Pre-processing Large Spatial Data Sets with Bayesian Methods

Saara Hyvönen, Esa Junttila, and Marko Salmenkivi

Helsinki Institute for Information Technology, Department of Computer Science
University of Helsinki, Finland
{saara.hyvonen,esa.junttila,marko.salmenkivi}@cs.helsinki.fi

Abstract. Binary data appears in many spatial applications such as dialectology and ecology. We demonstrate that a simple Bayesian modeling approach can be used in pre-processing large spatial data sets with missing or uncertain data. Our experiments on real and synthetic data show that conducting the pre-processing phase before applying conventional data mining methods, such as PCA, clustering or NMF, improves the results significantly.

Keywords: spatial data, pre-processing, Bayesian methods.

1 Introduction

While Bayesian methods have been used in confirmatory data analysis in various application areas, they have not been commonly applied to data mining tasks with relatively modest prior knowledge. In this paper we demonstrate that it is feasible to pre-process large spatial binary data sets with Bayesian methods and – what is the main point – to obtain good results in subsequent analyses.

In real applications the quality of raw data is often unsatisfactory. Missing data, noise, and uncertainty may distort the results of a straightforward application of many data mining methods. Bayesian methods provide tools for modeling, e.g., the missing data explicitly. Our experiments show that a relatively simple Bayesian model improves considerably the quality of the results achieved by different data mining methods. The applied model turns out to be practically "parameter-free": the specified prior distributions have very little influence on the results. The model is essentially based on the well-known Ising model [9], and it was first introduced in [7].

In this paper we investigate a large collection of geographical distributions of Finnish dialect words, and synthetic data sets. In particular, we study the influence of modeling the missing data with Bayesian methods on the results of three kinds of subsequent analyses: principal components analysis, k-means clustering and nonnegative matrix factorization.

Bayesian methods have been employed in spatial confirmatory data analysis in, e.g., ecology, epidemiology, and image reconstruction [1,2,5,9]. Hyvönen et al. analyze a part of the dialect data set used in this paper with several multivariate

J.N. Kok et al. (Eds.): PKDD 2007, LNAI 4702, pp. 498–505, 2007.

methods [6]. MDL-principle and association analysis methods have been applied to spatial presence-absence data [8,10]. These approaches ignore the problem of missing data.

The rest of the paper is organized as follows. We introduce the modeling approach and the dialect data set in Section 2. Section 3 compares the results of the subsequent analysis on the original and the pre-processed data. A general discussion is presented in Section 4. Section 5 is a conclusion.

2 Spatial Modeling with Markov Random Fields

Observations at two locations close to one another are often relatively similar, that is, they are spatially *autocorrelated*. Instead of trying to find out all the reasons for the similarity, spatial models often make assumptions of autocorrelation to cover the influence of these unobserved factors. Markov random fields (MRF) are typically employed to model autocorrelation.

Given a neighbor graph of regions, a probability distribution is an MRF, if $\text{PR}(X_i \mid \mathcal{X} \setminus \{X_i\}) = \text{PR}(X_i \mid \mathcal{N}(X_i))$, where $\mathcal{X}$ is the set of all variables, and $\mathcal{N}(X_i)$ is the set of variables associated with the regions that are neighbors of i. Thus, a random variable associated with region i is independent of the variables in all the other regions, given the values of variables in the neighbor regions.

Bayesian modeling requires setting up a joint distribution of the model parameters and data. It is not trivial to specify a valid distribution function that meets the Markov property in spatial domain. The function is valid (*Hammersley-Clifford theorem* [9]) if and only if it is of the form $\text{PR}(\boldsymbol{\omega}) = \frac{1}{Z} \cdot \exp((\sum_{C \in \mathcal{C}} V_C(\boldsymbol{\omega})))$. Here $\boldsymbol{\omega}$ is a vector of values of all variables in the MRF, and Z is a normalizing constant. Further, V_C is a *potential function* of clique C, and $\mathcal{C}$ is the set of all cliques in the neighbor graph. Given that $V_C(\boldsymbol{\omega})$ depends only on the values of the vertices in C, functions V_C may be chosen arbitrarily.

