CONJUGATE-PRIOR-REGULARIZED MULTINOMIAL PLSA FOR COLLABORATIVE FILTERING

Marcus Klasson, Stefan Ingi Adalbjörnsson*, Johan Swärd**, Søren Vang Andersen**

*Dept. of Mathematics, Faculty of Engineering, Lund University, Sweden
**Dept. of Mathematical Statistics, Lund University, Sweden
emails: klasson.marcus@gmail.com, {sia, js, sva}@maths.lth.se

ABSTRACT

We consider the over-fitting problem for multinomial *probabilistic Latent Semantic Analysis* (pLSA) in collaborative filtering, using a regularization approach. For big data applications, the computational complexity is at a premium and we, therefore, consider a *maximum a posteriori* approach based on conjugate priors that ensure that complexity of each step remains the same as compared to the un-regularized method. In the numerical section, we show that the proposed regularization method and training scheme yields an improvement on commonly used data sets, as compared to previously proposed heuristics.

***Index Terms*—** Recommender systems, collaborative filtering, conjugate prior regularization, probabilistic latent semantic analysis.

1. INTRODUCTION

Personalized recommender systems are commonplace in online marketplaces and media services, providing a useful tool for users to navigate the myriad of available items or media, and similarly useful for the vendors for increasing revenue [1]. Herein, we consider *Collaborative Filtering* (CF) [2–11], which is an approach within recommender systems, where the recommendations are personalized such that items are recommended for each consumer based on the preferences of other consumers that have similar historical consumption patterns [3, 12]. Alternatively, one may use content based methods, where the items meta-data are used to group items, so that items that are similar to previously consumed items can be recommended. However, when the prior consumption records for many users are available, it is well established that the CF based methods offer superior performance [13], with the best recent results coming from matrix factorization based methods [2–11]. This paper considers a probabilistic approach termed the *probabilistic Latent Semantic Analysis* (pLSA), which has many applications in information retrieval and filtering, and is an important tool for machine learning in text [3, 12]. In the recommendation setting, pLSA has been used for clustering users together based on their item preferences and similar tastes in order to make accurate rating predictions on items they have not yet consumed [3]. The model is exposed to user observations in order to learn which cluster the users belong to, as well as the distribution of the ratings for each movie in the clusters. For continuous rating data, the Gaussian emission pLSA model yields a low rank matrix that may be used to give predictions on all unconsumed items for all users; a similar prediction can be made for the multinomial case. Herein, we will consider the multinomial version of the model, which is applicable to a wide range of data, e.g., binary, such as when only a consumption is registered using a like or a dislike, or purely categorical, such as when a vote with an emoji is given on Facebook. The three main challenges with multinomial pLSA in CF are (i) the over-fitting problem, which results in less reliable model parameters and increases the prediction errors for unseen data [3, 12, 14], (ii) the sparsity of the data, i.e., most users only consume a small percentage of the available items and even fewer ratings, and (iii) the computational cost of the estimation and training of the system. In this work, we generalize the model presented in [3] by introducing a conjugate prior on the model parameters, to mitigate the over-fitting and to better handle the sparsity in the data. Similar approaches have been suggested for the Gaussian mixture model [15, 16] and the Gaussian emission pLSA model [17], which we herein generalize to the categorical pLSA model. The resulting model has the same complexity as the non-regularized version and can be seen as a simplified version of the model proposed in [14]. By adding bias to the model parameter values through the conjugate prior, users or items with few observations are less prone to be over-fitted, thus yielding better fit to unseen data. Furthermore, we show that, the pragmatic choice of performing *maximum a posteriori* (MAP) estimation with conjugate priors results in an *expectation maximization* (EM) algorithm that avoids the otherwise typical Markov Chain Monte Carlo or approximate inference [14]. We also compare the proposed regularized MAP pLSA to the multinomial pLSA model with early stopping presented in [3, 18], which is a commonly used heuristic in machine learning to counter act over-fitting.

