymbol{I}_{H_s}-1/H_s\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}$, $\boldsymbol{I}_{H_s}$ is $H_s\times H_s$ identity matrix, $\boldsymbol{1}$ is a vector of ones, and $\boldsymbol{J}$ is a block diagonal matrix composed of blocks $H_1\boldsymbol{J}_1,\ldots,H_S\boldsymbol{J}_S$. Realizing that $\boldsymbol{X}_s\boldsymbol{J}_s\boldsymbol{X}_s^{\mathrm{T}}=\sum_{h=1}^{H_s}(\boldsymbol{x}_{sh}-\bar{\boldsymbol{x}}_s)(\boldsymbol{x}_{sh}-\bar{\boldsymbol{x}}_s)^{\mathrm{T}}$, $\bar{\boldsymbol{x}}_s=\sum_{h=1}^{H_s}\boldsymbol{x}_{sh}$ is the covariance matrix of vectors in $\boldsymbol{X}_s$, the matrix $\boldsymbol{C}_{\mathrm{W}}=\boldsymbol{X}\boldsymbol{J}\boldsymbol{X}^{\mathrm{T}}$ is in fact the weighted sum of within covariances of each set $\boldsymbol{X}_s$ weighted according to the number of vectors it contains. Thus, the objective function (1) takes the form

$$J_{\text{NAP}}(\boldsymbol{F})=\mathrm{tr}(\boldsymbol{P}\boldsymbol{C}_{\mathrm{W}})=\mathrm{tr}(\boldsymbol{C}_{\mathrm{W}})-\mathrm{tr}(\boldsymbol{C}_F),\quad(4)$$

where $\boldsymbol{C}_F=\boldsymbol{F}_\perp^{\mathrm{T}}\boldsymbol{C}_{\mathrm{W}}\boldsymbol{F}_\perp$ is the within covariance after projecting each $\boldsymbol{x}_{ij}$ onto the column-space of $\boldsymbol{F}_\perp$. The criterion (4) is minimized when columns of $\boldsymbol{F}_\perp$ are eigenvectors of $\boldsymbol{C}_{\mathrm{W}}$ corresponding to highest eigenvalues – the highest within variance (most vulnerable to changes) is projected out.

2.1. NAP and least squares

To simplify following formulas let us assume that $H=H_1=\ldots=H_S$ and that the mean value $\bar{\boldsymbol{x}}_s$ was already subtracted from each vector in $\boldsymbol{X}_s, s=1,\ldots,S$. Now

$$1/H\ J_{\text{NAP}}=\mathrm{tr}\left(\boldsymbol{P}\boldsymbol{C}_W\right)=\mathrm{tr}\left((\boldsymbol{I}-\boldsymbol{F}_\perp\boldsymbol{F}_\perp^{\mathrm{T}})\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)$$
$$=\sum_i\mathrm{tr}\left(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{F}_\perp\boldsymbol{F}_\perp^{\mathrm{T}})\boldsymbol{x}_i\right)$$
$$=\sum_i||(\boldsymbol{I}-\boldsymbol{F}_\perp\boldsymbol{F}_\perp^{\mathrm{T}})\boldsymbol{x}_i||^2=\sum_i||\boldsymbol{x}_i-\boldsymbol{F}_\perp\boldsymbol{F}_\perp^{\mathrm{T}}\boldsymbol{x}_i||^2$$
$$=\sum_i||\boldsymbol{x}_i-\boldsymbol{F}_\perp\boldsymbol{z}_i||^2,\text{ and }\boldsymbol{z}_i=\boldsymbol{F}_\perp^{\mathrm{T}}\boldsymbol{x}_i,\quad(5)$$

hence $\boldsymbol{F}_\perp\boldsymbol{z}_i$ is an orthogonal projection of $\boldsymbol{x}_i$ onto the column space of $\boldsymbol{F}_\perp$. Thus, J_{NAP} can be solved also iteratively, an actual estimate of $\boldsymbol{F}_\perp$ is used to get projections $\boldsymbol{z}_i$ and subsequently a Least Squares (LS) problem is solved to get a new estimate of matrix $\boldsymbol{F}_\perp$. The iterative procedure does not guarantee the orthogonality of columns of $\boldsymbol{F}_\perp$, but we can use e.g. QR decomposition after each iteration to make columns of $\boldsymbol{F}_\perp$ orthogonal and to increase the robustness of the LS estimation algorithm.