2.1 Dialect Word Data

During the process of writing a comprehensive dictionary of Finnish dialects, a large set of maps describing the regional distribution of the dialect words have been compiled in electronic form. Combining these distributions yields a 17,100 $\times$ 563 binary matrix (words $\times$ municipalities), the proportion of 1s being 4%.

The overall collection is far from uniform. Roughly 15 % of the municipalities have been systematically surveyed, but even here the number of recorded words varies remarkably. The collections from the rest of the municipalities are often much smaller. For the purpose of compiling the dictionary the data is generally quite good, but the spotty coverage has proved to be one of the main issues in data analysis of the collections [6]. Social relationships spread dialect words, primarily through neighboring areas. Hence, it is likely that an assumption of autocorrelation can be utilized in modeling missing data.

Denote by $y_{m,d}$ the data item indicating, whether word d was recorded in municipality m, and by $x_{m,d}$ the unknown actual status of usage of d in m. We

specify a Bayesian model that estimates for each zero in the data the probability that the correct value is actually one. The likelihood $\text{PR}(y_{m,d} = 1 \mid x_{m,d},\ r_m) = x_{m,d} \cdot r_m$ of observing a word in m depends on the unobserved quantity r_m that can be interpreted as being related to the research activity in m. The greater the value of r_m, the greater the probability of observing a word in m, given that the word is used in m. We set the uniform prior distribution $r_m \sim \text{Unif}(0, 1)$.

An MRF (Ising model) is defined for each word to model the spatial dependencies: the larger the proportion of neighbors n of m having $x_{n,d} = 1$, the more evidence we have for $x_{m,d} = 1$. Two municipalities are defined to be neighbors, if they have at least one common point in their borders. Word-specific variables β_d control the strength of autocorrelation. If $\beta_d = 0$, the neighboring municipalites are ignored. Intuitively, the greater the value of β_d the more probability mass is given to the configurations of $x_{m,d}$ with coherent areas of zeros and ones.

Fig. 1 shows a graphical representation of the model. Denote by $\boldsymbol{x_d} = (x_{1,d}, x_{2,d}, \ldots)$ all the variables in the MRF associated with dialect word d. The joint distribution of the model $\mathcal{M}$ is $\text{PR}(\mathcal{M}) = (\prod_m \text{PR}(r_m)) \cdot \prod_d \text{PR}(\beta_d) \cdot \text{PR}(\boldsymbol{x_d} \mid \beta_d) \cdot \prod_m \text{PR}(y_{m,d} \mid x_{m,d},\ r_m)$. We next specify $\text{PR}(\boldsymbol{x_d} \mid \beta_d)$ in detail.

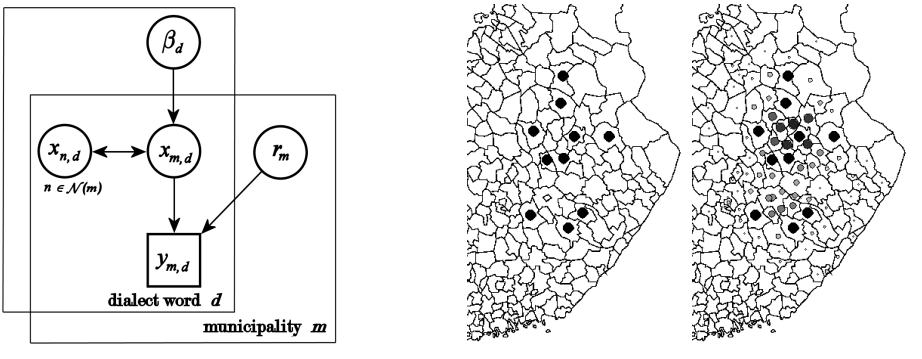


Fig. 1. Left: graph representation of a model with MRF dependencies between neighbor municipalities ($\mathcal{N}(m)$ is the set of municipalities that are neighbors of m). Right: observations of word *korahtaa*, and the approximated posterior probabilities of actual occurrences for the same word.