ISBN 978-0-9928626-7-1

2. MODEL DEFINITIONS

We use the following notation: define a set of users $\mathcal{U} = \{u_1, \ldots, u_m\}$ and items $\mathcal{I} = \{i_1, \ldots, i_n\}$. The users have the opportunity to rate items with a preference value from an explicit rating scale $\mathcal{R}$, where the given rating data $r_{u,i}$ is stored in an $m \times n$ matrix $\mathbf{R}$. In real-world applications, $\mathbf{R}$ is sparse, as users tend to only rate a small number of items [3, 12, 13]. For the latent states, we use $z \in \{z_1, \ldots, z_K\}$, where K denotes the number of possible states. The main idea with pLSA is to introduce latent states $z \in \{z_1, \ldots, z_K\}$. Unlike probabilistic user-clustering models wherein each user is associated with a single latent state, every single observation $\langle u, i, r\rangle$ is in pLSA connected to a latent state. Based on the multinomial pLSA model for discrete data presented in [3], a conditional probability for a user u is associated to state z, $P(z|u)$, and another conditional probability for the ratings given item i and state z, $P(r|i, z)$. Additionally, u is assumed to be, conditionally on z, independent of i, which leads to to the following mixture model for a single rating

$$P(r|u, i; \theta) = \sum_z P(r|i, z)P(z|u) \tag{1}$$

where θ is the unknown parameter vector; there are $K|\mathcal{U}| + K|\mathcal{I}||\mathcal{R}|$ multinomial probabilities in $\theta = \{P(r|i, z), P(z|u)\}$. The summation over z means summing over all possible latent states. Predicting the missing rating data, one may use the expected value of a rating given by [3]

$$\begin{aligned} \hat{r}_{u,i} = \mathrm{E}[r|u, i] &= \sum_{r\in\mathcal{R}} rP(r|u, i) \\ &= \sum_{r\in\mathcal{R}} r \sum_z P(r|i, z)P(z|u) \end{aligned} \tag{2}$$

Thus, the model may be used to offer suggestion for new items to be consumed on an individual basis, as the probability of an observation coming from a specific cluster is user specific, whereas the emission probabilities are item and cluster specific. The parameter learning optimization procedure is performed by optimizing the unknown parameters with the EM algorithm. Using EM for training model parameters is very common in latent variable models, such as pLSA, since the latent states z are unobservable and embedded in complicated manner in the log-likelihood function [3, 19].The data tends to be sparse, having comparably as many values as the number of parameters, and unbalanced, i.e., some users or items could be associated with only a few observations; this may lead to unreliable parameter estimates [12]. This means that users and items could be trapped in certain consumption or rating patterns, thus resulting in inaccurate predictions on unseen data. To mitigate this over-fitting, we here propose using a regularization method in order to balance and control the parameter values.

		Proposed	ES pLSA	Pop
RMSE	mean	1.2375	1.2727	1.3712
	std	0.0064	0.0070	0.0062
MAE	mean	0.9711	0.9834	1.0908
	std	0.0047	0.0052	0.0048

Table 1. The mean and standard deviation of the prediction error, resulting from the EachMovie data set.

3. MAP ESTIMATION WITH CONJUGATE PRIORS

As the conjugate prior of a multinomial distribution is the Dirichlet density, the posterior distribution when using such a prior will be Dirichlet distributed as well [15, 16]. The prior distribution may thus be expressed by a Dirichlet distribution of order $m \geq 2$, with parameters $\mathbf{w} = (w_1, \ldots, w_m)$, where $w_i \geq 0$, with $\sum_{i=1}^m w_i = 1$ [19]. The resulting probability density function may be expressed as

$$\mathrm{Dir}(\{\mathbf{w}\}|\{\gamma\}) = \frac{\Gamma\left(\sum_{i=1}^m \gamma_i\right)}{\prod_{i=1}^m \Gamma(\gamma_i)} \prod_{i=1}^m w_i^{\gamma_i - 1} \tag{3}$$

where $\gamma = (\gamma_1, \ldots, \gamma_m)$ are the hyperparameters and $\Gamma(\cdot)$ denotes the Gamma function[1] [18]. As a result, the prior distribution of the parameters θ may be expressed as