3. Factor analysis (FA)

FA is a latent linear Gaussian model of the form [8]

$$\boldsymbol{v}_i=\boldsymbol{F}\boldsymbol{y}_i+\boldsymbol{\epsilon}_i,\quad(6)$$

where $\boldsymbol{V}=[\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N]$ is a matrix of input vectors of dimension D which were normalized to zero mean beforehand. In FA an assumption is made that each $\boldsymbol{v}_i$ may be explained by some lower D_y dimensional latent representation $\boldsymbol{y}_i$ ($D_y<D$), which lies in a subspace spanned by columns of $\boldsymbol{F}$. The $D\times D_y$ matrix $\boldsymbol{F}$ is also called the *factor loading matrix*, entries of $\boldsymbol{y}_i$ are denoted as *factors*, and $\boldsymbol{\epsilon}_i$ is some residual noise following Gaussian distribution $\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$ with zero mean and diagonal covariance $\boldsymbol{\Sigma}$. Since both $\boldsymbol{F}$, $\boldsymbol{y}_i$ are unknown also the distribution for $\boldsymbol{y}_i$ has to be specified – it is given as a standard Gaussian distribution $\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. Hence, it is obvious that both

$$p(\boldsymbol{v}_i)\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}+\boldsymbol{\Sigma})\quad(7)$$

and $p(\boldsymbol{v}_i|\boldsymbol{y}_i)\sim\mathcal{N}(\boldsymbol{F}\boldsymbol{y}_i,\boldsymbol{\Sigma})$ follow Gaussian distibution.

The model is found utilizing Expectation Maximization (EM) algorithm in order to maximize (7). In the E-step mean and covariance of each latent variable $\boldsymbol{y}_i$ given $\boldsymbol{v}_i$ are evaluated using old values of $\boldsymbol{F},\boldsymbol{\Sigma}$:

$$\boldsymbol{\Psi}=(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}+\boldsymbol{I})^{-1},\quad(8)$$
$$\mathrm{E}[\boldsymbol{y}_i]=\boldsymbol{\Psi}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{v}_i,\quad(9)$$

$p(\boldsymbol{y}_i|\boldsymbol{v}_i)\sim\mathcal{N}(\mathrm{E}[\boldsymbol{y}_i],\boldsymbol{\Psi})$, and subsequently $\boldsymbol{F}$ and $\boldsymbol{\Sigma}$ are updated in the M-step:

$$\boldsymbol{F}=\left(\sum_{i=1}^{N}\boldsymbol{v}_i\mathrm{E}[\boldsymbol{y}_i^{\mathrm{T}}]\right)\left(\sum_{i=1}^{N}\mathrm{E}[\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}]\right)^{-1},\quad(10)$$

$$\boldsymbol{\Sigma}=\boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}^{\mathrm{T}}+\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{v}_i-\hat{\boldsymbol{v}}_i)(\boldsymbol{v}_i-\hat{\boldsymbol{v}}_i)^{\mathrm{T}},\quad(11)$$

where $\hat{\boldsymbol{v}}_i=\boldsymbol{F}\mathrm{E}[\boldsymbol{y}_i]\approx\boldsymbol{F}\boldsymbol{y}_i$ is the reconstructed vector $\boldsymbol{v}_i$, and $\mathrm{E}[\boldsymbol{y}_i]$ is in fact the MAP estimate of $\boldsymbol{y}_i$. Note that $\mathrm{E}[\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}]$ showing up in (10) is obtained realizing that $\boldsymbol{\Psi}=\mathrm{E}[\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}]-\mathrm{E}[\boldsymbol{y}_i]\mathrm{E}[\boldsymbol{y}_i^{\mathrm{T}}]$. From (11) it is easy to see that the noise variance captures the residual variance and in addition it grows also in regions where the covariance of latent variables given observed $\boldsymbol{v}_i$ is high.