Each municipality forms a clique, each pair of neighboring municipalities forms a clique of size two etc. In order to achieve the desired interaction between municipalities the cliques of size two are essential. Thus, for the cliques of the other sizes we set $V_C = 0$. For the cliques of size two we assign a word-specific potential function $V_{\{m,n\}}(\boldsymbol{x_d}) = \beta_d$, if $x_{m,d} = x_{n,d}$, and 0 otherwise. Values of β_d are treated as unknown parameters. We set the prior distribution $\beta_d \sim \text{Unif}(0, 10)$, which allows no autocorrelation as well as very high correlation. We obtain [9] conditional probabilities for $x_{m,d}$ as $\text{PR}(x_{m,d} = 1 \mid \beta_d,\ x_{s,d},\ s \neq m) = Q/(1 + Q)$, where $Q = \exp((\beta_d \cdot \sum_{j \in \mathcal{N}(m)} (2x_{j,d} - 1)))$ and $\mathcal{N}(m)$ is the set of municipalities that are neighbors of m.

Based on a priori knowledge of dialect words, we know that missing observations are, in practice, potential occurrences only within a reasonable distance from some observation. Ignoring this prior information leads to incorrect simulated occurrences, particularly in edge areas. Thus, we set the probability of occurrence to zero in the remote municipalities, that is, far from any observation. After conducting trials with different reasonable criteria for remoteness, we defined a municipality to be *remote* with respect to word d, if there are no observations of d within k steps in the neighbor graph. This practice eliminated the edge effects, while the results in other respects showed no significant changes when a value of $k \geq 3$ was used.

We employed MCMC methods (see, e.g., [4]) to approximate the posterior distribution and to obtain probabilities (expectations of variables x) of word usages in municipalities. A single run (110,000 sweeps, Linux 3 GHz) on the whole data set took approx. five days. We tested the convergence with several well-known methods (Gelman–Rubin, Geweke, Heidelberger–Welch, Raftery–Lewis, see, e.g. [3]). The exact computation of $\text{PR}(\boldsymbol{x_d} \mid \beta_d)$ is intractable, and we applied the common pseudo-likelihood approximation, see [5]. Fig. 1 illustrates the marginal posterior distributions of word occurrences of a single word.

3 Subsequent Analysis

Next we compare the performance of principal components analysis (PCA), nonnegative matrix factorization (NMF) and clustering on the original and preprocessed dialect data and a synthetic data set.

Principal Components Analysis. The aim of PCA is to capture the intrinsic variability in the data. Figs. 2a,d show PC 1 for the original and pre-processed data; it essentially tells about the number of words in each municipality. We discuss Figs. 2a,d in more detail in Sec. 4. The next component in Figs. 2b,e captures the east-west variation, which is known to be the dominant direction of Finnish dialect variation. PC 3 in Figs. 2c,f shows the north-south variation.

Pre-processing clearly improves the results: the plots for the original data are grainy while the plots for the pre-processed data exhibit a smooth variation. The graininess is due to the uneven sampling of the original data. The divisions of western/eastern and northern/southern dialects are much clearer in the preprocessed data than in the original data. A similar effect is present for the subsequent components as well.

Nonnegative Matrix Factorization. Given a data matrix $\mathbf{D} \in \mathbb{R}^{m \times n}$ NMF finds nonnegative matrices $\mathbf{W} \in \mathbb{R}^{m \times k}$ and $\mathbf{H} \in \mathbb{R}^{k \times n}$ such that $\mathbf{D} \approx \mathbf{WH}$. This means that each data vector is expressed as a linear combination of k nonnegative factors (columns of $\mathbf{W}$). These factors can be interpreted as corresponding to different dialect regions. A geographical distribution of a word is then expressed as a weighted combination of these factors. A large k yields relatively local dialect regions. We show the results for $k = 3$ in Fig. 3. Indeed we observe the eastern, northern and western dialect regions in Figs. 3a,d, 3b,e and 3c,f respectively.