$$\begin{aligned} P(\theta) &= \mathrm{Dir}(\{P(r|i, z)\}|\{\gamma_{i,r,z}\}) \cdot \mathrm{Dir}(\{P(z|u)\}|\{\gamma_{u,z}\}) \\ &\propto \prod_z \left[\prod_{i,r} P(r|i, z)^{\gamma_{i,r,z}-1} \prod_u P(z|u)^{\gamma_{u,z}-1} \right] \end{aligned} \tag{4}$$

where $\varphi = \{\gamma_{i,r,z}, \gamma_{u,z}\}$ are the hyperparameters of the conjugate prior distributions. These hyperparameters may be selected in different ways; for instance, if $P(z|u)$ is considered, they may be selected as

$$\gamma_{u,z} = n_{\{P(z|u)\}} + 1 \tag{5}$$

where $n_{\{P(z|u)\}}$ can then be interpreted as an additional set of artificial data points to regularize the estimates of $P(z|u)$.

In [3], the EM algorithm is used for minimizing the negative log-likelihood

$$-\ell(\theta; \mathcal{D}) = -P(\mathcal{D}|\theta) = -\sum_{\mathcal{D}} \log P(r|u, i; \theta) \tag{6}$$

where the summation is over all observed triplets, $\mathcal{D} = \langle u, i, r\rangle$. The EM procedure introduces *variational probability distributions* $Q(z|u, i, r; \theta)$ to equation (6), which models the probability for a latent state z to be associated with a certain observation. Therefore, the EM algorithm will allow Jensen's inequality to be used to form an upper bound on the

[1]The Gamma function is defined as $\Gamma(x) \equiv \int_0^\infty t^{x-1}e^{-t}\, dt$.

ISBN 978-0-9928626-7-1 © EURASIP 2017

		Proposed	ES pLSA	Pop
RMSE	mean	0.9193	0.9781	0.9847
	std	0.0076	0.0176	0.0090
MAE	mean	0.7241	0.7807	0.7870
	std	0.0049	0.0184	0.0057

Table 2. The mean and standard deviation of the prediction error, resulting from the MovieLens data set.

likelihood, which then is to be minimized in order to obtain the parameter estimates θ.

We can similarly extend the EM algorithm with the MAP approach by considering the negative log-posterior, or the sum of the likelihood function in (6) and the prior distribution in (4), such that the optimal regularized parameter estimates are given by

$$\begin{aligned}\theta^*_{MAP} &= \arg\min_{\theta} -\log P(\theta|\mathcal{D}) \\ &= \arg\min_{\theta} -\{\ell(\theta;\mathcal{D}) + \log P(\theta)\}\end{aligned} \tag{7}$$

where

$$\begin{aligned}\ell(\theta;\mathcal{D}) + \log P(\theta) = \sum_{\mathcal{D}}\sum_{z} Q(z|\mathcal{D};\theta)\log\frac{P(r|i,z)P(z|u)}{Q(z|\mathcal{D};\theta)} \\ +\log \text{Dir}(\{P(r|i,z)\}|\{\gamma_{i,r,z}\}) + \log \text{Dir}(\{P(z|u)\}|\{\gamma_{u,z}\})\end{aligned} \tag{8}$$

To summarize, the MAP-based EM algorithm alternates between the expectation and maximization step until the parameter values have converged. In the E-step, the optimal variational probabilities, denoted by Q^*, are estimated as [3, 15, 16]

$$Q^*(z|u,i,r;\theta) = \frac{P(r|i,z)P(z|u)}{\sum z'(P(r|i,z')P(z'|u)} \tag{9}$$

Thereafter, in the M-step, one obtains the new parameter estimates from equation (8) using the computed Q^*-distributions. The regularized parameter estimates are given by

$$P(z|u) = \frac{\sum_{\langle u',i,r\rangle:u'=u} Q^*(z|u,i,r;\theta) + (\gamma_{u,z}-1)}{\sum_{z'}\sum_{\langle u',i,r\rangle:u'=u} Q^*(z'|u,i,r;\theta) + (\gamma_{u,z'}-1)} \tag{10a}$$

$$P(r|i,z) = \frac{\sum_{\langle u,i',r'\rangle:i'=i,\, r'=r} Q^*(z|u,i,r;\theta) + (\gamma_{i,r,z}-1)}{\sum_{\langle u,i',r\rangle:i'=i} Q^*(z|u,i,r;\theta) + (\gamma_{i,r,z}-1)} \tag{10b}$$

where the prime signs under the summations denote a fixed variable for the computed conditional probability. It should be noted that this method reduces to the EM-algorithm presented in [3] in the case when all hyperparameters are set to 1.