An interesting property is that if the noise covariance $\boldsymbol{\Sigma}$ is fixed (e.g. it is derived a-priori, and it is not updated in each iteration), the FA training algorithm described in this section leads to the same result as the FA training utilizing the dataset $\tilde{\boldsymbol{V}} = [\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}_1, \ldots, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}_N]$ with

$$\tilde{\boldsymbol{\Psi}} = (\tilde{\boldsymbol{F}}^{\mathrm{T}}\tilde{\boldsymbol{F}} + \boldsymbol{I})^{-1}, \ \mathrm{E}[\tilde{\boldsymbol{y}}_i] = \tilde{\boldsymbol{\Psi}}\tilde{\boldsymbol{F}}^{\mathrm{T}}\tilde{\boldsymbol{v}}_i, \qquad (12)$$

yielding (up to some arbitrary rotation matrix $\boldsymbol{R}$ due to Gaussionality assumptions) $\boldsymbol{FR} = \boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{F}}$, $\tilde{\boldsymbol{\Psi}} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{R}$, and $\tilde{\mathrm{E}}[\boldsymbol{y}_i] = \boldsymbol{R}^{\mathrm{T}}\mathrm{E}[\boldsymbol{y}_i]$. It is easy to see that

$$\begin{aligned}&\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}_i = \tilde{\boldsymbol{v}}_i = \tilde{\boldsymbol{F}}\mathrm{E}[\tilde{\boldsymbol{y}}_i] = \\ &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{FRR}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{RR}^{\mathrm{T}}\mathrm{E}[\boldsymbol{y}_i] = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{F}\mathrm{E}[\boldsymbol{y}_i]. \qquad (13)\end{aligned}$$

Such an observation may bring some additional computational savings when implementing a FA system. The input dataset is normalized beforehand so that all the multiplications with $\boldsymbol{\Sigma}^{-1}$ in the FA update formulas will vanish.

3.1. FA and least squares

It is straightforward to show that minimizing the objective function

$$J_{\mathrm{FA}}(\boldsymbol{F}) = \frac{1}{2}\sum_{i=1}^{N}||\boldsymbol{v}_i - \boldsymbol{F}\mathrm{E}[\boldsymbol{y}_i]||^2 + \frac{N}{2}\mathrm{tr}(\boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}^{\mathrm{T}}), \quad (14)$$

having $\mathrm{E}[\boldsymbol{y}_i]$ and $\boldsymbol{\Psi}$ fixed leads to the update formula identical to (10). The formulation is known as the problem of regularized LS [9] and it brings a new insight into the concept of FA. The iterative estimation of $\boldsymbol{F}$ consists of two steps. At first actual $\boldsymbol{F}$ is used to get the mean and the covariance of latent variables $\boldsymbol{y}_i$ given $\boldsymbol{v}_i$, and subsequently new $\boldsymbol{F}$ is found solving the problem of regularized LS (14). Note that the noise covariance $\boldsymbol{\Sigma}$ appears only when evaluating $\mathrm{E}[\boldsymbol{y}_i]$ and $\boldsymbol{\Psi}$, thus it alters only the latent variables. Since (14) has to be minimized the second term in (14) – the *regularization term* – is used to push the directions in $\boldsymbol{F}$, in which the covariance of latent variables is high, toward zero. Thus, the rank of the matrix $\boldsymbol{F}$ can *decrease* below D_y (assuming that the rank of the sample covariance matrix is at least D_y).

4. FA and NAP

In previous sections it was shown how both NAP and FA may be solved in terms of Least Squares (LS). Now we are going to look on similarities and differences between both approaches.