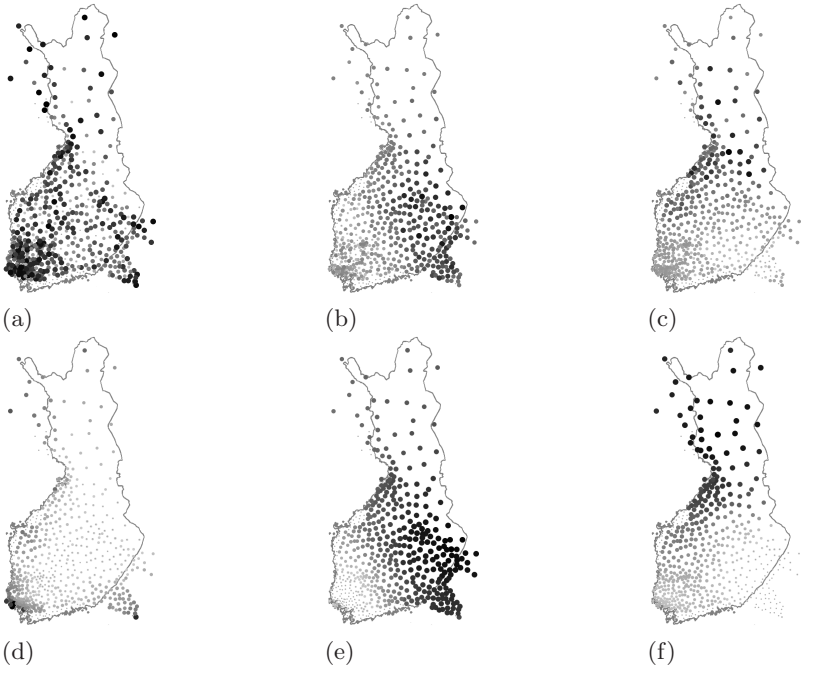


Fig. 2. The first 3 PCA components for the original (a-c) and pre-processed (d-f) data

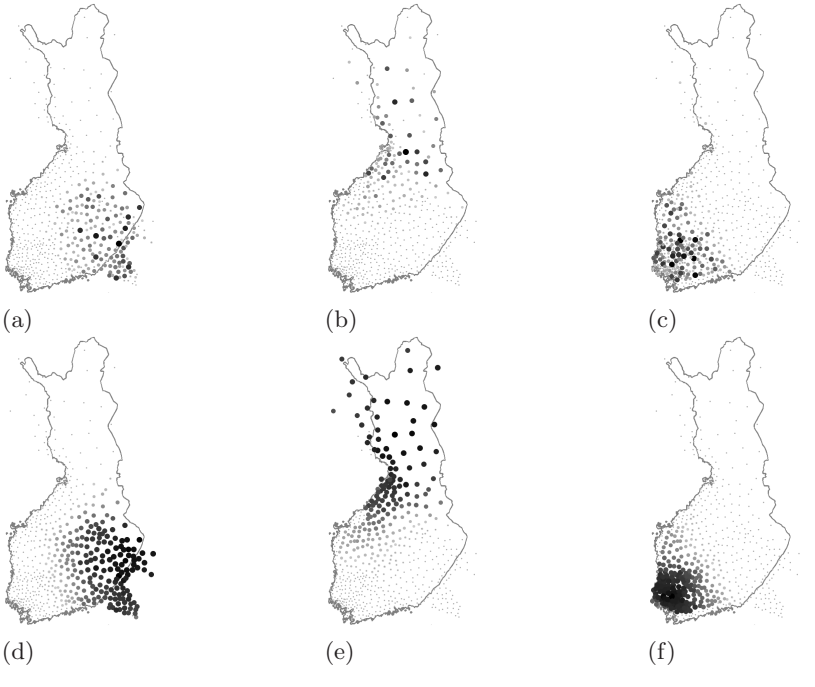


Fig. 3. NMF components ($k = 3$) for the original (a-c) and pre-processed (d-f) data

Again the factors computed on the pre-processed data are smoother and easier to interpret. This phenomenon is also apparent for different choices of k.