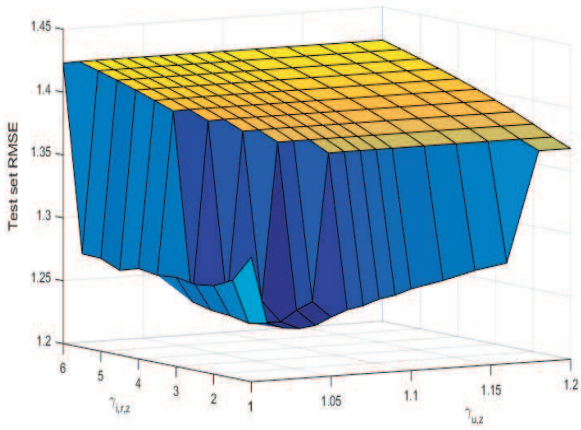

Fig. 1. Grid search over RMSE with respect to $\gamma_{u,z}$ and $\gamma_{i,r,z}$ for $K = 200$.

4. EXPERIMENTAL RESULTS

To investigate the proposed conjugate-prior-regularized MAP pLSAs capability to mitigate overfitting, we compare it to both pLSA using the standard EM with an early stopping (ES) condition and to the so-called Pop item-average estimator [13]. This evaluation is done using both the EachMovie data set, which contains $2,811,718$ ratings, entered by $61,265$ users, of 1623 items (movies) [20], and the MovieLens 1M dataset [21], which consists of $1,000,209$ ratings entered by $6,040$ users, rating $3,900$ items (movies). The data is randomly divided into three sets: training, validation, and test data. To reduce the variance, this subdivision was repeated ten times on the original data set. Users that have less than three observations and items which have not been rated are removed from the data sets. Partitioning the data into the different sets was performed with the leave-one-out algorithm, such that one observation from every user is randomly picked and left out from the training set. This procedure was executed twice to obtain a validation and a test set. To evaluate the models, the training procedures of the regularized pLSA model was terminated when the difference in log-likelihood between the two latest EM iterations was smaller than 10^{-3}. The ES pLSA was terminated, as in [3], before the prediction error of the validation set had increased, and then performed one last EM training step including the training plus validation data. When the final parameters were updated, the prediction errors are computed for the test set, using

$$\text{RMSE} = \sqrt{\frac{1}{|\mathbf{R}|}\sum_{\langle u,i\rangle\in\mathbf{R}}(r_{u,i}-\hat{r}_{u,i})^2}$$

$$\text{MAE} = \frac{1}{|\mathbf{R}|}\sum_{\langle u,i\rangle\in\mathbf{R}}|r_{u,i}-\hat{r}_{u,i}|$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

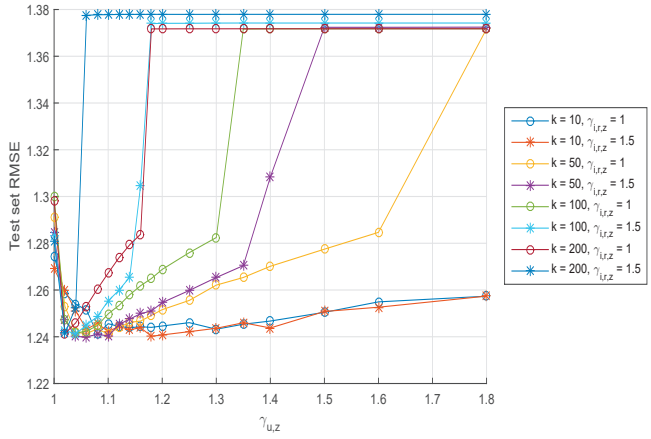

Fig. 2. Average RMSE on three different test sets with respect to $\gamma_{u,z}$, different state sizes K and $\gamma_{i,r,z} = 1.0$ and 1.5.