We will come out of conclusions made in [9], where it was shown that generative model (6) with isotropic noise covariance $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, which maximizes the likelihood (7) of input data, is given by the eigenvalue decomposition of the data covariance matrix. However we will use different approach enabling more insight into the problematic. Focusing on isotropic noise covariance formulas (8) and (9) change to

$$\begin{aligned}\boldsymbol{\Psi} &= \sigma^2(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} + \sigma^2\boldsymbol{I})^{-1}, \\ \mathrm{E}[\boldsymbol{y}_i] &= (\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{v}_i. \qquad (15)\end{aligned}$$

Let us make a small diversion and let us adjust the form of (14) utilizing the isotropic noise covariance. Let us decompose $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{DQ}$ by SVD decomposition so that $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{QQ}^{\mathrm{T}} = \boldsymbol{I}$, and since $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}$ is positive semi-definite $\boldsymbol{D} = [d_{ii}]$ is a diagonal matrix with $d_{ii} \geq 0$. For $\boldsymbol{F}_\perp = \boldsymbol{FQ}^{\mathrm{T}}\boldsymbol{D}^{-1/2}$ we get $\boldsymbol{F}_\perp^{\mathrm{T}}\boldsymbol{F}_\perp = \boldsymbol{I}$ (columns of $\boldsymbol{F}_\perp$ are orthonormal). For future use let

$$\boldsymbol{K}_1 = \left[\frac{d_{ii}^2 + 2d_{ii}\sigma^2}{(d_{ii}+\sigma^2)^2}\right], \ \boldsymbol{K}_2 = \left[\frac{d_{ii}\sigma^2}{d_{ii}+\sigma^2}\right] \qquad (16)$$

be diagonal positive semi-definite $D_y \times D_y$ matrices. At first, let us focus on the second term in (14). Substituting for $\boldsymbol{\Psi}$ we get

$$\begin{aligned}\mathrm{tr}(\boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}^{\mathrm{T}}) &= \mathrm{tr}(\sigma^2(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{DQ} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{DQ}) \\ &= \mathrm{tr}(\sigma^2(\boldsymbol{D} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{D}) = \mathrm{tr}(\boldsymbol{K}_2). \quad (17)\end{aligned}$$

Thus, the term $\mathrm{tr}(\boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}^{\mathrm{T}})$ influences only the *scaling* of directions in the latent space. Before the rearrangement of the first term in (14) note that

$$(\boldsymbol{I} - \boldsymbol{F}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{F}^{\mathrm{T}})^2 = \boldsymbol{I} - \boldsymbol{F}_\perp\boldsymbol{K}_1\boldsymbol{F}_\perp^{\mathrm{T}}. \quad (18)$$

Back to the first term in (14), substituting for $\mathrm{E}[\boldsymbol{y}_i]$:

$$\begin{aligned}&\sum_i ||\boldsymbol{v}_i - \boldsymbol{F}\mathrm{E}[\boldsymbol{y}_i]||^2 = \\ &= \sum_i \mathrm{tr}\left(\boldsymbol{v}_i(\boldsymbol{I} - \boldsymbol{F}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{F}^{\mathrm{T}})^2\boldsymbol{v}_i^{\mathrm{T}}\right) \\ &= \mathrm{tr}(\sum_i \boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}) - \mathrm{tr}(\boldsymbol{K}_1\sum_i \boldsymbol{F}_\perp^{\mathrm{T}}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{F}_\perp) \\ &= N\mathrm{tr}(\boldsymbol{C}_V) - N\mathrm{tr}(\boldsymbol{K}_1\boldsymbol{C}_F), \qquad (19)\end{aligned}$$

where $\boldsymbol{C}_V = 1/N\sum_i \boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}$, $\boldsymbol{C}_F = \boldsymbol{F}_\perp^{\mathrm{T}}\boldsymbol{C}_V\boldsymbol{F}_\perp$ is the covariance of the projected dataset $\boldsymbol{V}$. And finally, combining both we get

$$\frac{2}{N}J_{\mathrm{FA}}(\boldsymbol{F}_\perp, \boldsymbol{D}) = \mathrm{tr}(\boldsymbol{C}_V) - \mathrm{tr}(\boldsymbol{K}_1\boldsymbol{C}_F - \boldsymbol{K}_2). \quad (20)$$

Note that $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ depend on the diagonal matrix $\boldsymbol{D}$, and $\boldsymbol{C}_F$ depends on $\boldsymbol{F}_\perp$. Examining (20), Figure 1 and Figure 2 we can make conclusions on the role of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$. At first, note that the diagonal elements of $\boldsymbol{K}_1$ are lower and upper bounded by 0 and 1, respectively, whereas diagonal elements of $\boldsymbol{K}_2$ are only lower bounded by 0 (recall that $d_{ii} \geq 0$, $\sigma^2 \geq 0$). If $\sigma^2 >> d_{ii}$ than the corresponding directions do not contribute to minimize J_{FA}, and the task of $\boldsymbol{K}_2$ is to completely eliminate these directions.