Clustering. Dialects are traditionally divided into specific dialect regions, which makes clustering a natural approach to dialect data. The k-means clustering using the Euclidean distance fails for the raw data, because municipalities with few words resemble each other. Fig. 4 shows that the problem of clearly incorrectly clustered points is treated by the pre-processing ($k = 6$). This also holds for other choices of k. Employing the Cosine distance will get rid of the "scatter cluster" (+): in those cases the clusterings on both data sets look almost similar.

Synthetic Data. We have also experimented on different synthetic data sets. We present the results for the simple but illustrative case, in which we generate a rectangular map consisting 18 by 20 grid cells. These are divided into three distinct regions as shown in Fig. 5a. Each of the 3000 features is assigned to a single region. Now we remove a fraction f_j of the data in each cell s_j, where $f_j \sim \text{Unif}(0.1, 0.9)$. We then use the pre-processing approach to recover the missing data. We present the results for NMF in Figs. 5b-c. The original data can be expressed in terms of three factors, such that in each factor the elements corresponding to a particular region are equal to one and all others are zero. Introducing missing data makes the components noisy (Fig. 5b), but pre-processing effectively removes this noise (Fig. 5c). In the latter case the missing cells in the factors as well as those that have moved to another factor are cells near the corners of the region borders. We next discuss this issue.

4 Discussion

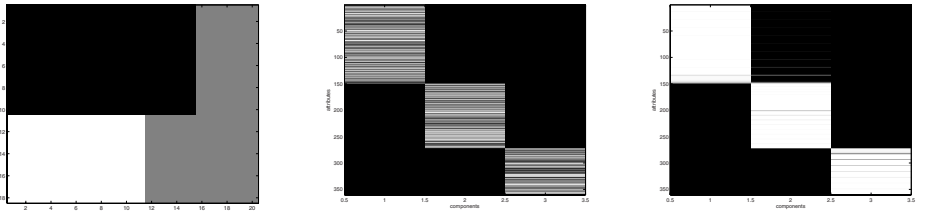
The edge areas are problematic for the plain Ising model: a small number of neighbors yields unrealistically high probabilities. In order to handle the problem we introduced the notion of remoteness in Sec. 2.1. In some cases this leads to the opposite effect, particularly in corner areas.

The first PC for the pre-processed data (Fig. 2d) correlates very strongly with the (expectation of the) number of words in each municipality. There are some municipalites with only a single neighbor and few recorded words. For instance, in the whole south-eastern region the research activity is low. Thus, the MRF structure cannot spread a lot of probality mass of new words into the southeasternmost municipality, since the nearest municipality with a large number of recorded words is remote. Bottlenecks also appear in the southwestern archipelago, where the number of neighbors is small, and the northernmost areas.

We compared our results to those of earlier linguistic research (see references in [6]). Our findings are in good agreement with them. For instance, the dark points in the north (Fig. 2f, Fig. 3e) agree with the earlier research, unlike the analyses on raw data distorted by low research rates in the most northern municipalities. Yet, one should be careful with large areas having few observations.