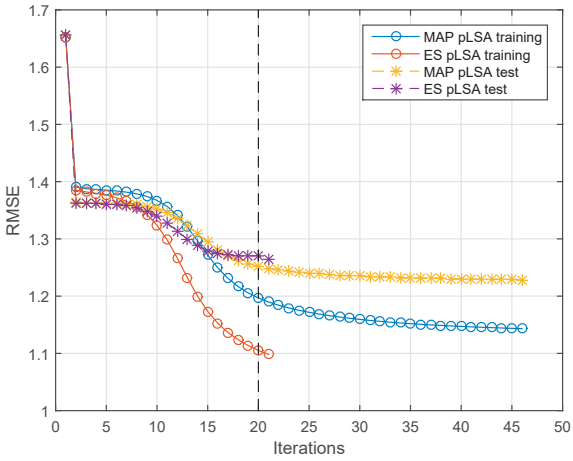

Fig. 3. RMSE for training and test data from the EachMovie data set over iterations. The state sizes are $K = 50$ and $K = 200$ for the MAP and ES pLSA respectively. The graph also shows when the early stopping occurs as the dashed line.

where $\mathbf{R}$ is the rating matrix containing all real observations and $\hat{r}_{u,i}$ is the predicted rating to the actual rating $r_{u,i}$.

To find suitable hyperparameters for the conjugate priors, we propose using a cross-validation method and a simple grid search, where the EM trainings were made with different hyperparameter values and evaluated by computing the RMSE from the trained pLSA parameters on a test set. After some test simulations, it was found that the hyperparameter $\gamma_{u,z}$ was more beneficial for decreasing the RMSE and that small variations on $\gamma_{i,r,z}$ did not have much impact to achieve smaller prediction errors. Thus, the hyperparameter values were selected to be 16 points in the interval $1.0 \leq \gamma_{u,z} \leq 1.2$ and eleven equidistant points in $1.0 \leq \gamma_{i,r,z} \leq 6.0$. As grid searches may be time consuming, the procedure were only evaluated on one test set and the state size was selected to be $K = 200$, since 200 states was determined to be a good state size selection for the multinomial pLSA model in [3]. The resulting grid search can be seen in Figure 1. Increasing the hyperparameter values results in that the MAP-based EM training reaches the convergence criteria after about five iterations compared to above 40 iterations, required otherwise. Since the conjugate priors smears out the conditional probabilities $P(z|u)$ and $P(r|i,z)$, the parameter values will be equally valued in the beginning of the learning procedure. As the hyperparameter value increases, it will be difficult for the pLSA parameters to learn from the training data, as the priors will dominate the observed data, which leads to very small changes of parameter values when the training starts. Thus, the stopping criteria might falsely indicate that the parameters already have converged, which results in poor predictions and larger errors on unseen data points. However, the RMSE do increase when the hyperparameter values are increased, which was proved by forcing the EM algorithm to perform 60 iterations.

From Figure 1, it is evident that the selection of hyperparameters is sensitive for $K = 200$. In order to obtain statistical significance from the grid search, further experiments with the grid search were performed to find the best possible values. Since $\gamma_{i,r,z}$ does not help much when decreasing the RMSE, only two values, $\gamma_{i,r,z} = 1$ and 1.5, were selected for the further experiments. We also investigated if better prediction accuracy could be found for the state sizes $K = 10$, 50, and 100, in addition to $K = 200$. Figure 2 shows the results when averaging three test sets. It is worth noting that a similar RMSE is achievable with each of the tested state sizes, although the computational time is shorter when using a smaller K. To find suitable hyperparameters to the conjugate priors, we propose using a cross-validation method, where the MAP pLSA parameters are trained with different hyperparameter value combinations, choosing the pair achieving the best results on the test sets. We examined the methods with 10 data sets picked with different random seeds and averaging the resulting prediction errors from each test set. From earlier experiments, we found that the lowest prediction errors for the ES pLSA are received with the latent state size $K = 200$. The conjugate-prior-regularized MAP pLSA performed best with $K = 50$; the results from both training procedures are shown for $\gamma_{u,z} = 1.08$ and $\gamma_{i,r,z} = 1.5$. Both model training procedures are shown in Figure 3. From this figure, it is clear that the proposed MAP pLSA has mitigated the over-fitting, since the gap between its training and test errors is reduced compared to the ES pLSA. Table 1 and 2 summarize the mean and standard deviation of the RMSE and MAE for both models and the POP estimator, where for the MovieLens data set, a similar hyperparameter grid search was performed.