Since $\boldsymbol{K}_1$, $\boldsymbol{K}_2$ perform only scaling of directions, in order to minimize (20) at first $\mathrm{tr}(\boldsymbol{C}_F)$ has to be maximized. This is done when columns of $\boldsymbol{F}_\perp$ are formed by eigenvectors of $\boldsymbol{C}_V$ corresponding to highest eigenvalues, see Section 2. A useful side effect is that $\boldsymbol{C}_F$ becomes diagonal with D_y highest eigenvalues λ_i of

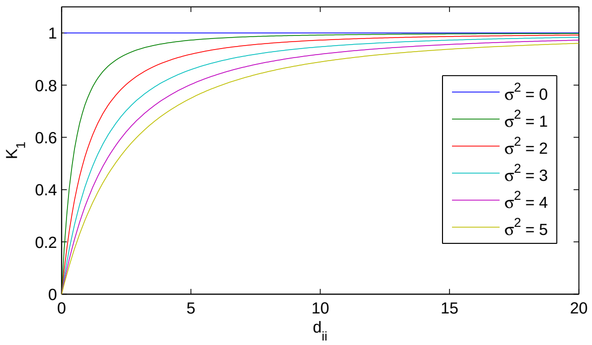

Figure 1: *The dependency of diagonal entries of $\boldsymbol{K}_1$ on different values of d_{ii} and σ^2.*

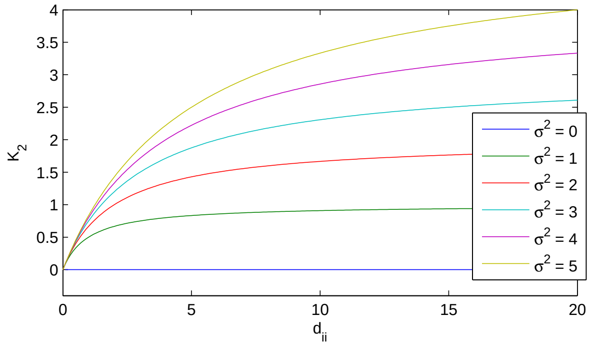

Figure 2: *The dependency of diagonal entries of $\boldsymbol{K}_2$ on different values of d_{ii} and σ^2.*

$\boldsymbol{C}_V$ on its diagonal. To find $\boldsymbol{D}$ one has to maximize tr$(\boldsymbol{K}_1\boldsymbol{C}_F - \boldsymbol{K}_2)$:

$$\frac{\partial}{\partial d_{ii}} \sum_{i=1}^{D_y} \frac{d_{ii}^2 + 2d_{ii}\sigma^2}{(d_{ii}+\sigma^2)^2}\lambda_i - \frac{d_{ii}\sigma^2}{d_{ii}+\sigma^2} = 0. \qquad (21)$$

After taking the derivative we get $d_{ii} = 2\lambda_i - \sigma^2$, $\boldsymbol{D}$ is given by eigenvalues of $\boldsymbol{C}_V$. Since $d_{ii} \geq 0$ condition $d_{ii} = 0$ if $\lambda_i \leq \sigma^2/2$ has to be introduced, which is in accordance with previous discussion on the role of $\boldsymbol{K}_2$.