Fig. 6 (right) shows the posterior expectations of r_m against the number of recorded words in each municipality. These factors correlate particularly strongly,

Fig. 4. Clustering results using k-means and squared Euclidean distance for original data (left), and pre-processed data (right) for 6 clusters



(a) (b) (c)

Fig. 5. (a) Synthetic data set, (b) NMF components with missing data introduced, and (c) NMF components for (pre-processed) recovered data set

when the number of recorded words is small. Still, three outliers are discerned: the southeasternmost municipality, and two munipalities in the southwest. The use of remoteness restricts the influence of autocorrelation and decreases the probability of word being used but not observed. Thus, it leads to increasing the probabilities of great values of r_m, since r_m is identical to the probability of observing a word in m, given that the word is used in m.

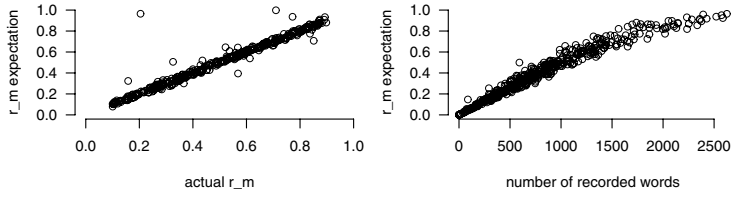


Fig. 6. Actual values vs. posterior expectations of r_m in the synthetic data (left); number of recorded words vs. posterior expectations of r_m in the dialect data (right)

In the case of the synthetic data Fig. 6 (left) shows how the model can very accurately estimate the known values of r_m for almost every cell m. There are a few outliers, always residing next to the border of two or three dialect areas.

Several neighbors of the ourlier cells have very low research rates, which increases the uncertainty. Our experiments indicated that the number of outliers decreased when the amount of data was increased. In real data the connection between the number of recorded words and values of r_m is not as straightforward as in the synthetic data, since the distributions of words are very heterogeneous.

5 Conclusion

We have demonstrated how Bayesian methods can be used as a pre-processing step in spatial data analysis. A relatively simple approach is sufficient to significantly reduce the effects of missing data. The method is practically "parameter-free", since the prior distributions have very little influence on the results. We have investigated synthetic data, and a Finnish dialect data set, that suffers from uneven sampling. We have applied the principal components analysis, nonnegative matrix factorization and clustering to both the original and pre-processed data sets. Some regions suffer slightly from edge effects. Importantly, those local effects cannot distort the results of the subsequent, global analyses.

References

1. Besag, J., York, J., Mollie, A.: Bayesian image restoration with two applications in spatial statistics. Ann. Institute of Statistical Mathematics 43(1), 1–59 (1991)
2. Best, N., Richardson, S., Thomson, A.: A comparison of Bayesian spatial models for disease mapping. Statistical Methods in Medical Research 14(1), 35–59 (2005)
3. Cowles, M., Carlin, B.: Markov chain Monte Carlo convergence diagnostics: a comparative review. J. of the American Statistical Association 91, 883–904 (1996)
4. Gamerman, D.: Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman–Hall, Great Britain (1997)
5. Heikkinen, J., Full, H.H.: Bayesian approach to image restoration with an application in biogeography. Applied Statistics 43(4), 569–582 (1994)
6. Hyvönen, S., Leino, A., Salmenkivi, M.: Multivariate analysis of Finnish dialect data – an overview of lexical variation. Literary and Linguistic Computing (2007)
7. Junttila, E., Salmenkivi, M.: Modeling missing data with Markov random fields in large data sets. In: Proc. of IADIS European Conference on Data Mining, Lisbon (2007)
8. Papadimitrou, S., Gionis, A., Tsaparas, P., Väisänen, R., Mannila, H., Faloutsos, C.: Parameter-free spatial data mining using MDL. In: Proc. of the 5th IEEE Int. Conf. on Data Mining (ICDM 2005), pp. 346–353. IEEE Computer Society Press, Los Alamitos (2005)
9. Winkler, G.: Image Analysis, Random Fields and Dynamic Monte Carlo Methods. Springer, Berlin (1995)
10. Yoo, J., Shekhar, S.: A joinless approach for mining spatial colocation patterns. IEEE Transactions on Knowledge and Data Engineering 18(10), 1323–1337 (2006)