ISBN 978-0-9928626-7-1 © EURASIP 2017

5. REFERENCES

[1] I. MacKenzie, C. Meyer, and S. Noble, "How retailers can keep up with consumers." "{http://www.mckinsey.com/insights/consumer_and_retail/how_retailers_can_keep_up_with_consumers}", October 2013.

[2] Y. Koren and R. Bell, *Recommender Systems Handbook*, ch. Chapter 5 Advances in Collaborative Filtering, pp. 145–186. Springer, 2011.

[3] T. Hofmann, "Latent Semantic Models for Collaborative Filtering," *ACM Transactions on Information Systems*, vol. 22, pp. 89–115, Jan. 2004.

[4] A. Theodoridis, C. Kotropoulos, and Y. Panagakis, "Music Recommendation using Hypergraphs and Group Sparsity," in *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 56–60, May 26-31. 2013.

[5] K. Alodhaibi, A. Brodsky, and G. A. Mihaila, "A Confidence-Based Recommender with Adaptive Diversity," in *IEEE Symp. on Computational Intelligence in Multicriteria Decision-Making*, April 11–15 2011.

[6] G. Dror, N. Koenigstein, and Y. Koren, "Web-Scale Media Recommendation Systems," *Proceedings of the IEEE*, vol. 100, pp. 2722–2736, Sept. 2012.

[7] Y. Koren, "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model," in *KDD*, Aug. 24–27 2008.

[8] X. Li and T. Murata, "Using Multidimensional Clustering based Collaborative Filtering Approach Improving Recommendation Diversity," in *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, (Macau), Dec. 4–7 2012.

[9] M. Nazim, J. Shrestha, and G. S. Jo, "Enhanced Content-based Filtering using Diverse Collaborative Prediction for Movie Recommendation," in *First Asian Conf. on Intelligent Information and Database Systems*, (Dong Hoi, Quang Binh, Vietnam), April 1–3 2009.

[10] Y. Song, S. Dixon, and M. Pearce, "A Survey of Music Recommendation Systems and Future Perspectives," in *Proc. 9'th Int. Symp. Computer Music Modelling and Retrieval*, June 19–22. 2012.

[11] Y. X. Zhu, W. Zeng, and Q. M. Zhang, "The Effect of Rating Variance on Personalized Recommendation," in *5th Int. Conf. on Computer Science & Education*, (Hefei, China), Aug. 24–27 2010.

[12] N. Barbierir, G. Manco, and E. Ritacco, *Probabilistic Approaches to Recommendations*. Morgan & Claypool, 2014.

[13] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds., *Recommender Systems Handbook*. Springer, 2011.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[15] J. Chu and Y. Lee, "Conjugate Prior Penalized Learning of Gaussian Mixture Models for EMG Pattern Recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (San Diego, CA), pp. 1093–1098, 2007.

[16] J. Chu and Y. Lee, "Conjugate-Prior-Penalized Learning of Gaussian Mixture Models for Multifunction Myoelectric Hand Control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, pp. 287–297, June 2009.

[17] S. I. Adalbjörnsson, J. Swärd, M. Ö. Berg, S. V. Andersen, and A. Jakobsson, "Conjugate Priors For Gaussian Emission pLSA Recommender Systems," in *24th European Signal Processing Conference*, 2016.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[19] J. Chien and M. Wu, "Adaptive Bayesian Latent Semantic Analysis," vol. 16, pp. 198–207, January 2008.

[20] Digital Equipment Corporation, "EachMovie recommendation data set." http://www.gatsby.ucl.ac.uk/~chuwei/data/EachMovie/eachmovie.html, 2004.

[21] GroupLens Research Group, "MovieLens dataset." "http://grouplens.org/datasets/movielens/1m/", 2003.

ISBN 978-0-9928626-7-1 © EURASIP 2017