Now let us replace the general dataset $\boldsymbol{V}$ with the dataset $\boldsymbol{X}$ from Section 2, where vectors $\boldsymbol{x}_{sh}$ were already normalized to zero mean, and the covariance matrix $\boldsymbol{C}_V$ with the within covariance matrix $1/N\boldsymbol{C}_\text{W}$ utilized in NAP. This means that the latent variables $\boldsymbol{y}_i$ will now describe the channel/session space. Recall that in the case of NAP the solution is also given by the eigenvalue decomposition of $\boldsymbol{C}_\text{W}$, thus if the noise model in FA is isotropic the solutions (more precisely the estimated subspaces) for NAP and FA become identical. However, criteria J_NAP and J_FA will still differ in some extent ($\boldsymbol{K}_1$ and $\boldsymbol{K}_2$). Recall that in the case of NAP we had (rewriting (4)):

$$\frac{1}{N} J_\text{NAP}(\boldsymbol{F}) = \text{tr}(\boldsymbol{C}_\text{W}) - \text{tr}(\boldsymbol{C}_F). \qquad (22)$$

If $\sigma^2 = 0$ than $\boldsymbol{K}_1 = \boldsymbol{I}$, $\boldsymbol{K}_2 = \boldsymbol{0}$ and both criteria become equivalent. The same is true if we put an orthonormal restriction on columns of $\boldsymbol{F}$, hence $\boldsymbol{F}^\text{T}\boldsymbol{F} = \boldsymbol{Q}^\text{T}\boldsymbol{I}\boldsymbol{Q}$ and $d_{ii} = 1$. Now, both $K_1 = (1+2\sigma^2)/(1+\sigma^2)^2$ and $K_2 = \sigma^2/(1+2\sigma^2)$ become constants independent on the choice of $\boldsymbol{F}$, and $J_\text{FA} = \alpha_1 J_\text{NAP} + \alpha_2$ becomes a scaled version of J_NAP for some constants α_1, α_2. Otherwise, the FA criterion does incorporate also the influence of noise, thus the value of the criterion differs from J_NAP even if the resulting subspaces are identical.

It should be stated that the previous discussion can be used also when comparing Principal Component Analysis (PCA) and FA. The only difference between NAP and PCA is that PCA takes any data covariance matrix and performs the eigenvalue decomposition, whereas NAP requires within class covariance matrix.

To get an idea what is going on when the noise covariance $\boldsymbol{\Sigma}$ is diagonal we can turn to (12). Hence, at first the input data are rescaled according to the given covariance matrix $\boldsymbol{\Sigma}$ (the feature space is adjusted to promote dimensions with low amount of noise), and subsequently the previous discussion can be followed assuming $\sigma^2 = 1$.

5. Conclusions

Factor Analysis and Nuisance Attribute Projection were analyzed and compared. We have shown in the light of least squares when NAP and FA criteria as well as their solutions (generated subspaces) equal, and how does FA in addition treat noise.

6. Acknowledgements

This work was supported by the Grant Agency of the Czech Republic under Project GAČR P103/12/G084.

7. References

[1] Kenny, P., “Joint factor analysis of speaker and session variability: theory and algorithms”, Tech. report, Centre de Recherche Informatique de Montral (CRIM), 2006

[2] Dehak, N., Kenny, P., Dehak, R., et al., “Front-end factor analysis for speaker verification”, IEEE Transactions on Audio, Speech and Language Processing, 2010

[3] Prince, S. and Elder, J., “Probabilistic linear discriminant analysis for inferences about identity”, IEEE 11th International Conference on Computer Vision, 1-8, 2007.

[4] Solomonoff A., Quillen, C. and Campbell, W., “Channel compensation for SVM speaker recognition”, Odyssey, 219–226, 2004.

[5] Kenny, P. and Dumouchel, P., “Disentangling speaker and channel effects in speaker verification”, ICASSP, 37–40, 2004.

[6] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel, P., “Factor Analysis Simplified”, ICASSP, 637–640, 2005.

[7] Harville, D. A., “Matrix algebra from a statistician’s perspective”, Springer, 1st edition, 1997.

[8] Comon, P. and Jutten, Ch., “Handbook of blind source separation”, Academic Press, 1st edition, 2010.

[9] Bishop, Ch. M., “Pattern Recognition and Machine Learning”, Springer, 2007.